SAPIENZA
UNIVERSITÀ DI ROMA

On the wide applicability of Bayesian hierarchical models

Department of Statistical Sciences

XXXIV Ph.D. program in Methodological Statistics

Candidate

Marco Mingione

ID number 1527640

Thesis Advisor

Prof. Giovanna Jona Lasinio

2021/2022

Thesis defended on February 22, 2022
in front of a Board of Examiners composed by:

Prof. Salvatore Ingrassia (chairman)

Prof. Maura Mezzetti

Prof. Domenico Vistocco

On the wide applicability of Bayesian hierarchical models
Ph.D. thesis. Sapienza – University of Rome

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

Author's email: marco.mingione@uniroma1.it

A Nonna Maria, con tutto il cuore.
To that part of us we give to others.

Abstract

This dissertation attempts to gather the main research topics I engaged during the past four years, in collaboration with several national and international researchers from "La Sapienza" and other universities. The primary focus is the application of Bayesian hierarchical models to phenomena in several domains such as economics, environmental health, and epidemiology. One common point is the attention to their fast implementation and results' interpretability. Typically, these two main goals are challenging to be simultaneously achieved in the Bayesian setting for two main reasons: on the one hand, the fast implementation of Bayesian machineries requires an oversimplification of the modeling structure, which does not necessarily reflect the complexity of the analyzed phenomenon; on the other hand, if the estimation of complex models is sought, parameters' interpretation may not be straightforward, especially when intricate dependence structures are present. The reader must be aware that all the presented applications with related solutions stemmed from these premises.

The first chapter of this dissertation introduces the advantages of adopting the *hierarchical paradigm* for the model formulation from a conceptual perspective.

Following this conceptual introduction, the second chapter delves more into the technical aspects of hierarchical model formulation and estimation. Far from being exhaustive, it provides all the essential ingredients for a thorough understanding of their theoretical foundations and optimal implementation. These first two chapters pave the road for the four original developments presented thereafter.

In particular, the third chapter describes a new statistical protocol aiming at variable selection within a Beta regression model for the estimation of food losses percentages at the country-commodity level. The work has been carried out in collaboration with the Food and Agricultural Organization of the United Nations, which started in 2017 for my Master's thesis and led to the recent publication by Mingione et al. (2021b).

The fourth chapter includes an extended version of the work developed during my Visiting Research period at the University of California, Los Angeles. It describes a modeling framework for the fast estimation of temporal Gaussian processes in the presence of high-frequency biometrical sampled data. Nowadays, such data are easily collected using new non-invasive wearable devices (e.g., accelerometers) and generate substantial interest in monitoring human activity. The work is currently under review and is available in Alaimo Di Loro et al. (2021b) as a pre-print.

The fifth chapter presents two modeling proposals to estimate epidemiological incidence indicators, typically collected during an epidemic for surveillance purposes. The methodology was applied to the Italian publicly available data for the monitoring of the COVID-19 epidemic. Both proposals consider probability distributions coherent with the nature of the data, which are *counts*, and adopt a generalized logistic function for the parametrization of the mean term. However, the second proposal allows for a latent component accounting for dependence among geographical units. Notice that, in the first work by Alaimo Di Loro et al. (2021a), the inference is pursued under a likelihood-based framework. This work helps highlighting even more the advantages of using a Bayesian approach, as subsequently described by Mingione et al. (2021a).

The last chapter summarizes the main points of the dissertation, underlining the most relevant findings, the original contributions, and stressing out how Bayesian hierarchical models altogether yield a cohesive treatment of many issues.

Contents

Chapter 1

Motivation and Introduction

> "The best thing about being a statistician is that you get to play in everyone's backyard"
>
> *John Tukey*

When I attended my first year at the university, I never imagined I would find myself dazzled by the flair of statistics and today, in truth, I do not understand why every person does not wish to become a statistician. Indeed, among all *disciplines*, statistics is probably the most comprehensive in the sense that it serves all other *sciences*, continuously attempting to shape the world and society as a whole. With this work, although being a grain of sand, I hope to give my contribution in this field and I wish I could say this is only the beginning.

Among the countless stimulating challenges statistical research poses, and the plethora of possible approaches to the solution, I often ended up dealing with the advanced handling of complex phenomena through Bayesian hierarchical models. Loosely speaking, complex phenomena may be described by the combination of many sub-components interacting with each other. Truth be told, the majority of real-world events are the result of entangled relationships, regardless of the field of interest. Therefore, whenever studying these relationships from a statistical perspective, their intrinsically complex nature must be taken under proper consideration. In strict statistical terms, this complexity usually arises in different forms, sometimes apparent and other times concealed, such as: heterogeneity of data sources, presence of missing data, unobserved variability. In this respect, statistics experienced giant leaps forward in the statistical methodology and analysis to keep pace with the scientific and technological progress momentum in other fields such as climatology, ecology, environmental health, and economics. Since the early 2000s, we have been assisting to a paradigm shift where intently gathered experimental data gave way to the increasing availability of observational data. A thorough examination of complex systems using such data, often requires integration of multiple sources of information and necessitates to look at the big picture as a sequence of smaller frames, each one with its own peculiarity. Following the scheme *divide-rule-combine*, the goal is to provide *ad hoc* solutions to practical problems which require the analysis of data that are highly multivariate, geographically referenced and/or temporally correlated, with specific attention to the computational aspects and the interpretability of the results. To this scope, the Bayesian paradigm for statistical analysis provides a convenient framework for combining complex data models and external knowledge by modeling both observed data and any unknown as random variables, allowing for a proper quantification of uncertainty in the process of decision making. As it will become even more clear in the following chapters, this dissertation will deal with the technical and computational aspects of formal inference using Bayesian techniques

and the adoption of the **hierarchical paradigm** for the model formulation will be essential. Further details about the enhancements with respect to the state of the art will be given in each specific section. However, regardless the framework adopted, hierarchical modeling should be seen as a conceptual and philosophical approach of doing science since it emphasizes model construction and depicts the easiest and most general way of going from the problem to the solution.

1.1 The *hierarchical paradigm*

Generally speaking, the term *hierarchical* refers to "a system in which people or things are arranged according to their importance" (Cambridge Dictionary). In the statistical field, the concept of hierarchy does not go far beyond its original definition, but it usually indicates the data structure and/or the model formulation. In the scientific literature, especially in social sciences, hierarchical analysis is also referred to as (or considered within) *multi-level analysis*, and related methodologies have been developed strongly before they were actually implemented. However, the existence of such hierarchies is neither accidental nor ignorable.

For instance, let us assume there are variables describing individual characteristics, but these individuals belong to larger categories (each one consisting of more than one individual), and there are also variables describing these categories. More practically, let us assume the interest lies in estimating the average score for humanistic subjects for high schools students in 2020. The first step is to select a sample of students and record their grades in philosophy, history, literature and law. If possible, it could also be useful to record the average number of hours they spent studying per week, reasonably expecting that the more the time spent studying, the higher the score. If we estimate a standard linear regression model, we would be neglecting the fact that the students are grouped in classes, and each class has its teacher, with varying teaching skills. Moreover, classes may belong to different schools, which in turn may be located in different neighborhood or cities, and so on. We have variables describing classes (e.g. size), variables describing schools (e.g. school building) and also variables describing neighborhoods (e.g. economic status). These *higher-order* variables are assigned to the individual since the analysis is conducted at the individual level. Generally speaking, we can talk about primary units or populations, secondary units, groups sub-populations, and simple units or individuals. Eventually, this kind of structure may allow for further levels according to the complexity of the problem at hand. To overlook such kind of dependence structure would lead to dramatically wrong conclusions and flawed inference.

The previous example can be interpreted as a kind of *individual within group* hierarchy, in which a nesting structure could be identified up to the fourth level (e.g. students within classes within schools within neighborhoods). Other similar data structures arise in *longitudinal data*, as often occurs in econometric applications, or when *repeated measurements* are recorded for the same unit, as it is common practice in clinical trials to study the effect of a treatment or human growth. A further type of hierarchy includes the so-called *non-nested* structures, where individual observations are nested within *groups/clusters*, but neither of them can be ordered/is above the other in a hierarchical sense.

At last, even though falling into abeyance, hierarchical models are also known as *random-effects* or *mixed-effects* models. In particular, the term "random effects" refers to the regression coefficients that are considered random outcomes themselves. This is in contrast with the term "fixed effects", which refers either to parameters that do not vary (e.g. individual-specific intercepts) or to parameters that vary but

are not modeled themselves (e.g. indicators for categorical explanatory variables). A "mixed-effects" model includes instead both fixed and random effects[1]. As for these informal definitions, it is evident however that there is also no commonly agreed mathematical formulation for such kind of models, even though the literature is full of attempts in solving this issue (Kreft et al., 1998; Searle et al., 1992; Green and Tukey, 1960; LaMotte, 2014; Snijders and Bosker, 2011). In general, the terms "fixed", "random" and "mixed" are confusing and often misleading, so we will avoid their use in what follows, unless it is strictly necessary.

In the last three decades, there has been a substantial growth in the usage of hierarchical models due to their flexibility and wide applicability in a rapidly expanding range of fields, such as agriculture (Henderson, 1984; Robinson et al., 1991), educational statistics (Bock, 2014), social sciences (Longford, 1995; Kreft et al., 1998; Goldstein, 2011; Snijders and Bosker, 2011; Raudenbush and Bryk, 2002) and environmental sciences (Clark and Gelfand, 2006; Royle and Dorazio, 2008) among others. Most cited works discuss hierarchical model formulation from the frequentist perspective, yet understanding the Bayesian standpoint is somehow more thorough, as we will see in the next chapter. Generally speaking, we can say that all multilevel models are Bayesian to the extent that parameters are not unknown and fixed quantities, but random variables whose inference is sought. A complete and rigorous overview of hierarchical models theory and implementation with `R` is available in Gelman and Hill (2006).

Before moving into the core part of these dissertation, the reader should understand that the hierarchical paradigm is utterly helpful in clarifying the nature of the inference problem in a mathematically and statistically precise way, by focusing on its conceptually and scientifically distinct components. Hence, while hierarchical models yield a cohesive treatment of many technical issues, they also foster the fundamental activities of model building and inference.

1.2 Content of the thesis

In this brief introduction, I hope the reader already got the scent of the potential of hierarchical thinking and the advantages of the Bayesian approach to statistical analysis. To provide an even clearer description of these tools, Chapter 2 is entirely focused on Bayesian hierarchical modeling formulation and estimation from both the qualitative and quantitative perspectives. At first, the concept of hierarchy in Bayesian statistics is discussed, naturally leading to the model's generic formulation. Secondly, the main estimation techniques are described, highlighting their pros and cons. Chapter 2 does not have the presumption of including all the exhaustive details about the subject at hand, neither it has the rigour of a book in describing the methods; however, the goal is to provide a comprehensive conceptual account of their theoretical foundations, the intuition behind them and their optimal implementations. I hope the dedicated reader will acquire a solid grasp of the motivation that leads to Bayesian hierarchical models' choice for the proposed applications, why they work, when they succeed, and perhaps most notably, when they fail.

Then, in the following chapters of this dissertation, I present the four most valuable publications of my whole scientific production. The *fil rouge* bringing all of them together is the application of Bayesian hierarchical models, with special attention to their fast and efficient implementation. Each original work was motivated by a real

[1]Please, note that this latter distinction is more appropriate in the frequentist setting rather than in the Bayesian one, as in this case, both the data and the parameters are considered **random** quantities

data problem and intended to provide the *best* solution from the methodological perspective, without overlooking the interpretability of the results.

Chapter 3 presents the research work following my Master's thesis project and now published in Mingione et al. (2021b). In this work, we proposed an alternative approach to what is officially reported in the State of Food and Agriculture (FAO, 2019) to estimate yearly food losses percentages at the country-commodity level. In particular, our approach is based on a Bayesian Beta regression model with a variable selection step. Proper estimation of food losses is key in the calculation of the *Food Loss Index*, which is used by the Food and Agricultural Organization of the United Nations to monitor progress towards the Sustainable Development Goal n. 12: responsible consumption and production. Identification of the most important factors explaining what drives food losses dynamics worldwide is equally crucial for implementing prevention policies.

Chapter 4 describes the original work developed during my visiting research period at University of California, Los Angeles (UCLA) between September 2019 and March, now under review and published by Alaimo Di Loro et al. (2021b) as a preprint. The research covered the efficient implementation of *Nearest Neighbor Gaussian Process* (Datta et al., 2016a; Finley et al., 2019), and its application to estimate physical activity level trajectories on a large scale population study, using data collected through modern accelerometer and GPS devices.

Chapter 5 introduces a novel parametric regression model to fit *incidence indicators* typically collected during epidemics. This work is the result of a joint project of a group of statisticians who share the same commitment to the social role of statistics, but are aware of the pitfalls that can stem from poor quantitative communication. In particular, the first part of this chapter is dedicated to the first proposal by Alaimo Di Loro et al. (2021a), developed during the first outbreak of COVID-19 epidemic; in the second part, the focus instead relates on the extension of the work mentioned above, now published in Mingione et al. (2021a).

A general discussion is given in Chapter 6, highlighting the important findings and the contributions. Nevertheless, this dissertation still leaves room for many questions and open problems, hence some thoughts regarding promising directions for future research will be discussed.

Chapter 2

Bayesian hierarchical modeling

"Diviser pour régner"

Luigi XI

The first hints about the notion of a hierarchy date back to the 1960s, when the classification of *kinds* of probability from the Bayesian perspective was debated both from a philosophical (Good, 1959), and mathematical point of view (Good and England, 1965). In subsequent work, it was noted that different stages of probability arise naturally, whether in the theory of physical probabilities, subjective probabilities, or a mix of both (Good, 1980). Nevertheless, the widespread usage of hierarchical models had to wait another decade before standing out other methodologies. Since then, a huge amount of related literature was produced, some more focused on the theory (Berliner, 1996; Gelman et al., 2013; Cressie, 2015; Banerjee et al., 2014; Gelfand et al., 2019), other on the applications (Raudenbush and Bryk, 2002; Clark and Gelfand, 2006; Royle and Dorazio, 2008; Congdon, 2019).

Formal definitions of hierarchical modeling are plenty. However, the most comprehensive and endorsed by researchers is given by Gelman (2006), who states:

> *"Hierarchical modeling is a generalization of linear and generalized linear modeling in which regression coefficients are themselves given a model, whose parameters are also estimated from data."*

The above definition suits both the frequentist and the Bayesian approach, even though the debate about this matter is still open (Allen, 2017).

The goal of this chapter is to provide the basic ingredients of Bayesian hierarchical modeling formulation and implementation, in order to make even more clear the methodological choices proposed in the following applications to the thoughtful reader. Given these premises, Section 2.1 describes the generic formulation of Bayesian hierarchical models from the conceptual and technical perspectives. Section 2.2 instead introduces the main Markov Chain Monte Carlo methods for the estimation of models' parameters, highlighting pros and cons of each of the described algorithm.

2.1 The generic formulation

When it comes to building a model which has to account for complex structures and a large variety of random quantities, it may be helpful to break it into little pieces. From a statistical point of view, this means that a manageable *joint* probability distribution for all the random quantities involved may not be derived straightforwardly, and the problem should instead be tackled from a *conditional* perspective. Indeed, restoring to basic probability theory, any complex joint distribution can be factorized into simpler conditional distributions. As we will see hereafter, this

approach is intrinsically hierarchical, as these conditional distributions are somehow naturally ordered. Moreover, although the model can also be formulated from the frequentist perspective, following the Bayesian paradigm facilitates the inclusion of *prior* beliefs on the outcome and may help in better quantifying the uncertainty of the final estimates. Note that the adopted stochastic models are only approximations of the complex processes affecting real phenomena, hence some error will always be introduced. The benefit of the transparency implicit in this way of building models is that it allows to determine *where* and *how* to introduce the error.

When specifying a Bayesian model, the main assumption is that both the data (*data*) and the parameters (*pars*) describing the data generative mechanism are random variables, and the uncertainty should be quantified in terms of their joint distribution. Specifically, the inference is based on the posterior distribution of the *pars* given the *data*, which represents the statistical compromise between the prior knowledge and the observed information. By applying the Bayes' Theorem, this can always be expressed as:

$$\pi\left(pars|data\right) = \frac{J\left(data, pars\right)}{m\left(data\right)} \propto \mathcal{L}\left(data|pars\right) \cdot \pi\left(pars\right), \tag{2.1}$$

where $\pi(\cdot)$ and $\pi(\cdot|data)$ represent the prior and the posterior distribution of the *pars*, respectively; $J(\cdot,\cdot)$ indicates the joint distribution; $m(\cdot)$ is the normalizing constant which does not depend on the *pars*; $\mathcal{L}(\cdot|pars)$ is the likelihood of the data. In general, the likelihood is chosen to be coherent with the nature of the data. At the same time, the prior can either follow as a convenient combination with the likelihood (i.e., conjugate families) and/or consider pre-existing information (e.g., expert opinion or past statistical analysis).

Equivalently, Equation (2.1) can be seen as a *two-level* structure:

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

where each level may envision the presence of additional sub-levels as, for example, the specification for the typically unknown *pars* (e.g., the *hyperprior*).

Albeit simple, the consideration of such *two-stage* hierarchical structure has been revolutionary for statistical modeling[1], especially for models requiring the specification of complicated dependence structures. Following this approach, it is possible to keep the classical independence assumption at the *data* level, averting the specification of elaborate dependence structures directly on the outcome variable (Gelman et al., 2013). In some applications, these intricate dependence structures can be dealt with the addition of a latent process (*proc*). In practice, this means placing a further level to the hierarchy in between the likelihood and the prior specification. Following Berliner (1996), we can imagine a *three-stage* hierarchical specification:

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

[1]It changed the way of doing science in many fields: today, with the power of computers, applications of Bayes' Theorem go from climatology (e.g. weather forecast, Abramson et al. (1996); Di Narzo and Cocchi (2010)) to computer science (e.g. spam detection, Eberhardt (2015); Rathod and Pattewar (2015)).

where the form of $\pi\left(proc|pars\right)$ is arbitrary, although *Gaussianity* represents a convenient choice in most of the applications, due to its several probabilistic properties.

This framework decomposes a complicated generative process into three primary components linked by simple probability rules. Moreover, this partitioning allows specifying simple models at each stage that, when combined, can describe very complex joint data, process, and parameters distribution (Gelfand et al., 2019). As thoroughly discussed in Gelfand (2012), either the first or the second stage of the hierarchical specification can be shaped according to the peculiarities of the problem under consideration. The process specification can, in turn, envision the inclusion of latent components either independent or correlated to account for unobserved heterogeneity.

The ultimate interest is performing inference on the model parameters, (sometimes) on the latent process, and providing predictions for the outcome at unobserved units. In general, all these tasks can be pursued in terms of the parameters' and process' posterior distributions:

$$\pi(proc, pars|data) \propto \mathcal{L}(data|proc, pars) \cdot \pi(proc|pars) \cdot \pi(pars) \tag{2.2}$$

and of the posterior predictive distribution:

$$\pi(\widetilde{data}|data) = \int \pi(\widetilde{data}|proc, pars, data) \cdot \pi(proc, pars|data) dpars\, dproc. \tag{2.3}$$

From the mathematical perspective, hierarchical modeling formulation involves the specification of the distributional model (e.g., the likelihood) $f(y|\theta)$ for the data $y = (y_1, \ldots, y_n)$ given a vector of unknown parameters $\theta = (\theta_1, \ldots, \theta_k)$, where we suppose that θ is a random quantity sampled from a prior distribution $\pi(\theta|\lambda)$, and λ is a vector of hyperparameters. In practice, also λ is unknown, therefore the hyperprior $\pi(\lambda)$ will often be required, leading to the generic expression of Equation (2.2) as:

$$\pi(\theta|y) = \frac{\pi(y, \theta)}{\pi(y)} = \frac{\int f(y|\theta)\pi(\theta|\lambda)\pi(\lambda)d\lambda}{\int f(y|\theta)\pi(\theta|\lambda)\pi(\lambda)d\theta d\lambda}, \tag{2.4}$$

and the posterior predictive distribution in Equation (2.3) as:

$$\pi(\widetilde{y}|y) = \int \pi(\widetilde{y}|\theta, y) \cdot \pi(\theta|y)d\theta. \tag{2.5}$$

A useful instrument that can be used in the construction of such models to ease the understanding of the underlying structure of the problem is a Directed Acyclic Graph (DAG). This tool is extremely valuable whenever the goal is to represent a set of variables and their conditional dependencies in a hierarchical structure. In particular, the nodes in the graph correspond to data or parameters (or any random variable in the Bayesian sense), while directed edges between the nodes represent conditional distributions. For example, the DAG-based representation of Equation 2.4 is reported in Figure 2.1[2]. Obviously, this structure can be arbitrarily complicated, yet this simple representation highlights how interpretable and explainable may be even the most complex model.

Nevertheless, DAG structures arise even more the computational concerns about the estimation of such models. Indeed, aside from specific and rare cases, a DAG representation illustrates how posterior distributions lack a closed-form solution,

[2]Please note that this is a toy example, where each observation depends on just one parameter. However, in practice, a single parameter often can condition more than one observation.

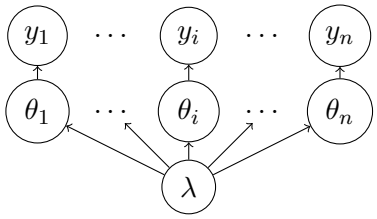


Figure 2.1. DAG representation of a simple Bayesian hierarchical model.

like the ones in Equation (2.4) and Equation (2.5). Their computation, therefore, relies on simulation techniques, such as Markov Chain Monte Carlo (MCMC, Brooks et al. (2011)), including Hamiltonian Monte Carlo (HMC), either of which will be introduced in Section 2.2. Although, in principle, these equations can be simplified by suitable marginalization/integration, the advantage of this specification lies in the convenience of formulation, ease of interpretation, and, often, in the facilitation of model fitting. Nevertheless, according to the size of the dataset and the complexity of the specifications, such model fitting can be very challenging, sometimes unfeasible. Limitations of hierarchical modeling will become more of a constraint as we seek models that stretch the limits of our computing capabilities. However, they represent powerful tools, and their application still brings more advantages than drawbacks.

Because of the above, hierarchical modeling has unsurprisingly taken over the landscape in contemporary stochastic modeling. It has been chosen to be the primary methodological tool for the applications presented in the rest of this dissertation.

2.2 Markov Chain Monte Carlo methods

As discussed in Section 2.1, the main object of interest for carrying out Bayesian inference is the posterior distribution or any of its summaries. Although its definition (at least up to a proportionality constant) is straightforward from a theoretical point of view, obtaining closed-form expressions is not trivial in most cases. Unless working with *conjugate* families, if we try to calculate the posterior distribution analytically for such models, the algebra starts to overwhelm the statistical science almost entirely, making the full Bayesian analysis too cumbersome for most practical applications. Fortunately, a battery of powerful methods has been developed over the past few decades for approximating integrals and simulating from probability distributions. Integration arises for calculating the normalizing constant for the posterior as in Equation (2.4), the posterior predictive distribution in Equation (2.5) or posterior summaries of $\pi(\theta|y)$, such as expectation, credible intervals, etc. These techniques are known as *Monte Carlo* methods (Robert and Casella, 2013) and fall into the so-called *simulation based* inference, generally referring to those algorithms which allow simulating random processes.

Following Brooks et al. (2011), the typical problem consists in the evaluation of

$$\mathbb{E}_\pi \left[g\left(\theta\right) \right] = \int_\Theta g\left(\theta\right) \pi\left(\theta\right) d\theta,$$

where Θ is the domain of the random variable θ, which usually coincides with the support of the density function $\pi(\cdot)$, and $g(\cdot)$ is any valid function of θ.

The idea is to simulate a sample $\{\theta_1, \ldots, \theta_M\}$ from the density $\pi\left(\cdot\right)$ and approximate the integral by the empirical average:

$$\bar{g}_M = \frac{1}{M} \sum_{m=1}^{M} g(\theta_m),$$

which converges almost surely to $\mathbb{E}_\pi[g(\theta)]$. Theoretical justification for this solution relies on the application of the Strong Law of Large Numbers for M *sufficiently large*. Moreover, when $\mathbb{E}_\pi\left[g^2(\theta)\right] < \infty$, estimation of the asymptotic variance of $\bar{g}_M$ can be obtained as

$$\text{var}\left[\bar{g}_M\right] = \frac{1}{M^2} \sum_{m=1}^{M} \left(g(\theta_m) - \bar{g}_M\right)^2,$$

hence using the Central Limit Theorem

$$\frac{\bar{g}_M - \mathbb{E}_\pi[g(\theta)]}{\sqrt{\text{var}\left[\bar{g}_M\right]}} \xrightarrow{d} \mathcal{N}(0,1),$$

it is possible to build confidence bounds and convergence tests.

The advantage of using *probabilistic integration* rather than deterministic numerical methods (e.g., trapezoidal or Simpson's rule, Smith (1991)) is twofold: the latter fail to spot the region of importance for the integrating function, wasting computational effort in the evaluation of the integral at unimportant areas, and present the problem of multi-modality, which largely affects their accuracy. This implies that numerical methods cannot easily face the high dimensionality of the probability distributions involved in most of the statistical problems. This issue is also known as the *curse of dimensionality*, meaning that the volume of the sample space increases exponentially with the number of parameters.

However, simulation-based methods also present some limitations, mostly related to the ability of simulating from the target distribution π. That is typically inconvenient in Bayesian hierarchical modeling, especially when elaborate dependence structures are present, as these models produce highly complex probability distributions that are difficult to sample from directly. A suitable solution is provided by Markov Chain Monte Carlo (MCMC) methods. In principle, they were used mainly by chemists and physicists to simulate particles movement but later became essential for applied (Bayesian) statisticians. They can be seen as a subset of Monte Carlo methods, which comprise an extensive class of algorithms primarily used to calculate multidimensional integrals' numerical approximations. The novelty is that they provide an alternative whereby the sampling occurs directly from the posterior, deriving the sample estimates of the quantities of interest, namely performing the integration implicitly (Brooks, 1998).

The idea of MCMC sampling was first introduced by Metropolis et al. (1953) and later generalized by Hastings (1970). Suppose that the target distribution distribution $\pi(\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^k$, is only known up to some multiplicative constant. If π is sufficiently complex that we cannot sample it directly, an indirect method to obtain samples from π is to construct an **aperiodic** and **irreducible** Markov chain with state-space Θ, whose stationary (or invariant) distribution is $\pi(\theta)$ itself (Smith and Roberts, 1993). Then, if we run the chain for a sufficiently long time, simulated values can be treated as a dependent sample from the target distribution and used as a basis for summarizing important features of π. More technically, a MCMC method for the simulation of a distribution π is any method producing an **ergodic** Markov chain whose stationary distribution is π.

MCMC techniques are fundamental to solve the problem of simulating from intricate models by sampling from the target distribution indirectly (conditionally),

but hinder the possibility of constructing independent samples[3]. As it will be more clear in the next section, high-autocorrelation is present because the simulation scheme involves repeated occurrences of the same value. However, the lower the autocorrelation, the greater the amount of information contained in a given number of draws from the posterior; this is referred to as the *efficiency* or *mixing of the chain*. Controlling for the autocorrelation is important since, if present, it affects the precision of our estimates. In particular, the autocorrelation of a Markov-chain inflates the standard error of the sample mean by a factor, increasing more than exponentially as the autocorrelation approaches its maximum value. The value of this inflation factor gives an idea of the *effective size* of the Markov-chain needed to provide an unbiased picture of the target distribution. For example, an autocorrelation equal to 0.95 inflates the standard error by a factor which is $\simeq 40$. That means that we would need roughly forty times as many points as are required for the same precision as with an uncorrelated sequence (Gelman et al., 1992). This comes with computational concerns because long chains are needed to achieve stationarity/good approximation. Such computational concerns limited the widespread implementation of MCMC algorithms until recently, when poor computing methods and inadequate processing infrastructures were replaced by more efficient and comprehensive tools. Statisticians were eventually able to estimate sophisticated models providing accurate representations of the observed data, instead of settling for simpler models.

The formal definition of a Markov chain and its properties are beyond the scope of this work. The theory behind these methods is well-establish and has been thoroughly studied. The general idea of MCMC sampling provided above is sufficient to understand the main passages of the following sections. However, the author points the more keen reader to Meyn and Tweedie (2012) and Robert and Casella (2013) for a detailed technical introduction about Markov chains theory and theoretical results on the convergence of MCMC algorithms, respectively.

2.2.1 Gibbs sampler

Gibbs sampler is probably the most straightforward MCMC sampling technique. Its implementation depends on two iterative steps, which can only be computed if the full conditional distributions of the parameters are available. It was proposed by Geman and Geman (1984) who chose the name "Gibbs sampler" because the distributions used in their context (i.e., image restoration, where the parameters were the colors of pixels on a screen) were Gibbs distributions (Gibbs, 1902), and it was made famous in the statistical community by Gelfand and Smith (1990).

More in detail, let us suppose our model envisions a set of k parameters, $\theta = (\theta_1, \ldots, \theta_k)$. As aforementioned, to implement the Gibbs sampler we must assume that samples can be generated from each of the full or complete conditional distributions $\{\pi(\theta_i|\theta_{j\neq i}, y)\}_{i=1}^{k}$ in the model. These samples might either be available directly (e.g., in closed-form from known probability distributions) or indirectly (e.g., obtained using other sampling schemes, as for example, the adaptive rejection sampling algorithm of Gilks and Wild (1992)). In both cases, the joint posterior distribution $\pi(\theta|y)$ is entirely determined (under mild conditions) by the collection of full conditional distributions. Consequently, all marginal posterior distributions $\pi(\theta_i|y), i = 1, \ldots, k$ are also determined. Following Banerjee et al. (2014); Robert and Casella (2013), the generic formulation to obtain M posterior samples for the

[3]Note that convergence properties are still valid, by means of the *ergodic theorem* (Robert and Casella, 2013).

Algorithm 1: Gibbs sampling scheme.

0: Initialization: define an arbitrary set of starting values $\theta^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_k^{(0)}\right)$

1: Simulation:

for $m = 1, \ldots, M$ **do**

  Draw $\theta_i^{(m)}$ from $\pi\left(\theta_i | \theta_1^{(m-1)}, \ldots, \theta_{i-1}^{(m-1)}, \theta_{i+1}^{(m-1)}, \ldots, \theta_k^{(m-1)}, y\right)$, $\forall i$

end

vector parameter θ using the Gibbs sampler is very straightforward and can be described as in Algorithm 1.

As proved by Geman and Geman (1984) in their seminal paper, or by Smith and Roberts (1993) in a review, the generated k-tuple at iteration m, $\{\theta_1^{(m)}, \ldots, \theta_k^{(m)}\}$, converges in distribution to a draw from the true joint posterior distribution $\pi(\theta_1, \ldots, \theta_k | y)$, as long as some weak regularity conditions hold. This result implies that for m sufficiently large (e.g. larger than a threshold $\tilde{m}$), the set of k-tuple $\{\theta^{(m)}\}_{m=\tilde{m}+1}^{M}$ is essentially a (correlated) sample from the *true* posterior, from which any posterior quantities of interest may be estimated.

2.2.2 Metropolis-Hastings algorithm

The ease of implementation and understandability of the Gibbs sampler described above comes at a cost: it is mandatory to be able to sample from each of the full conditional distributions promptly. This is rarely the case when the prior distribution and the likelihood are not a conjugate pair (Diaconis et al., 1979), as it is cumbersome to derive closed-form expressions for these full conditionals. Nevertheless, the latter are often available up to a proportionality constant that does not depend on θ. The Metropolis (or Metropolis-Hastings) algorithm fits in such context as it precisely tackles this issue. It was firstly proposed by Metropolis et al. (1953) and later generalized by Hastings (1970). The Metropolis sampler was not developed for statistical purposes in principle, but conceived by physicists to simulate the fluid particle movements in equilibrium with its gas phase. It is based on a *rejection step* for which a *candidate density* must be chosen, and it only requires a function proportional to the distribution to be sampled.

Let us suppose our main interest is to obtain posterior samples for the vector parameter $\theta = (\theta_1, \ldots, \theta_k)$. In other words, we wish to generate from the joint posterior distribution

$$\pi(\theta | y) \propto h(\theta) \equiv f(y|\theta)\pi(\theta).$$

First of all, there is the need to specify the candidate (also called *proposal*) density, which from now on will be referred to as $q(\cdot|\theta)$. It has to be a valid density function for every possible value of the conditioning variable θ, and it should be relatively easy to simulate from. Following Banerjee et al. (2014); Robert and Casella (2013), the generic formulation to obtain M posterior samples for the vector parameter θ using the Metropolis-Hastings can be described as in Algorithm 2.

It can be proved that, under the same mild conditions required for the Gibbs sampler, the generated k-tuple at iteration m with the Metropolis-Hastings, $\{\theta_1^{(m)}, \ldots, \theta_k^{(m)}\}$, converges in distribution to a draw from the true joint posterior distribution $\pi(\theta_1, \ldots, \theta_k | y)$.

Algorithm 2: Metropolis sampling scheme.

0: Initialization: define an arbitrary set of starting values $\theta^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_k^{(0)}\right)$

for $m = 1, \ldots, M$ **do**

1: Draw θ^* from $q(\cdot|\theta^{(m-1)})$

2: Accept

$$\theta^m = \begin{cases} \theta^* \text{ with probability } \rho \\ \theta^{(m-1)} \text{ with probability } 1-\rho \end{cases}$$

end

where

$$\rho = \min\{1, \frac{h(\theta^*)}{h(\theta^{(m-1)})} \cdot \frac{q(\theta^{(m-1)}|\theta^*)}{q(\theta^*|\theta^{(m-1)})}\}$$

This algorithm always accepts the value θ^* such that its *plausibility*, ρ, is increased compared with the previous value, although it may also accept values θ^* such that the ratio is decreasing. Obviously, ρ is only defined when $h(\theta^{(m)}) > 0$. However, if the chain starts with a value $\theta^{(0)}$ such that $h(\theta^{(0)}) > 0$, it follows that $h(\theta^{(m)}) > 0$ for every $m \in \mathbb{N}$ since the values of such that $h(\theta^*) = 0$ lead to $\rho = 0$, and are therefore rejected by the algorithm.

One important thing to highlight is that, since full conditional distributions for well-defined models are unique, the steps of the Gibbs sampler are fully determined by the considered statistical model. On the other hand, when using the Metropolis algorithm, the choice of the candidate density is crucial. If appropriately selected, the algorithm affords substantial flexibility and approximates the posterior distribution of more sophisticated models. However, even though we are free to pick almost any candidate density from a strictly theoretical point of view, in practice, only *good* choices lead to sufficient candidate acceptances. Theoretically speaking, an *optimal* choice of q would lead to an empirical acceptance ratio equal to 1, with no apparent waste of candidates. This is exactly what happens when the proposal comes from the full conditional distribution and is always accepted, defining the Gibbs sampler as a particular case of the Metropolis-Hastings. However, a correct specification of q is rather more subtle: accepting all or nearly all candidates is often the result of an overly narrow candidate density. The latter implies a poor exploration of the parameter space, as the proposal will slowly move around the support for the target distribution, leading to high acceptance and autocorrelation in the sampled chain. An overly dispersed candidate density will struggle likewise, proposing leaps to places far from the core part of the support of the posterior, leading to high rejection and, again, high autocorrelation.

A convenient approach considers symmetric candidate distribution, namely satisfying $q(\theta^*|\theta^{(m-1)}) = q(\theta^{(m-1)}|\theta^*)$. Typically, the common solution is to set

$$q(\theta^*|\theta^{(m-1)}) = \mathcal{N}\left(\theta^*|\theta^{(m-1)}, \tilde{\Sigma}\right), \tag{2.6}$$

since this distribution obviously satisfies the symmetry property, and is *"self correcting"* (candidates are always centered around the current value of the chain). Equation (2.6) is usually referred to as a "Gaussian random walk" proposal density. Specification of q then comes down to specification of $\tilde{\Sigma}$. A simple option would be

setting $\tilde{\Sigma}$ equal to an empirical estimate derived from a preliminary run. However, this does not guarantee any optimal property. Hence, in general, it is common practice *tuning* $\tilde{\Sigma}$ so that roughly 25 – 40% of the candidates are accepted (Gelman et al., 1997), with the optimal value depending both on the the dimension and the *true* posterior correlation structure of θ. However, such a procedure may be time-consuming, and, in practice, a solution has to be obtained in a reasonable amount of time. A possible remedy involves the implementation of *adaptive strategies*, and has been proposed in the literature in different ways (Evans, 1991; Gilks and Wild, 1992; Gelfand and Sahu, 1994; Gilks et al., 1998; Haario et al., 1999, 2001). In other words, the basic idea is to use the history of the process in order to tune the proposal distribution suitably. For the scope of this dissertation, we will only introduce the adaptive proposal by Haario et al. (1999), which has been considered for the application presented in Chapter 4, and point the interested reader to Laine et al. (2008) and Brooks et al. (2011) for further details.

Adaptive Metropolis. Specification of the adaptive proposal in the Metropolis-Hastings algorithm comes with a slight modification of Algorithm 2. In particular, the candidate density $q(\cdot)$ may depends on the whole history (or a part of it) of the sampled chain. Hence, we refer to this adaptive proposal as $q_m(\theta^* \mid \theta^{(1)}, \dots, \theta^{(m)})$. It is now crucial to establish *how* the proposal depends on the history. For this purpose, Haario et al. (1999) noted that the sampled set at iteration m can be written as $\{\theta^{(1)}, \dots, \theta^{(m-H+1)}, \dots, \theta^{(m)}\}$, where H is a fixed integer representing the *memory* parameter. The proposal distribution q_m is chosen as follows:

$$q_m(\theta^* \mid \theta^{(1)}, \dots, \theta^{(m)}) \sim \mathcal{N}\left(\theta^{(m)},\, s_k^2 \tilde{\Sigma}^{(m)}\right) \tag{2.7}$$

where $\tilde{\Sigma}^{(m)}$ is the $k \times k$ covariance matrix determined by the H points $\theta^{(m-H+1)}, \dots, \theta^{(m)}$ and the scaling factor s_k depends only on the dimension of the vector parameters k. In particular, the choice of the scaling parameter s_k can either be heuristic or match the theoretically optimal mixing properties (Gelman et al., 1996; Roberts and Rosenthal, 2009). Here, we note that the stochastic chain obtained using Equation (2.7) is not Markovian anymore, however Haario et al. (2001) proved that, under mild conditions, it is still ergodic. Being effective and easy to implement, the inclusion of an adaptive step partially solves the issue related to convergence (poor mixing) of the chains. However, there are still several problems to deal with, among which autocorrelation, the choice of the starting values, the choice of the number of the chains, etc.

Summing up, even the *ideal* MCMC technique may poorly explore the support of the posterior distribution, especially in high-dimensional spaces, yielding to very imprecise estimators regardless of the tuning. In the worst-case scenario, it is not guaranteed that we could even attain the center of the target distribution in the finite computational time at our disposal, and the resulting MCMC estimators will be highly biased. Consequently, the direct application of MCMC machinery to hierarchical models must be considered carefully, as their scalability to high-dimensional settings (i.e., a large number of parameters) of practical interest is not trivial and often requires advanced solutions. The main reason why the *guess-and-check* strategy of Metropolis-based samplers is doomed to fail in high-dimensional spaces is that the number of guesses increases exponentially, lowering the chances of these guesses being accepted. The basic idea to overcome such a problem and possibly attain unexplored regions of the target distribution is to exploit information

about the *geometry* of the target distribution itself. Hamiltonian[4] Monte Carlo, firstly proposed in Duane et al. (1987), is the unique procedure for automatically generating this coherent exploration for sufficiently well-behaved target distributions and it will be discussed in Section 2.2.3.

The complex set of MCMC techniques just described becomes easy when we learn how to use several ready-made libraries of programs for their implementation. The most common are WinBUGS (Lunn et al., 2000), BUGS (Spiegelhalter et al., 1996), JAGS (Plummer, 2003), and the very recent NIMBLE (de Valpine et al., 2017). All these are connected to `R` through specific packages: `R2WinBUGS` (Sturtz et al., 2005), `R2jags` (Su et al., 2015) and `nimble`.

2.2.3 Hamiltonian Monte Carlo

As for its alternatives, described in Section 2.2.1 and 2.2.2, Hamiltonian Monte Carlo (HMC) was firstly introduced in physics by Alder and Wainwright (1959) to describe the molecules' motion (*Hamiltonian dynamics*) following Newton's laws. Subsequently, in the seminal paper by Duane et al. (1987), HMC was applied to lattice field theory simulations of quantum chromodynamics, and the authors referred to it as "Hybrid Monte Carlo." Not much later, the first statistical applications of HMC were proposed for neural networks and regression models (Neal, 2012; Ishwaran, 1999; Schmidt, 2009). However, its spread among researchers and practitioners began immediately after the release of `Stan` (Carpenter et al., 2017). This probabilistic programming language provides full Bayesian inference for continuous-variable models through (but not only) HMC sampling.

The basic ingredients required for the HMC are the k-dimensional vector of parameters θ, which is also referred to as *position* vector, and an auxiliary *momentum* variable, r, having the same dimension of θ. Altogether, they constitute an augmented $2k$-dimensional parameter space, also known as *phase space*. Combining the phase space with Hamilton's equations, able to describe the conservative dynamics in physical systems, it is possible to establish the conceptual framework HMC stemmed from.

In the following paragraph, HMC is introduced following the conceptual overview given by Betancourt (2017) and some technicalities provided by Neal et al. (2011) and Betancourt and Girolami (2015).

Hamilton's equations In the context of Hamiltonian dynamics, a physical system can be described by a function of θ and r, known as the *Hamiltonian*, generally referred to as $H(\theta, r)$. The partial derivatives of H determine how θ and r change over time t, according to Hamilton's equations:

$$\begin{cases} \dfrac{\partial \theta_i}{\partial t} = \dfrac{\partial H}{\partial r_i} \\ \dfrac{\partial r_i}{\partial t} = -\dfrac{\partial H}{\partial \theta_i} \end{cases} \quad , \qquad i = 1, \ldots, k. \tag{2.8}$$

For any time interval of duration t', these equations define a mapping from the state at any time t to the state at time $t + t'$. In particular, these dynamics corresponding to such mapping uphold to theoretical properties crucial for the validity of MCMC updates (e.g. *reversibility*, *conservation*, *volume preservation* and *symplectiness*). For the sake of brevity, the discussion of such points is overlooked in this work, complete technical details are in Neal et al. (2011).

[4] a.k.a. *hybrid*.

The Hamiltonian function can always be decomposed in the sum of two independent terms, as follows:

$$H(\theta, r) = U(\theta) + K(r), \tag{2.9}$$

where $U(\theta)$ is called the *potential energy*, and will be defined to be proportional to minus the log probability density of the distribution for θ that we wish to sample; while $K(r)$ is called the *kinetic energy*, and is usually defined to be $r^\top \Sigma^{-1} r/2$. Here, Σ indicates a symmetric, positive-definite matrix and yields a $K(r)$ proportional to minus the log probability density of the zero-mean Gaussian distribution with covariance matrix Σ.

MCMC from Hamiltonian dynamics The conjunction between MCMC and Hamiltonian dynamics involves the concept of *canonical distribution*, borrowed from statistical mechanics. More specifically, the canonical distribution can be seen as the quantity linking the potential energy function to the distribution we wish to sample from, representing the probability distribution over all the possible states (θ, r). It has the generic probability density function:

$$\pi(\theta, r) = \frac{1}{Z} \exp(-H(\theta, r)/T), \tag{2.10}$$

where T is the temperature of the system, and Z is a normalizing constant. If the Hamiltonian is defined as Equation (2.9), than we can rewrite Equation (2.10) as

$$\pi(\theta, r) = \frac{1}{Z} \exp(-U(\theta)/T) \exp(-K(r)/T). \tag{2.11}$$

Without loss of generality, by setting $T = 1$ and inverting Equation (2.10) with respect to $H(\theta, r)$, we obtain:

$$H(\theta, r) = -\log(\pi(\theta, r)) - \log(Z).$$

By looking at this equivalence, it is easy to see that if we want to sample from

$$\pi(\theta \,|\, y) \propto \pi(\theta, r) = \pi(\theta) f(y \,|\, \theta),$$

we can express the posterior distribution as a canonical distribution, where the potential energy is set to be minus the log-posterior, i.e.:

$$U(\theta) = -\log\left[\pi(\theta) f(y \,|\, \theta)\right]$$

Leapfrog integrator Solution to Equation (2.8) is seldom derived analytically. Hence numerical approximation is required. For implementation purposes, Equation (2.8) must be accurately approximated through discretization of time, using a small stepsize ϵ. The general idea is starting at t_0 and iteratively compute the approximate position-momentum state at times ϵ, 2ϵ, 3ϵ, ..., for a reasonable amount of stepsizes[5]. Among the set of powerful methods able to approximate the solution of a system of differential equations, excellent results can be obtained with the *leapfrog integrator*. Although valid for any specification of the kinetic energy function, we here assume $K(r) = r^\top \Sigma^{-1} r/2$ is assumed, with $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_k)$ for simplicity of notation. The leapfrog integrator consists of the following steps, which are executed repeatedly for a predetermined number of times L:

[5]Theoretical justification for the optimal integration time is discussed in Betancourt (2016b).

$$
\begin{aligned}
r_i(t+\epsilon/2) &= r_i(t) - (\epsilon/2)\frac{\partial U}{\partial \theta_i}(\theta(t)) \\
\theta_i(t+\epsilon) &= \theta_i(t) + \epsilon \frac{r_i(t+\epsilon/2)}{\sigma_i} \\
r_i(t+\epsilon) &= r_i(t+\epsilon/2) - (\epsilon/2)\frac{\partial U}{\partial \theta_i}(\theta(t+\epsilon)).
\end{aligned} \tag{2.12}
$$

The leapfrog integrator guarantees volume preservation, is time-reversible, and provides stable solutions as long as $\epsilon < 2$ (Neal et al., 2011). Its approximation error, and hence HMC performances, strongly depend on choosing suitable values for ϵ and L[6]. Indeed, if ϵ is too large, the simulation will be inaccurate, yielding low acceptance rates. On the other hand, if ϵ is too small, computational effort will be wasted taking many small steps. At the same time, if L is too large, HMC will generate trajectories retracing their steps. Even worse, if L is chosen so that the parameters jump from one side of the space to the other at each iteration, the Markov chain may not even be ergodic (Neal et al., 2011). Eventually, if L is too small, subsequent samples will be close to each other, resulting in undesirable random walk behavior and slow mixing. In light of these considerations, tuning these parameters for any specific problem requires some expertise and usually one or more preliminary runs. Moreover, there is not a simple objective metric for establishing whether a trajectory is "correct" or not. Therefore, it is common practice to rely, for example, on heuristics based on autocorrelation statistics from preliminary runs.

Hamiltonian Monte Carlo sampling scheme The sampling algorithm generating the Markov chain via HMC consists of two main steps, which are repeated iteratively. The first step only affects the momentum; namely, new values are randomly drawn from their Gaussian distribution (recall that $K(r) = r^\top \Sigma^{-1} r/2$ is assumed), independently of the current values of the position variables at each iteration. The second step may change both position and momentum. Indeed, a Metropolis update is performed, using Hamiltonian dynamics to propose a new state. Starting with the current state (θ, r), Hamiltonian dynamics are simulated for L steps using the leapfrog method (or some other reversible, volume-preserving method), with a stepsize of ϵ. The momentum variables at the end of this L-step trajectory are then negated, giving a proposed state (θ^*, r^*). This proposed state is accepted as the next state of the Markov chain with probability

$$
\rho = \min\{1, \exp\left(-H(\theta^*, r^*) + H(\theta, r)\right)\}. \tag{2.13}
$$

Following Hoffman et al. (2014), the standard implementation of HMC is presented in Algorithm 3.

If the proposed state is rejected, the next state is the same as the current state. The negation of the momentum variables at the end of the trajectory makes the Metropolis proposal symmetrical, as needed for the acceptance probability above to be valid. In practice, this negation is not really necessary since $K(r) = K(-r)$. It is now clear the "hybrid" nature of HMC in the sense that the simulation is done alternating the joint update of θ and r via Hamiltonian dynamics and updating r via Gibbs sampling. The real advantage of HMC is that Hamiltonian dynamics for (θ, r) can produce a value for θ with a much different probability density (equivalently, a much different potential energy, $U(\theta)$), leading to a more efficient exploration of the target distribution.

[6]Implications and limiting cases are comprehensively discussed in Leimkuhler and Reich (2004).

Algorithm 3: Hamiltonian Monte Carlo sampling scheme.

0: Initialization: define an arbitrary set of starting values $\theta^{(0)}$, ϵ, L, $\mathcal{L}$ and M

for $m = 1, \ldots, M$ **do**

1: Draw $r^{(0)} \sim N(0, \Sigma)$

2: Set $\theta^{(m)} \leftarrow \theta^{(m-1)}$, $\tilde{\theta} \leftarrow \theta^{(m-1)}$, $\tilde{r} \leftarrow r^{(0)}$

for $i = 1, \ldots, L$ **do**

3: Set $\tilde{\theta}$, $\tilde{r} \leftarrow \text{Leapfrog}(\tilde{\theta}, \tilde{r}, \epsilon)$

end

4: Accept $\theta^{(m)} \leftarrow \tilde{\theta}$ and $r^{(m)} \leftarrow -\tilde{r}$ with probability ρ

end

where

$$\rho = \min\{1, \exp\left(-H(\theta^*, r^*) + H(\theta, r)\right)\}$$

The algorithms introduced in this chapter are the driving engines behind the studies which will be presented in the sequel. Their direct and indirect implementation constitutes the bridge that binds each problem to each solution, always winking to the computational efficiency.

Chapter 3

Measuring and modeling food losses

The United Nations (UN) defined **sustainable development** as "*development that meets the needs of the present without compromising the ability of future generations to meet their own needs*". In this respect, one aspect is to promote a "**Responsible Consumption and Production**" (SDG 12). The third Target under this goal (Target 12.3) states "*By 2030, [to] halve per capita global food waste at the retail and consumer levels and reduce food losses along production and supply chains, including post-harvest losses*". The Food and Agricultural Organization of the United Nations (UNFAO) and UN Environment are the custodians of SDG 12.3 and, for consistency with policy objectives, relevance, and measurement, the indicator has been split into two distinct sub-indicators that focus on losses on the supply side and waste on the demand and consumption side of the food systems, respectively. My work only focuses on the first sub-indicator, namely the one on losses.

In particular, this chapter includes the joint work with Prof. Giovanna Jona Lasinio and Dr. Carola Fabi, Statistician at FAO, with whom I had the pleasure to collaborate during the internship that led to my Master's thesis project, now published by Mingione et al. (2021b). During the internship, we developed the basic methodology for the *Food Loss Index* (FLI), which is now officially implemented in FAO Statistical Working System. Then, soon after my Master's thesis discussion, we carried out a parallel project which aimed at providing an alternative standardized protocol for the estimation of yearly food losses at the country-commodity level to compile the FLI and explain what drives food losses dynamics at the global level. A major challenge was the lack of data, which dictated many methodology decisions. Therefore, the objective of the work was to present a possible improvement to the modeling approach used by FAO in estimating the annual percentage (over production) of food losses by country and commodity. Our proposal combines robust statistical techniques with the strict adherence to the rules of the official statistics, and focuses on cereal crops, which currently have the highest (yet incomplete) data coverage allowing for more ambitious modeling choices. The estimation work is twofold: it aims at selecting the most important factors explaining losses worldwide, comparing two Bayesian model selection approaches, and at predicting losses with a Beta regression model in a fully Bayesian framework.

We point out that we did not envision a model for immediate use in the production of official data, but rather run an academic study. Several enhancements have been made by FAO in the data collection effort and the food loss database is now extremely different from the one used in Mingione et al. (2021b). The latter uses data which are now outdated and unofficial: therefore, results do not reflect FAO views or policies. However, in accordance with FAO Statistics Division, we tried our model on the official estimates that are now available in FAOSTAT and obtained comparable results. In particular, 99% of the official data fall into our prediction intervals (see Figure A.5 for an example).

3.1 Introduction

The *2030 Agenda for Sustainable Development* (Cf, 2015), approved by all the Member States of the United Nations (UN) in September 2015, officially came into force on 1st January 2016. The Agenda includes 17 *Sustainable Development Goals* (SDGs) and 169 *Targets* supported by a global monitoring framework with 231 *Indicators*, which were established to track progress. Some of these goals are groundbreaking: indeed, for many of them, there was no specific indicator, methodology, nor underlying data for the measuring. Food losses reduction fell into this category. More precisely, the indicator to measure and monitor food losses, associated with Target 12.3, was initially classified in Tier III, meaning that an indicator and data collection method needed to be developed on purpose. One significant challenge is the lack of reliable estimates of the level of losses (and waste) worldwide, particularly in developing countries, for numerous reasons (Fabi et al., 2018). Preliminary work indicates that food losses and waste remain unacceptably high, impacting economic efficiency and natural resource usage and contributing to inefficient food systems. The widely quoted advocacy study "*The Global Food Loss and Waste – extent, causes, and prevention*" (Gustavsson et al., 2011), published by the Food and Agricultural Organization of the United Nations (FAO), estimates that yearly global food loss and waste account for 30% of the overall production, which is equivalent to almost USD 1 trillion. Recently, model-based estimates in the *State of Food and Agricultural* (SOFA) Report (FAO, 2019) confirmed that food losses on the supply side alone (after harvest and up to but excluding the retail level) are equal to almost 14% of agriculture production and are worth at least USD 400 billion in 2016.

Literature on the measurement and estimation of global food losses

Food loss measuring and monitoring is not a novel issue among experts in both the private and public sectors. The UN General Assembly addressed the problem back in 1975 and passed a resolution calling for "a 50 percent reduction of post-harvest losses by 1985". In 1976, FAO identified the major constraints causing post-harvest losses focusing on staple crops, including grains and pulses. Two years later, the FAO produced an action program that led to developing a standard terminology and suitable methodology for the measurement of post-harvest losses formalized in the milestone publication, "*Postharvest Grain Loss Assessment Methods*" (Harris, K. L. and Lindblad, C. J. , 1978). Some methods and techniques explained in the manual were later revised by Boxall, R.A (1986) over the period 1980-1986, in an attempt to simplify them. Additionally, in 1980 FAO published the guidelines for the "*Assessment and Collection of Data on Post-Harvest Food-Grain Losses*" (FAO, 1980) to support the implementation of a statistical methodology combining objective measurements with statistical survey sampling techniques to collect data and produce accurate survey-based post-harvest loss estimates. Many other studies followed these efforts in modeling and estimating losses. The most relevant and recent ones that have received the highest level of worldwide consensus are: "*The African Post-harvest Losses Information System*" (APHLIS), which developed a calculator to estimate cumulative post-harvest losses over the entire value chain, as a percentage of production for nine cereals in Sub-Saharan African countries (SSA); "*The Global Food Loss and Waste – extent, cause, and prevention*" report (Gustavsson et al., 2011), that changed the world perception over food loss and waste and which uses a mass balance approach to quantify the volumes of food loss and waste at the global level; "*Imputation of Loss Ratios*", a technical report by an FAO consultant, Klaus

Grünberger (2013, unpublished), who developed an econometric model to estimate loss using causal factors and covariates such as countries infrastructure, national income level, geographic region, and commodity groups. The causal factors were not significant, hence the model was abandoned.

All these efforts have been hindered by little available data, which reflects the low priority given to post-harvest losses until recently and to the objective complexity and cost of food loss data collection. These constraints persist and affect the quality of the estimates and, consequently, the reliability of results. The dire lack of data, an international definition of food losses and a recognized methodology to monitor loss reduction underpinned the need to develop a standardized approach for measuring, collecting data, and modeling food losses. A comprehensive methodology including a measurable definition, an indicator, an aggregation method, an estimation model and a range of data collection methods and tools has been developed by FAO to help countries measure food losses and monitor progress against SDG target 12.3 (FAO, 2019).

The scope of this work is to present an improved model capable of estimating food losses at the country-commodity level. The new model considers a set of explanatory variables that scientific literature has consistently identified as the causes or proxies of causes of losses in all countries of the world. The purpose of using explanatory variables is to link losses with their causal factors at the country-commodity level to support decisions on interventions, investments, and policy-making. Our model's main feature is that it builds on the finding of previous efforts and works toward overcoming their weaknesses.

Definitions

In recent years, Food Loss and Waste (FLW) became a priority issue on the global agenda, for both the public and private sector, as one aspect of sustainable global food systems. In the absence of a commonly agreed definition, the various stakeholders have developed their definitions of food loss and waste, albeit a pre-condition for a harmonized methodological approach and data comparability is to agree on the terminology. For this reason FAO, under the aegis of the Save Food initiative, undertook the development of a "*FLW Definitional Framework*" in consultation with national and international stakeholders building on the previous definitions found in the literature and laying the foundation for a consistent methodology. In what follows, we will only report the most important definitions required for a proper comprehension of this work. For further details, we highly recommend to take a look at the whole document (FAO, 2014).

In particular, the main definitions include:

- **Food Supply Chain (FSC)**: the connected series of activities to produce, process, distribute, and consume food;
- **Food loss**: the decrease in quantity or quality of food.

For the sake of measurability and consistency with the SDG 12.3 target formulation, an operational definition of "Food Loss" was added to the *Definitional Framework* in 2016 (unpublished) drawn from FAO's annual questionnaire on agriculture production whereby:

- **Food Losses** are crop and livestock product losses that cover all quantities along the supply chain for all utilizations (food, feed, seed, industrial, other) up to, but not including, the retail/consumption level. Losses of the commodity

as a whole (including edible and non-edible parts) and losses direct or indirect, which occur during storage, transportation, and processing, also of relevant imported quantities, are therefore all included;

- **Food Waste** occurs from retail to the final consumption/demand stages.

3.1.1 The Food Loss Index

At the global level, the *Global Food Loss Index* (GFLI) is a composite indicator, built as a weighted average of countries' Food Loss Indices (FLIs). A country FLI is a fixed-base index that aggregates the losses of 10 key commodities in the five main food groups using economic weights (value of production in the base period). FAO is partnering with national and international stakeholders to foster data collection along the supply chains and build the evidence base for these commodities. Although the FLI uses aggregated percentage losses along the supply chain, more disaggregated data at different stages of the value chain (e.g., farm, transportation, storage, processing, and wholesale) is needed to decide on appropriate interventions. The countries' FLIs summarise complexities of food loss and their dynamics to provide decision-makers with an overview of the magnitude of the problem at the national level and an overall monitoring indicator.

The food loss dataset can be treated as a longitudinal dataset for a multivariate outcome across different countries. In the ensuing sections, we will refer to the observed (or estimated) loss percentage for country i, commodity j at year t as l_{ijt}. Therefore, the FLI for country i in year t is defined as:

$$FLI_{it} = \frac{\sum_j (l_{ijt} \cdot q_{ijt_0} \cdot p_{jt_0})}{\sum_j (l_{ijt_0} \cdot q_{ijt_0} \cdot p_{jt_0})} \cdot 100, \tag{3.1}$$

where t_0 is the reference year; q_{ijt_0} the production quantities by country and commodity in the reference year, available in FAO's corporate statistical database (FAOSTAT); p_{jt_0} the fixed price (in USD) set by commodity for the $(t_0 - 1)$-$(t_0 + 1)$ average. At present, the reference year is set to 2015 (the year in which countries adopted the SDGs), while l_{ijt} can be either survey-based or model-based. The food loss percentages at the commodity or country level can be interpreted as the average percentage of supply that does not reach the retail stage. The FLI shows the relative change in percentage food loss for country i over time, compared to the base year. Finally, using weights proportional to the total value of agricultural production in the base year, the country indices can be aggregated to build the Global Food Loss Index (GFLI). To achieve SDG 12.3, both GFLI and FLIs should ideally show that post-harvest losses decrease compared to the base period from a base value of 100.

Basic data constraints

Primary data on losses are seldom compiled within the national statistical systems worldwide: only 39 countries out of 185 reported losses for one product or more in FAO's annual Questionnaires on Agricultural Production, including a section on product utilization. Moreover, reporting on losses has increased slightly in recent years and data was even more scarce in the past period. Data on utilization, including losses, stock changes, and food supply, is used to compile the Supply Utilization Accounts (SUA) and Food Balance Sheets (FBS). The FBS framework defines agricultural production net of harvest losses and collects loss estimates net of harvest losses. Noteworthy, only 7% of loss data in FAOSTAT's FBS domain (FAOSTAT, 2016) was officially reported by the countries in the period 1990-2016.

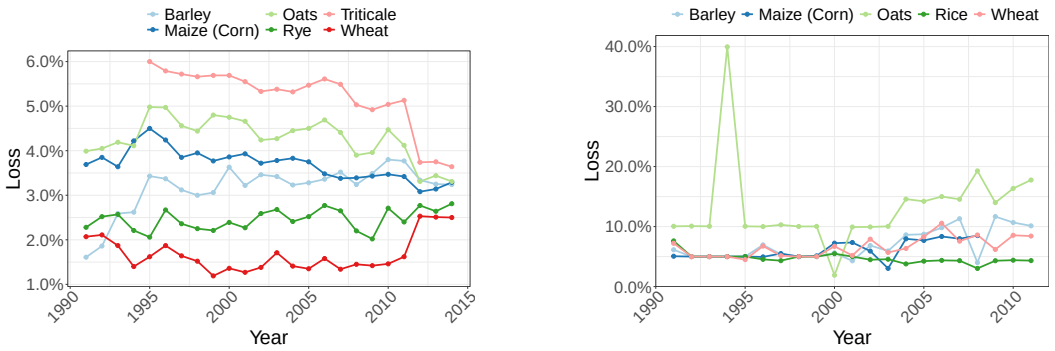
The remaining 93% of records are estimated or considered null. In conclusion, since representative data on losses are very scarce, the FLI will be model-based. However, with the strong emphasis put on SDG 12.3 and the need for evidence-based policy-making, one has to expect an increase in data availability in the future years. The methodology provided herein attempts to refine further the 2016-18 FAO developed model, described in the next paragraph.

3.1.2 FAO modeling approach: SOFA 2019

FAO developed a random effects model able to exploit panel data information, i.e., in a cross-section – by commodity and country – and longitudinally over time to estimate missing loss data and compile the FLI of all countries (FAO, 2019). The model is part of the methodology for monitoring progress against SDG 12.3[1]. Results were first published in the State of Food and Agriculture 2019 edition (SOFA 2019) and stated that global food losses along the supply chain, up to but excluding the retail level, are almost 14% of 2016 total production. At present, FAO can disseminate loss estimates at the global, regional, and commodity group level. The model supplements the 7% officially reported loss data along the supply chain with two additional data sets. The first one is a dataset of food losses built from a literature review to increase the coverage. The second one is a dataset composed of over 200 possible explanatory variables from various international sources (International Energy Association, World Bank, FAO, and more), possibly representing causal factors or proxy variables for the causes of losses. These causal factors can be grouped under common categories to be easily managed by a model. These categories are welcome . The random forest algorithm was used to standardize variables' selection and choose the 5 most important ones by commodity grouping. The purpose was to capture better the variation in the causes of losses by country or region and commodity. Where the observations by country and commodity are fewer than three, a bare minimum to run the model for a country-commodity combination, available information has been clustered by commodity group on the assumption that causes and rates of losses are more similar within the groups than across them (for example losses of maize and lentils are more similar than losses of maize and fresh milk). The same assumption applies to the value chain (e.g., traditional, capital-intensive, vertically integrated, and more) and solutions (improved farm practices, infrastructures, cool chain, and others). Clustering scarce data evened out the impact of outliers on the results. The coexistence of country-level estimates and cluster-level estimates required a model hierarchy to fill in the results matrix. All the methodological choices have been dictated by the need to overcome data scarcity.

The rest of the paper is structured as follows: Section 3.1.3 describes the available data and some preliminary results on data consistency; Section 3.2 delves into the methodology. In Section 3.3, we report and analyze the model results. Section 3.4 is dedicated to the concluding remarks and discussion.

[1] see SDG indicators metadata at `https://unstats.un.org/sdgs/metadata/`

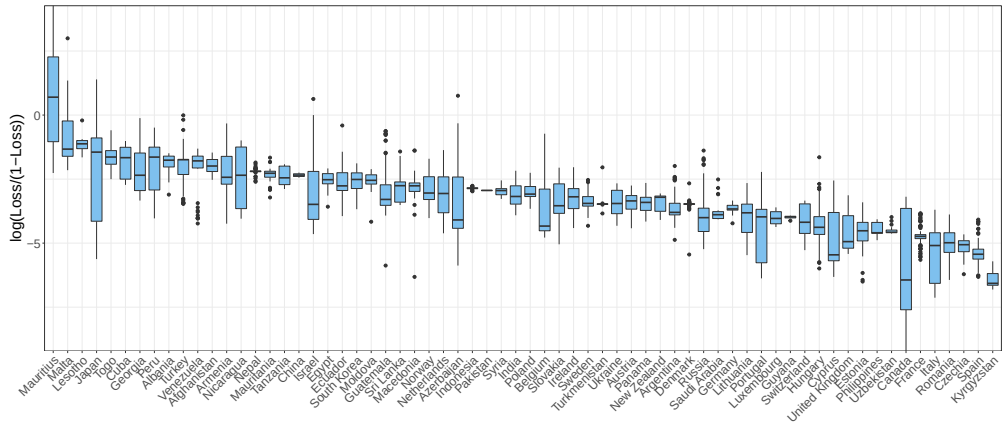


(a) (b)

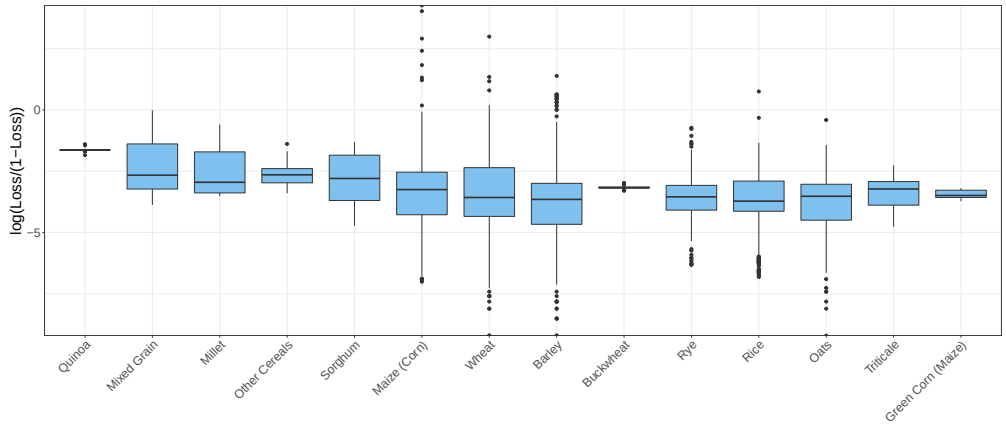
Figure 3.1. Available time series of cereals in Austria (left panel) and Ecuador (right panel).

3.1.3 Available data

We worked with official loss data extracted from FAOSTAT (2016). The dataset covered 138 Countries and 145 different commodities, with the most extended time series starting in 1961 and ending in 2015, and included a total number of 18,472 records. A preliminary exploratory analysis of the data highlighted some critical issues. First, some country-commodity combinations presented loss levels larger than their total production: these losses characterizes several import-dependent countries where domestic supply consists mainly of imported produce. The FLI methodology deals with import-dependent countries by changing the denominator in the ratio. In this work, we did not introduce any exception and therefore we excluded these records from the analysis. Second, 4264 of country-commodity-year combination records were equal to 0. Zero losses on such a large scale are unlikely and instead point at under-coverage or missing data interpreted as nil amounts. Moreover, the comparison of FAOSTAT data to loss factors found in the literature showed a systematic difference. The SUA seems to underestimate the actual losses within the countries and the explanation is manifold (Fabi et al., 2018). Case studies in the literature tend to focus on countries where losses are high, and the problem is more acute, representing an upper boundary. On the contrary, nation-wide estimates average losses across all value chains, including the more efficient ones. Also, losses are sometimes obtained as the balance for quantities that cannot be accounted for in the SUA. Therefore, SUA data can be considered the lower boundary. Indeed, FAOSTAT data showed a global loss average of 7.2% over the whole dataset that is 9.4% when excluding zero values. Another data constraint and challenge to the modeling framework is that countries tend to use carry-forwards estimates on loss percentages, on the ground that systemic losses do not change quickly over time, but at the same time removing any trend from the time series (see Figures 3.1a and 3.1b). Additional information was gathered from more than 300 publications and reports from various sources to increase observations and reduce the noise in the data. These sources included reports from international organizations, such as the World Bank, GIZ (Gesellschaft fur Internationale Zusammenarbeit), FAO, IFPRI (International Food Policy Research Institute), sub-national reports, and academic institutions (Fabi et al., 2018). All data have been consolidated in a database that is continuously updated and accessible at `http://www.fao.org/platform-food-loss-waste/flw-data/en/`.



(a)



(b)

Figure 3.2. Distribution of loss percentages (on the logit scale) by country (a) and crop (b).

Cereals and cereal products This food category includes the largest share of available loss data. Hence, we focused our modeling efforts on cereals.

More precisely, cereals data include 66 countries and 14 different commodities, amounting to 196 country-crop combinations ($< 66 \times 14$ as not all countries produce all cereals), with the longest time-series going from 1991 to 2014. A simple average over the available data gives a loss percentage of 5.6%, but there is variability both by country and commodity (for further details see Figures 3.2a and 3.2b). Figure 3.3 represents a clear snapshot of the data availability: each square identifies a country-crop combination, and it is colored according to the country-crop temporal average.

The majority of estimates are provided by countries of Northern America and Europe (NAE). In particular, 111 out of the 196 country-crop combinations (more than 50%) come from NAE, whose average loss percentage is the lowest (only 3.24%), as reported in Table 3.1. Sub-Saharan Africa (SSA) records the highest loss percentages, with losses amounting to 24% of total production, but only five SSA countries reported data on losses. This unbalanced data distribution does not introduce any bias in our methodology because the estimation is carried out

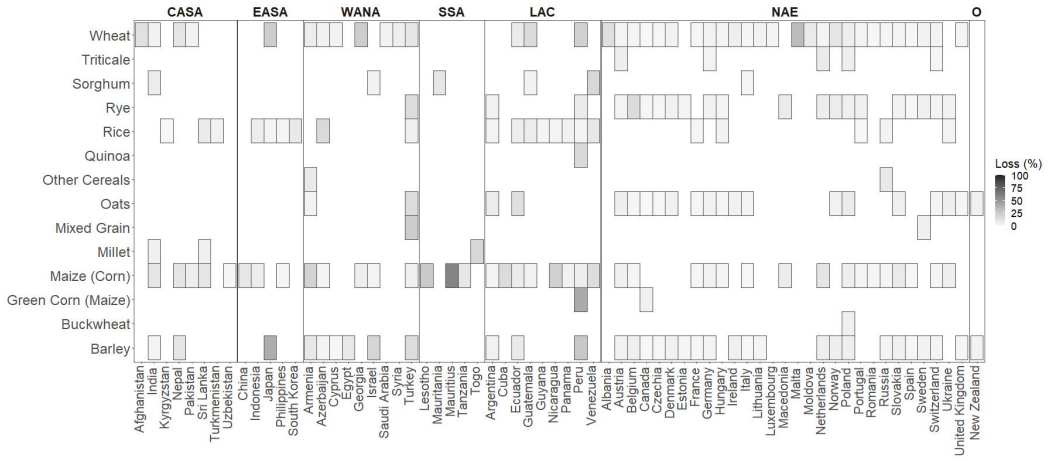


Figure 3.3. Available data for cereals. Each square represents the temporal average loss by country and commodity. Countries have been ordered on the x-axis with respect to the SDG region they belong Central Asia and Southern Asia (CASA), Eastern Asia and South-eastern Asia (EASA), Western Asia and Northern Africa (WANA), Sub-Saharan Africa (SSA), Latin America and the Caribbean (LAC), North America and Europe (NAE), Oceania (O).

at the country-crop level. However, some bias may be introduced when losses are aggregated at the global level. In this case, the weights used to calculate the GFLI should be proportional to each country's agricultural sector size. Nevertheless, the representativeness of single countries or single macro-regions can differ significantly. We do not want to compile the global losses by estimating missing country data exclusively from countries within the same region (e.g., losses in Asian countries estimated only losses of the Asian countries). Few available countries will determine the estimates too heavily. If the few informants are not representative of the region, the regional and ultimately the global estimates will underestimate/overestimate the actual loss level.

	CASA	EASA	WANA	SSA	LAC	NAE	O
Loss (%)	5.86	10.53	8.58	24.09	9.67	3.24	3.27
nObs	16	9	24	5	29	111	2

Table 3.1. Average loss (%) and number of distinct country-crop combination by SDG region: Central Asia and Southern Asia (CASA), Eastern Asia and South-eastern Asia (EASA), Western Asia and Northern Africa (WANA), Sub-Saharan Africa (SSA), Latin America and Caribbean (LAC), North America and Europe (NAE), Oceania (O).

3.2 Methodology

This section will describe our proposal. We will first define the steps in our statistical protocol and then delve into the model in detail. Our model will estimate losses by country, commodity, and year for cereals in a full Bayesian framework, so as to reduce the impact of critical issues in the data described in Section 3.1.3. After dealing with missing data in the available predictors, we build a Beta regression model with a latent component (Wu, 2009). The latent component captures variations at

the country-crop level due to missing information on other known causal factors such as the time of harvest, rainfall on crop areas, and other variables that should be measured at crop level.

Missing predictors imputation

A total number of $K = 34$ explanatory variables were considered in the loss estimation model. Most variables are proxies for relevant explanatory factors commonly found in the scientific literature. Similar to the loss imputation model of the SDG methodology, these factors can be grouped into categories relating to *Energy*, *Economic Factors*, *Transportation and Logistics*, *Building Materials* and *Weather and Crop Cycles* (see Table A.1). However, not all the variables are available for all countries and years, highlighting a severe missing data issue as the *missingness* in the set of predictors is almost 19%. More in detail, 16 out of the 34 variables contain at least one missing value and 13 out of these 16 ones have more than 30% of missing values overall. Assuming a MAR mechanism, we consider three non-parametric missing value imputation methods: the `missForest` algorithm (Stekhoven and Bühlmann, 2012), Multiple Imputation by Chained Equations (MICE) approach (White et al., 2011) and k-nearest neighbours (K-NN) as in Franzin et al. (2016). With each imputed dataset, we estimate the model in Equation (3.5) and compare results in terms of variable selection and prediction accuracy. We do not report all the details, but we simply note that the three imputation methods produce similar outputs in terms of final model performances. We decided to keep the imputed dataset with `missForest` for its flexibility with respect to assumptions on data collection and distribution. Besides, in the seminal paper, Stekhoven and Bühlmann (2012) show that `missForest` generally outperforms the two other imputation methods. Also, as demonstrated empirically not only with our set of data (Waljee et al., 2013; Cihan, 2018), the `missForest` algorithm yields better results, especially in terms of out of sample prediction error. This happens because of its non-parametric nature, which allows for the imputation of mixed-type data. Being based on a random forest algorithm (Breiman, 2001), it has no need for tuning parameters nor does it requires any assumptions about the distributional aspects of the data. Eventually, it offers a way to assess the quality of an imputation without the need of setting aside test data nor performing cross-validations. In particular, the full potential of `missForest` is deployed when the data include complex interactions or non-linear relations between variables of unequal scales, as it is in our case study.

Beta regression

Food losses are expressed as percentage of the total production, hence the Beta distribution is the most natural assumption for their modeling. Indeed, the class of Beta regression models, firstly proposed by Ferrari and Cribari-Neto (2004), is commonly used to model random variables that take values in the open standard unit interval $(0, 1)$. The main assumption is that the dependent variable is Beta-distributed and its mean $\mu \in (0, 1)$ is related to a set of regressors through a linear predictor with unknown coefficients and a link function $g : (0, 1) \to \mathbb{R}$, strictly increasing and twice differentiable. The model also includes a precision parameter $\phi > 0$ (independent from μ), which may be constant or may depend on a set of predictors through a link function as well. This approach has the advantage naturally incorporating features such as heteroskedasticity or skewness which are commonly observed in data taking values in the standard unit interval.

We assume that our outcome variable $y_i, \ldots, y_n$ is the realization of a random sample such that $y_i \sim Beta(\mu_i, \phi)$, $i = 1, \ldots, n$. Then, the Beta regression model is defined as $g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$; where $\boldsymbol{\beta}$ is the $k \times 1$ vector of unknown parameters and $\mathbf{x}_i$ is the vector of k predictors. The `logit` function represents the most common choice as link function for $g(\cdot)$, due to its shape and ease of interpretation, such as in any typical *Generalized Linear Model* (GLM) framework.

3.2.1 Bayesian variable selection

Given the pretty high dimensionality of the problem, which counts 34 predictors, our first objective is to find a robust selection method for the most relevant factors that can explain losses' behaviour. The goal is to find the subset of variables that can simultaneously catch the dependencies and dynamics driving food losses, but that are also meaningful for policy making. This issue is of paramount importance and raises several challenges. For example, a known problem when the number of relevant variables is large, is to account for possible collinearity in order to avoid conflicting results when assessing the importance of strongly correlated predictors (Ijarchelo et al., 2016).

In a Bayesian perspective, variable selection falls in the more general framework of *model choice* and can be addressed with various possible approaches. Two main classes of methods can be identified: discrete mixtures (Mitchell and Beauchamp, 1988; George and McCulloch, 1993) and shrinkage priors (Tibshirani, 1996). Methods belonging to the class of discrete mixtures model prior knowledge on coefficients βs with a prior comprising both a point mass at zero and an absolutely continuous alternative; on the contrary, methods belonging to the class of shrinkage prior model βs prior distribution with absolutely continuous *shrinkage priors* centered at zero.

We notice that discrete mixtures offer the correct representation of sparse problems by placing positive prior probability on the event $\beta_k = 0$, but pose several difficulties. These include foundational issues related to the specification of priors for trans-dimensional model comparison, and computational issues related both to the calculation of marginal likelihoods and to the rapid combinatorial growth of the solution set. Shrinkage priors, on the other hand, can be very attractive computationally. But they create their own set of challenges, since the posterior probability mass on $\beta_k = 0$ (a set of Lebesgue measure zero) is never positive. Truly sparse solutions can therefore be achieved only through artifice. In general, the choice of one approach or the other involves a series of trade-offs. However, although the latter is computationally tractable and seems to outperform its competitors in a variety of applications, the discrete-mixture approach represents a methodological ideal.

It is under this premises that, in this work, we tested and compared the two aforementioned estimation alternatives, choosing the most popular statistical technique for each one of them: the spike and slab technique (within the class of discrete mixtures) in the formulation by Mitchell and Beauchamp (1988), and the horseshoe prior (within the class of shrinkage priors), introduced by Carvalho et al. (2010).

Spike and slab

Spike and slab is considered as the gold standard to combine variable selection with the estimation of the regression parameters. With this technique, variables are chosen by estimating the *posterior* probability of all the models within the considered class (O'Hara et al., 2009), based on the *a priori* knowledge or expectation that only few variables truly impact on the outcome.

The main assumption is that the prior distribution of the k-th regression parameter is a mixture of two components: a probability mass either exactly at or around zero (spike) and a flat distribution (slab) elsewhere. Therefore, this prior is often written as:

$$\beta_k|\gamma_k, c, \epsilon \sim \gamma_k \cdot N(0, c^2) + (1 - \gamma_k) \cdot N(0, \epsilon^2), \tag{3.2}$$

where $\epsilon \ll c$ and where $\gamma_k \in \{0, 1\}$, denoting absence or presence of the k-th variable in the model. If ϵ is set to 0, then the spike is taken to be a Dirac δ_0 at the origin.

In Kuo and Mallick (1998), γ_k is embedded in the regression equation as follows:

$$y_i = \sum_{k=1}^{K} \beta_k \gamma_k x_{ik} + \epsilon_i. \tag{3.3}$$

Independent priors are typically assumed for β_k, γ_k and the response variance. In particular, $\gamma_k \sim Ber(p_k)$, namely a Bernoulli distribution with success probability p_k that reflects the preference for including the k-th predictor in the model building: e.g., $p_k = 0.5, \forall k$ is associated to prior belief of the equally likely relevance of all possible 2^k sub-models.

Once the model has been set up, it is usually fitted using Markov Chain Monte Carlo (MCMC) and the variable selection part of the model entails estimating γ_k. As a result, the posterior inclusion probability $\mathbb{E}[\gamma_k|\mathbf{y}]$ can easily be calculated as the mean value of the indicator γ_k as follows:

$$\mathbb{E}[\gamma_k|\mathbf{y}] = P(\gamma_k = 1|\mathbf{y}) = \frac{1}{T} \sum_{t=1}^{T} \gamma_k^{(t)},$$

where T denotes the total number of posterior samples. The selection rule consists merely of keeping those variables with posterior inclusion probability larger than a given threshold. If the threshold is set at 0.5, then the selection criterion is known as the *Median Probability Model* (MPM) by Barbieri et al. (2004). This criterion is known to be robust as it is the optimal predictive model under a squared error loss function with certain regularity conditions and the selected variables appear in at least half of the visited models (Barbieri et al., 2004). The orthogonality of the design matrix is required in all the sub-model scenarios to satisfy these conditions. If this is not the case, inference based on marginal inclusion probability could be incorrect.

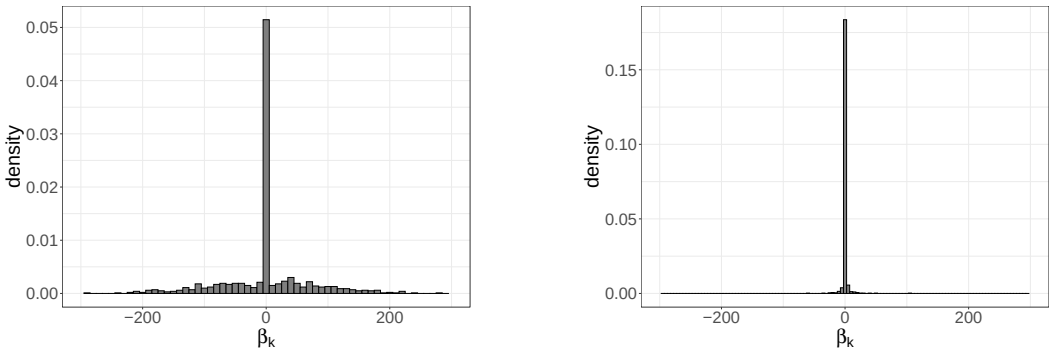
Horseshoe prior

This approach assumes that each coefficient β_k is *a priori* distributed as a scale mixtures of Normal distributions:

$$\begin{aligned} &\beta_k|\lambda_k, \eta \sim N(0, \lambda_k^2 \eta^2), \\ &\lambda_k \sim C_+(0, 1), \qquad \eta \sim C_+(0, 1) \end{aligned} \tag{3.4}$$

where $C_+(0, 1)$ represents the half-Cauchy distribution, λ_k is called *local* shrinkage parameter and η is the *global* shrinkage parameter (Carvalho et al., 2009). The horseshoe is named after the shape of the so-called *shrinkage coefficient*, which is $\frac{1}{(1+\lambda_k^2)} \sim Beta(\frac{1}{2}, \frac{1}{2})$ and can be interpreted as the posterior amount of weight that the posterior mean of β_k places on 0. Horseshoe prior's main advantage lies in its flat tails allowing for strong signals to remain large *a posteriori* and in its infinitely tall spike at the origin that severely shrink the β_ks that are very likely to be zero. It can

be easily noticed that setting $\epsilon = 0$ in Equation (3.2), generates a prior distribution very close to Equation (3.4) that allows for only two values, i.e. 0 and 1, instead of assigning continuous priors to γ_k, as in the case of the horseshoe. For the sake of clarity, Figure 3.4 shows the distribution of the prior on β_k in the case of the spike and slab (3.4a) and the horseshoe (3.4b).



(a) **(b)**

Figure 3.4. Prior distribution of the k-th regression coefficient in the case of spike and slab prior (a) and horseshoe (b).

3.2.2 Model proposal

Our two alternative proposals combine the procedure described in Section 3.2 with the ones described in Section 3.2.1. Let l_{ijt} be the observed loss percentage in country i, for cereal j at year t and x_{ikt} be the value of the k-th explanatory variable for country i at time t. The model with variable selection is expressed as follows:

$$\begin{aligned} & l_{ijt} \sim Beta(\mu_{ijt}, \phi), \quad \forall i, j, t \\ & \text{logit}(\mu_{ijt}) = \log\left(\frac{\mu_{ijt}}{1-\mu_{ijt}}\right) = \beta_i t + \sum_k x_{ikt}\beta_k^* + \nu_{ij}, \quad \forall i, j, t \\ & \nu_{ij} \sim \mathcal{N}(0, \tau^2), \qquad \phi \sim \mathcal{U}(5, 150), \qquad \tau \sim \mathcal{G}(4, 0.1), \end{aligned} \tag{3.5}$$

where $\beta_i t$ is a temporal linear trend specific to country i and ν_{ij} is the latent component describing the nested *commodity within country* effect. We consider different trend parameters for each country to capture different general behaviors dictated by country-specific policies or climate conditions, or other unobserved factors. The temporal trend was always included to detect generally well- or bad-performing countries in terms of the FLI. Parameter ϕ is known as the precision parameter, since for a given μ_{ijt}, a larger ϕ implies a smaller variance for l_{ijt}. We also adopted a constant precision τ across countries and commodities after estimating several models with different precision parameters (e.g., country-specific, crop-specific, or their sum) that did not yield significantly different estimates. Finally, according to the prior distribution ascribed to the regression coefficients β_k^*, we can obtain either the spike and slab model (Equation 3.3) or the horseshoe model (Equation 3.4). While hyperparameters for the shrinkage priors represent standard statistical choices commonly used in the Bayesian variable selection procedures, hyperparameters for ϕ and τ are set to obtain weekly informative priors. In the spike and slab model, we set $\gamma_k \sim Ber(p_k)$, adding a further level to the model by treating p_k with a Beta(5,5) so that all the models were equally likely to be selected *a priori*. The prior distribution on the trend coefficients β_i is $\mathcal{N}(0, 1000)$.

WAIC

We compared the performances of the two variable selection procedures using the *Watanabe-Akaike Information Criterion* (WAIC) proposed by Watanabe (2010). The main assumption, which should hold in our model setting, is that the observed values are conditionally independent given the parameters. If the model fits our data, then parameter estimation should minimize the expected log-pointwise predictive density. More precisely, let $lpd = \sum_{i=1}^{n} log \int p(y_i|y)p(\boldsymbol{\theta}|y)d\boldsymbol{\theta}$ is the log-pointwise predictive density and $p = \sum_{i=1}^{n} V_{post}[log(p(y_i|\boldsymbol{\theta}))]$ is the estimated *effective number of parameters*, i.e. the sum of the posterior variance (V_{post}) of the log-predictive density for each data point. Following Vehtari et al. (2017) the expected log-pointwise predictive density is given by $elpd = lpd - p$. The WAIC is then obtained as $WAIC = -2 \cdot elpd$. We use the WAIC instead of the *Deviance Information Criterion* (DIC) for two reasons: (i) WAIC has the desirable property of averaging over the posterior distribution rather than conditioning on a point estimate, which is particularly relevant in a predictive context (Gelman et al., 2014); and (ii) the DIC has a weaker theoretical justification (Celeux et al., 2006; Spiegelhalter et al., 2014). Furthermore, since the final objective is to predict food losses, the WAIC is more appropriate choice it is asymptotically equivalent to the Bayesian leave-one-out cross-validation (Watanabe, 2010) and hence it can be seen as a measure of a model's predictive performance.

3.3 Application

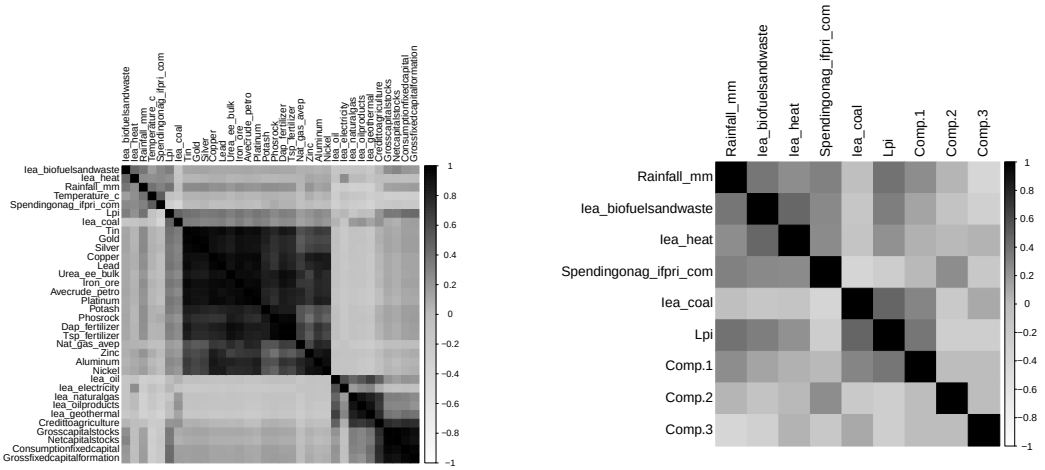
In this section, we present the results of our modeling effort. We will start with results of data pre-processing, including dimensionality reduction; we report the results of the variable selection afterwards, and finally show the out-of-sample predictions. Note that all the variables were standardized before the quantitative analysis.

We notice that the estimation of the posterior inclusion probabilities in the spike and slab framework is not reliable in the presence of severe collinearity. In particular, following Bhadra et al. (2019a), optimality can be achieved in terms of parameters' estimation if the design matrix is well-conditioned (e.g., orthogonal). The design matrix orthogonality ensures that no information is shared among the predictors, while collinearity has the effect of blurring distinctions between predictors in the variable selection process.

To this purpose, we first computed the correlation matrix (Figure 3.5a). Three different groups of strongly correlated variables can easily be pointed out: the first one (the big black square at the center of Figure 3.5a) includes all metals' prices (e.g., potash, silver, iron, gold, lead, etc.); the second one includes the prices of electricity, natural gas oil and derived products provided by the International Energy Agency (IEA); the third group (bottom right corner) includes all the economic variables from national accounts (such as net capital stocks) and credit to agriculture. We carried out a preliminary dimensionality reduction on the predictors using a simple principal component analysis (PCA) on 27 out of the 34 standardized variables in the three groups leaving out the seven variables (i.e., rainfall (mm), temperature (C), biofuels, heat, coal, LPI, spending on agriculture) in the top left square of Figure 3.5a, as they are not highly correlated with the others or among them. Results show that three components can explain 81% of the total variance. The first (principal) component can be interpreted as a proxy for input prices with metal prices (iron, silver, copper, etc.) for implements and infrastructure, and fertilizers' prices (potash, urea, etc.)

Figure 3.6. Variables associated to the first principal component, with loadings greater than 0.2 in absolute value. The size of each variable is proportional to its loading.

for growing crops (see Figure 3.6). The second component can be interpreted as a proxy for investment in agriculture (capital stocks, credit to agriculture, and more), but with negative loadings (see Figure A.1a). In other words, lower values of this component correspond to higher values of the considered variables. The third component is a proxy for energy's price (oil, natural gas, electricity, and more, see Table A.1b. The final dataset includes 10 almost completely uncorrelated variables (3 components + 7 standardized variables), whose correlation matrix is shown in Figure 3.5b.



(a) **(b)**

Figure 3.5. Graphical representation of the correlation matrices: (a) before dimensionality reduction of 27 variables out of 34, (b) after dimensionality reduction.

Estimation

The estimation was carried out using JAGS (Plummer, 2003), a well-known software for Bayesian model estimation, which uses Gibbs Sampler and the Metropolis Within Gibbs Sampler algorithms. For each variable selection approach, we ran the MCMC algorithm with two chains, $120,000$ iterations, a burn-in of $60,000$ iterations, and a thinning of 10, keeping $6,000$ samples from each chain for inferential purposes. Coding examples for both the estimation and prediction of the model with the different prior settings are available in Algorithms A.1 and A.2. From now on, we will refer to the model with spike and slab priors as **M1** and to the model with horseshoe prior as **M2**.

The WAIC is equal to $-19,510.6$ for M1 and $-19,522.8$ for M2, meaning that the two selection approaches are substantially equivalent, with M2 performing slightly better in terms of goodness of fit.

Recall that the spike and slab procedure allows for estimating the posterior inclusion probability for each predictor. We chose the *Median Probability Model* as the selection rule, hence we kept all the variables with posterior inclusion probability larger than 0.5.

The horseshoe priors do not provide a straightforward variable selection technique, hence we decided to keep all variables with coefficients whose 95% posterior credible intervals did not include the zero value. In other words, the whole point for shrinking priors is to shrink to zero coefficients that are not significant, according to the classical likelihood definition. Figure 3.7 illustrates the outcome of the selection step. Both techniques select four variables, i.e., biofuels (price), spending on agriculture (i.e., the agriculture share in GDP, which is a proxy for the agricultural sector relative importance in the national economies), the second principal component (comp.2), and the third principal component (comp.3). Both the computed point estimates and the credible intervals are comparable. In particular, according to M1, biofuels, spending on agriculture, and comp.2 are included in the model with probability equal to 1, and comp.3 is included with probability equal to 0.56[2]. The temperature is selected only by M2, while according to M1, its posterior inclusion probability $\hat{\gamma}_k$ is equal to $0.0075 \simeq 0$.

Biofuels has the most considerable effect (in absolute value) on the outcome with a positive sign. This variable, measured by the IEA, represents solid biofuels, liquid biofuels, and biogases produced with industrial and municipal waste. Biofuels are a possible utilization of cereals, both the full grains and its waste or discarded quantities. In this respect, a commissioned study by FAO (Kuiper and Cui, 2020) found that reducing food losses could decrease agricultural prices, which would benefit the production of meat and biofuels, with lower agricultural input prices. This study can help explaining the correlation between biofuels price and losses. An increase in biofuels price would increase the demand for input crops and absorb larger amounts of cereals for industrial uses, thus reducing the quantities ultimately lost. Unfortunately, biofuel data are often based on small sample surveys or other incomplete information. Thus, the data give only a broad overview of the biofuel sector and are not strictly comparable across countries (IEA, 2019). Spending on agriculture has the second-largest effect and with a negative sign, while the coefficient associated with comp.2, which we recall is a proxy for investment, is positive. However, the component has negative loadings on the original variables, which means that the higher the investments (or capital intensity), the lower the losses; it also means that investing more in agriculture would reduce losses. Comp.3 gives a small positive contribution, both for M1 and M2, which is consistent with

[2] see Table A.2 for the exact numerical results.

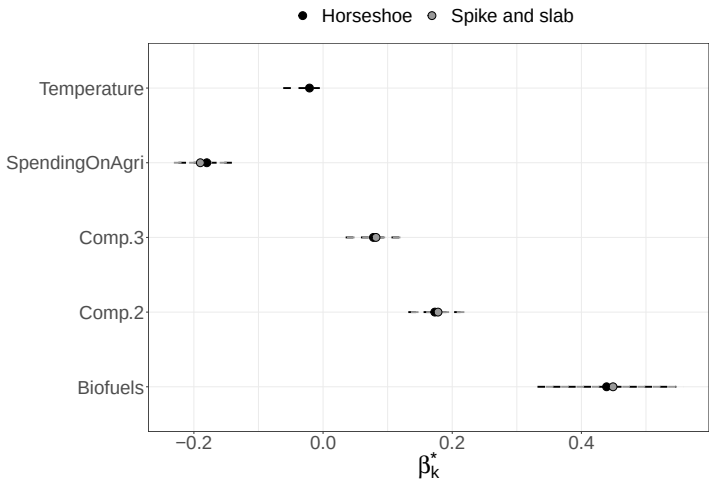


Figure 3.7. Point estimates and 95% credibility intervals for β_k^* associated to the selected variables by M1 (Spike and slab) and M2 (Horseshoe).

the other energy-related variable (i.e., biofuels). Eventually, the temperature effect, selected only in M2, is the smallest. Moreover, the temperature is not measured at the crop level and is a simple yearly average. Therefore, we cannot interpret the relation as a direct effect of annual temperature on losses, but rather as the combined phenomenon by which countries with higher average temperature tend to experience smaller losses (at least in our sample).

Recall that in Equation (3.5), we set a different time trend for each country. M1 estimates 24 countries out of 66 with statistically significant trends: 9 of them show an increasing trend and 15 of them show a decreasing trend. M2 instead identifies 30 countries with significant trends (22 of which are the same as in M1), with 10 countries showing an increase and 20 a decrease in the food loss percentages over time. The largest estimated random effect by both M1 and M2 belongs to Malta for wheat and is equal to 1.07 on the logit scale, which corresponds to 70% percentage losses. However, the final loss factors, including the covariates' effect, are around 18%, close to the observed value in FAOSTAT. We would like to point out that Malta is an island country that imports around 90% of its wheat consumption. Loss percentages in import-dependent countries should be calculated based on domestic supply to include imports and correct extreme results. The import-dependency has been dealt within the FLI methodology but was overlooked in this work because it was not relevant in this research context.

At the opposite hand of the scale, M1 estimates the smallest effect for oats in Armenia, while M2 does so for maize (corn) in Cuba. The point estimates are -5.45 and -5.58 on the logit scale, which correspond to 0.43% and 0.37% of loss percentage, respectively (see Table A.2). In Cuba's case, the final estimates range between approximately 15% (M1) and 25% (M2), on a similar level to the country's reported losses.

The estimated values for the variance of the random effects τ^2 and the variance of the outcome ϕ are $\simeq 0.1$ and $\simeq 71$ (thus a dispersion $\phi^{-1} \simeq 0.014$), respectively, both for M1 and M2, meaning that the estimates are precise. Finally, the estimated value for the global shrinkage parameter η^2 is equal to 0.01[3].

[3]For η^2; we use the Maximum A Posteriori (MAP) estimator since its posterior distribution is not symmetric. For details on variance's parameters, see Figures A.2 and A.3

Validation

To evaluate our models' predictive performance, we split the sample into training and test sets. The test set includes 953 data-points (i.e. 25% of the whole sample) and was built by removing the last five observations from the time series of each country-crop combination when the time series length was larger than eight years, only two observations were set aside for prediction purposes otherwise. We used the *Relative Mean Squared Error* (RMSE) to measure the difference between predicted and observed values. The RMSE is computed as the ratio between each model's prediction error (at the numerator) and the error that would have resulted by using the simple predictor (e.g. sample average). It can be computed as:

$$\frac{\sum_{i=1}^{n_{te}}(l_i^{te} - \hat{l}_i)^2}{\sum_{i=1}^{n_{te}}(l_i^{te} - \bar{l}_{tr})^2}, \tag{3.6}$$

where l_i^{te} are the observed losses in the test set, $\hat{l}_i$ are the predicted losses and $\bar{l}_{tr}$ is the sample average of observed losses in the training set. Predictions are obtained using the selected variables in M1 and M2. For each model, we ran two chains with $80,000$ iterations each, applied a burn-in of $40,000$, and a thinning of 10, then kept $4,000$ samples from each chain for inference. The overall RMSE is 0.359 and 0.358 for M1 and M2, respectively, confirming again that the two approaches are equivalent. Figure 3.8 shows the observed losses in the test set and their predicted values. Perfect predictions would lie on the dashed red line (or the identity line $y = x$). Both models show a good performance, especially for losses smaller than $\simeq 20\%$ (the majority in the sample). Besides, the average coverage of the prediction intervals for both M1 and M2 is greater than 90%. In particular, it is equal to 92.34% for M1, while it is equal to 92.55% for M2.

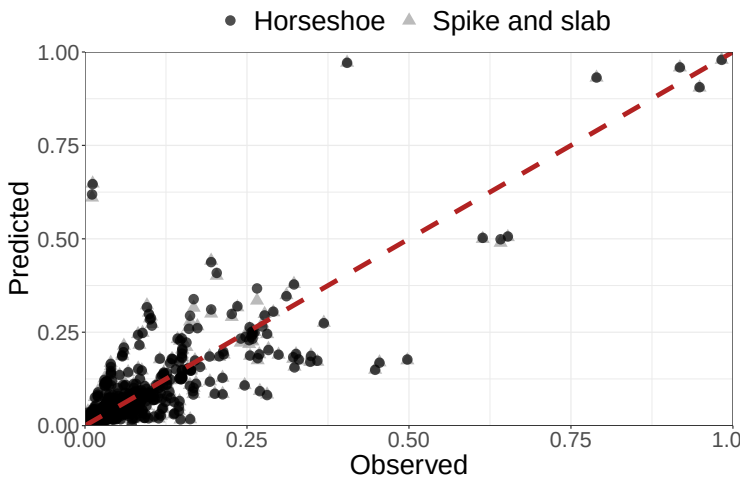
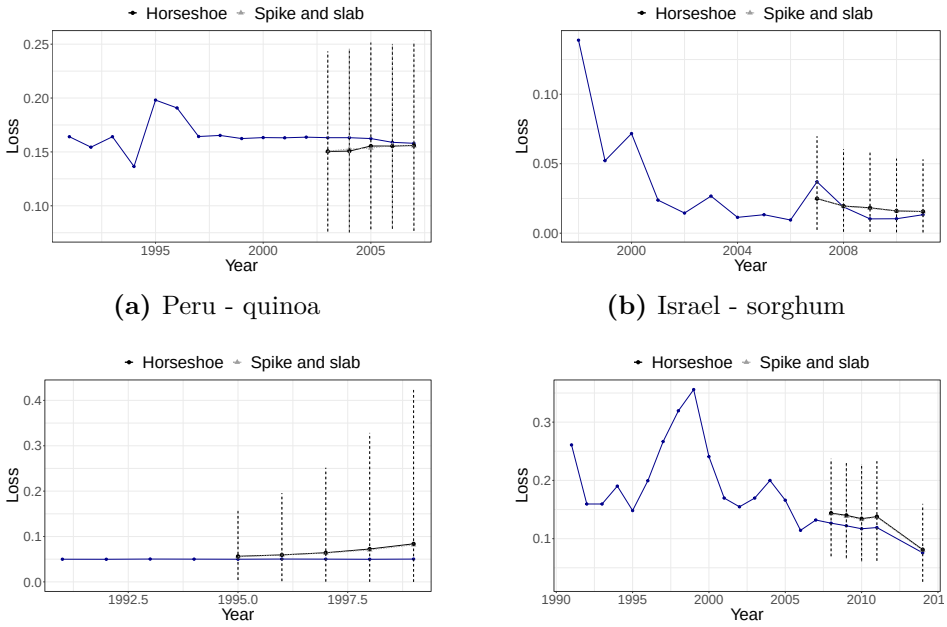


Figure 3.8. Observed vs predicted losses by the two sub-models. The dashed red line represents the identity line $y = x$.

M1 and M2 have comparable predictive performances when the error is evaluated separately by country and commodity, although the RMSE is not uniformly distributed across countries or across commodities. In particular, Pakistan and quinoa are the country and the commodity with the highest RMSE, respectively. For Pakistan, we only have loss data for one commodity (maize), with an approximately

flat time series at about 5%, as shown by the blue line in Figure 3.9c. The point predictions produced by our models struggle to reproduce the flat trend in this country, suggesting some weird behavior of one of the predictors (the spending on agriculture halved over the period) or an issue with the target variable itself; nevertheless, observed values fall into the 95% prediction intervals for both models. Quinoa losses are only observed for Peru, as reported in Figure 3.9a. The models underestimate losses for this commodity; however, also, in this case, observed values fall into the 95% prediction intervals. Good prediction performances are also shown in Figure 3.9b and 3.9d: both M1 and M2 catch the *plateau* and the decreasing trend in the observed time series. For these two country-crop combinations, the prediction error is about 0.00005 for Israel-sorghum and 0.0003 for Togo-millet. Notice that the error for Israel-sorghum would have been equal to 0.0036 had we used the sorghum mean for the prediction, or 0.0058 had we predicted losses with Israel mean; for Togo-millet, the error would have been equal to 0.00037 had we used the millet mean, or equal to 0.008 had we used with Togo mean.



(a) Peru - quinoa

(b) Israel - sorghum

(c) Pakistan - maize (corn)

(d) Togo - millet

Figure 3.9. Observed time series and predicted values by the two sub-models with 95% predictive intervals for four different country - crop combinations.

3.4 Discussion and further developments

In this work, we present a substantial improvement over the previous global food losses modeling efforts. First, the proposed distributional framework is highly coherent with the nature of the data: since food losses are expressed as proportions, the Beta distribution represents a much more appropriate choice for describing their behaviour. Second, the proposed approach is very flexible: the model in Equation (3.5) could be applied to other food groups when data will be available. Third, the hierarchical modeling structure covers most food loss dynamics and is easily scalable when needed (e.g., to estimate losses at supply chain step).

The two proposed variable selection techniques provided equivalent results. Although being computationally more demanding, the spike and slab priors allow for the computation of posterior inclusion probabilities for all the variables, providing a rigorous and straightforward way towards variables' choice. On the other hand, the horseshoe priors require the choice of *a posteriori* selection criteria, but they are considerably less demanding in terms of computational effort. Hence horseshoe priors should be preferred when the latter poses a serious issue.

One *caveat* on the use of this model is that it is more demanding in terms of the number of observations than the hierarchical mixed-effect model developed for the SDG 12.3 methodology and the SOFA 2019 report. Our model could be developed and tested for cereals only, which account for the largest share of available data. It is not a viable option to date to compile the FLI, which needs to cover all five commodity groups, eventually with very few observations.

We would like to further remark that quality and reliability issues affect both the explaining variables and the outcome in our case study. We dealt with these issues using a Bayesian approach, which allows for the modeling of parameters' uncertainty at the prior level, but the correction of such values is not within the scope of this work and would require additional investigation. In this regard, we are aware that, in general, multiple imputation should be used, as suggested in Sinharay et al. (2001). Indeed, single imputation techniques usually underestimate the imputation process's uncertainty, and the imputed data may display a smaller variance. To handle this issue, we can think of building a model that includes a measurement error term for the imputation step. However, the increase in computational complexity does not seem to justify this solution. For these reasons, as also suggested by an anonymous reviewer, we believe that the imputation of missing data in such context could become in itself a good method paper, hence we leave this for future developments.

We also expect, when a larger amount of data will be available, to obtain the same results if we estimate the model using a maximum likelihood approach. In this work, we decided to test only Bayesian techniques because they allow to perform probabilistic uncertainty quantification in the model choice process unlike, for example, with a lasso regression. Furthermore, the lasso's optimality (theoretical properties) is only guaranteed in the framework of standard linear regression (e.g., Gaussian outcome). There is a very interesting paper by Groll et al. (2019) in which the authors propose a lasso-type penalization for generalized additive models, but in the discussion, they state that "the number of true parameters is partly overestimated". An extensive comparison between the lasso and the horseshoe is given in Bhadra et al. (2019a). Here, the authors argue that even though the lasso estimation procedure is typically computationally faster, the horseshoe prior performs better in terms of estimation thanks to its heavy tails, making it adaptive to sparse data and robust to large signals. Moreover, Polson and Scott (2010), Polson and Scott (2012) and Datta and Ghosh (2015) have shown that horseshoe empirically outperforms lasso in terms of out-of-sample predictive sum of squares errors. Last but not least, the lack of speed can be easily overcome, as proposed in Terenin et al. (2019) and Bhadra et al. (2019b).

Overall, all the proposed models produced promising results, in terms of (i) the explanatory variables that were selected; (ii) the possibility to use country-level estimates instead of clustered or global estimates; (iii) the estimated trends (see SOFA 2019 for comparison with this paper). More extended tests will be carried out when the data collection effort that should be undertaken by the national and international stakeholders to support policy-making towards the achievement of SDG 12.3 will produce significant improvements in data availability.

Chapter 4

Modeling physical activity using actigraph data

The methodology and the application presented in this chapter have been developed between September 2019 and March 2020, during my Visiting Research period at UCLA. The work has been carried out in collaboration with Ph.D. Pierfrancesco Alaimo Di Loro from "La Sapienza" and the research team of the Fielding School of Public Health of UCLA, under the supervision of Prof. Sudipto Banerjee. The motivation of the study concerned the advanced modeling of Actigraph data, which are usually analyzed using standard statistical techniques because of their huge sample size. Building upon recent developments in this field, we construct temporal processes using directed acyclic graphs (DAG) on the line of the *Nearest Neighbor Gaussian Process* (NNGP) (Datta et al., 2016a), account for spatial heterogeneity through penalized spline regression, and develop optimized implementations of the *collapsed* 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 behaviour. 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.

Abstract

Rapid technological developments in accelerometers have generated substantial interest in monitoring human activity. Wearable devices, such as wrist-worn sensors that monitor gross motor activity (actigraph units) continuously record the activity levels of a subject, producing massive amounts of high-resolution measurements. Analyzing actigraph 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 devise a Bayesian hierarchical modeling framework for spatial-temporal actigraphy data 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 levels of physical activity.

4.1 Introduction

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 (PA) 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 (PA). Given the well-established relationships between lack of PA and several obesity-related 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 PA and to establish public health programs that promote more PA[1].

Technologies for monitoring spatial energetics (James et al., 2016; Drewnowski et al., 2020) and promoting physical activity continue to emerge. *Actigraphy* broadly refers to the monitoring of human rest and activity cycles using wearable devices. Actigraphy data are gathered directly from wearable sensors or indirectly through smart-phone mobile applications and record repeated measurements at very high resolution. Accelerometers, in particular, are motion sensors that measure acceleration along different axes and are able to collect large amounts of data (Plasqui and Westerterp, 2007; Sikka et al., 2019). They are increasingly conspicuous because of their affordability, accuracy, and availability in smart-phones, smart-watches and other wearable devices. Many devices include Global Positioning System (GPS) sensors that reference measurements with location tracking along trajectories, or paths, traversed by the subject. Collected data can be quickly downloaded and promptly analyzed to obtain insights into their pattern and structure.

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. Analyzing such data is sought for several reasons: (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 any proposed new trajectory for a given set of health attributes. Researchers have cogently demonstrated the benefits of an active lifestyle over a sedentary one on physical and mental well-being and longevity (Lee et al., 1995). Therefore, actigraphy tracking is especially attractive as it allows for a better understanding of what behavioral and environmental factors influence population and individual health and, hence, aid in public health recommendation and policy.

Given that actigraphs generate large amounts of spatial-temporal data, it is natural to choose from the rich classes of such models (Cressie, 2015; Gelfand et al., 2019). However, actigraph data present some notable challenges (Kestens et al., 2017): they exhibit dependence along trajectories and must be accounted while predicting PA along arbitrary (unobserved) trajectories. We argue against a customary spatio-temporal process over $\mathbb{R}^2$ and disentangle spatial effects from dependence along trajectories. The balance of the paper is organized as follows. Section 4.2 introduces the PASTA-LA data set with insights into accelerometry data. The model for the temporal correlation is introduced in Section 4.3, while spatial

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

effects are discussed in Section 4.3.4. An extensive simulation study validating our model is proposed in Section 4.3.5. Data analysis, including model assessment and comparisons, are presented in Section 4.4. Finally, we conclude with a discussion in Section 4.5.

4.2 PASTA-LA project

4.2.1 Available data

Our dataset is compiled from the original **P**hysical **A**ctivity through **S**ustainable **T**ransport **A**pproaches in **L**os **A**ngeles (PASTA-LA) study conducted on a cohort of 460 individuals monitored between May 2017 and June 2018. 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+). We focus on the values recorded by Actigraph and GPS devices worn by 134 subjects for two one-week periods (one in 2017 and one in 2018). Study protocol for safeguarding participant information received necessary institutional review board (IRB) approval from the UCLA Human Research Protection Program. The data were stored on a secure computer and a redacted version was created for purposes of data sharing and research collaboration. While we do not pursue all of the aims of the PASTA-LA study, we build and test the framework in Section 4.3 for modeling high-frequency actigraph data related to different individuals.

Questionnaires The online questionnaires included two baseline and four follow-up surveys: one baseline and two follow-ups for each collection period of the actigraph and GPS data. Each survey consisted of responses pertaining to the participant's demographics and transportation habits. Not all participants completed all questionnaires. Hence, we considered only the surveys available for all the participants. This survey is the *first baseline questionnaire* and contains personal information such as sex, age, BMI, ethnicity and other socioeconomic factors. A user ID was 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 unit is an accelerometer roughly the same size and weight of the average wrist-watch that can be worn on the wrist, hip, and thigh and measure the directional acceleration at a pre-specified time frequency (generally 10 Hz to 30 Hz). The Actigraph GT3X+ model used for the PASTA-LA study (see Figure 4.1) can detect movement 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 analysis through a proprietary software. During download, the 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. The proprietary software precludes recovering the raw data once they have been processed in the download phase. Our study asked the 134 participants to keep the Actigraph unit on them (wrist) at all times other than during

Figure 4.1. Actigraph GT3X+ model used for the PASTA-LA study.

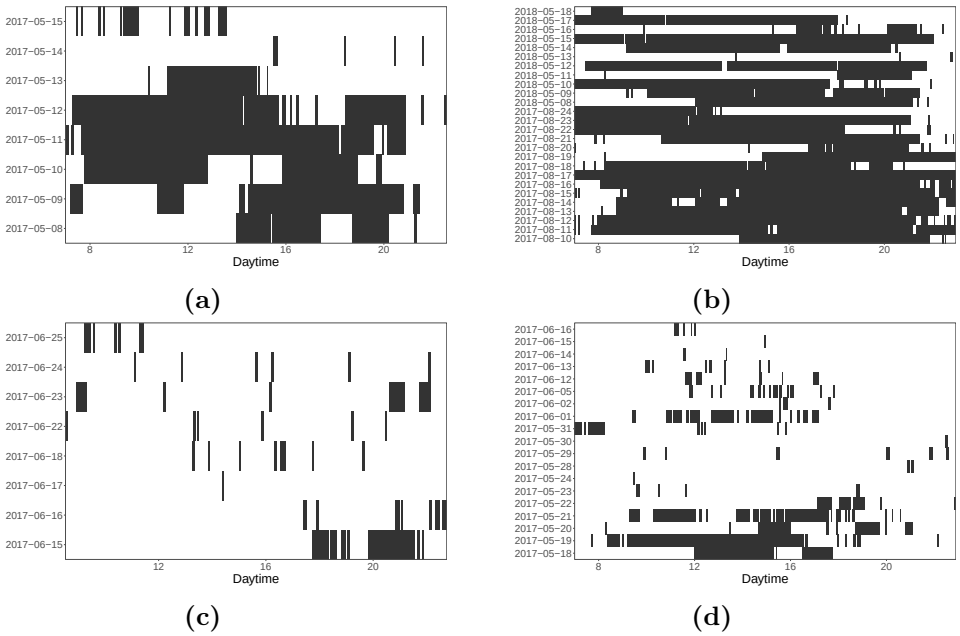


Figure 4.2. Derived missing data patterns for actigraph data during the daytime for 4 individuals arising from querying the inclinometer and eliminating low values: blank spots are missing data, black spots are observed values.

bathing and sleeping (awake time was assumed approximately from 7am to 11pm). Observations recorded outside this daily time-window were excluded.

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). 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). In the context of this paper, we settle on modeling solely the active hours of the individual, with our interest lying in detecting how individual or environmental covariates affect the physical activity level during the active time. During download the data were aggregated in sample epochs spanning 10 seconds. Measurements included the activity counts for the three axles and step counts. Time-stamps for final measurements (hour, minute, and second) were referenced by the mid-point between the beginning and the end of the epoch.

The Actigraph GT3X+ automatically records a measure of inclinometer values on how many of the 10-second epochs have been spent by the individual 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. Inclinometers[2] have reported sizable error rates up to 30% depending upon where they are worn and are 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. We checked that

[2]see https://actigraphcorp.com for details

large values of *inclinometer.off* corresponded to low (~ 0) values of activity in all the possible endpoints. Then, we dropped observations with *inclinometer.off* larger than $5s$ (i.e., the accelerometer was inactive for more than half of the epoch). This yields "missing data" from querying the inclinometer and eliminating low values. These resulted in $\approx 6.4 \times 10^6$ scattered observations out of the potential $\approx 12.3 \times 10^6$ ones[3], exhibiting missingness patterns as in Figure 4.2.

GPS The *Global Positioning System* (GPS) is a satellite-based radio-navigation system that does not require the user to transmit data and operates independently of any telephonic or internet reception. Any GPS unit can be set to record and store the spatial location at a pre-specified time frequency so that they could be downloaded and subsequently analyzed in a second moment. Obstacles, such as mountains and buildings, can block the relatively weak GPS signals and prevent the device from functioning accurately. In particular, the GPS *GlobalSat DG-500* was provided to the 134 subjects, which recorded the subject's location (latitude and longitude) every 15 seconds and comprised date and time of localization and speed in kilometers per hour (computed as distance over time through linear interpolation). In order to avoid a geographical imbalance that could bias and invalidate the model estimates, for the current analysis we restricted attention to subjects living and working in the Westwood neighborhood of Los Angeles. This helped excluding 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 aligned with the questionnaires' master-key to facilitate joining across all ID types (including email) while redacting and encrypting user data. The first baseline questionnaire, Actigraph and GPS were available for our group of 134 individuals. Henceforth, we refer to this specific group of units. We then build two different data sets:

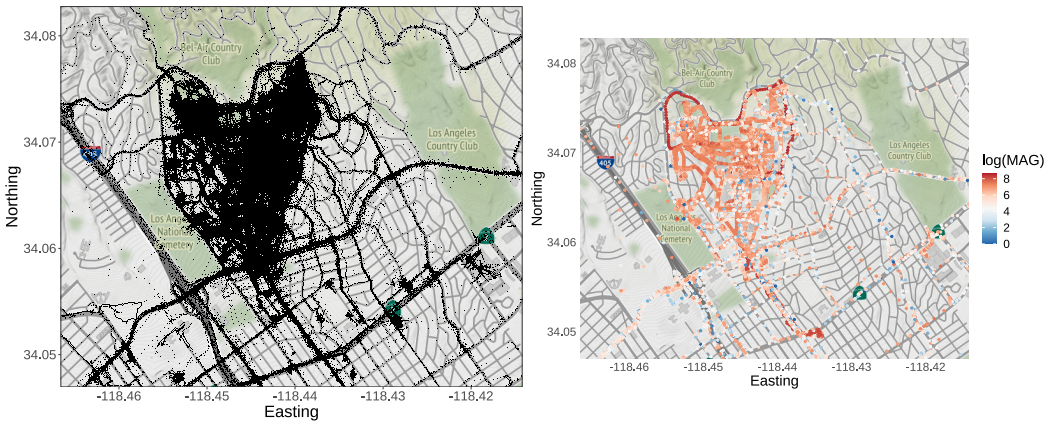
- The first dataset, D_1, comprising $N \simeq 5 \times 10^6$ measurements is obtained by joining the first baseline questionnaire with Actigraph data and includes the MAG (Section 4.2.2) at the different timestamps and all the individual predictors, but no spatial information.
- The second dataset, D_2, consists of $N \simeq 5 \times 10^5$ measurements (Figure 4.3) and is obtained by joining D_1 with GPS data. Actigraph and GPS data were temporally misaligned: the alignment was achieved by linear interpolation of the GPS locations on the temporal grid from the actigraph, keeping only those interpolated values with subsequent GPS measurements less than 30 seconds apart. This is a reasonable assumption, given that the individual's trajectory is well-approximated by a piece-wise linear GPS trajectory.

4.2.2 A measure of physical activity

Our primary endpoint of analysis is the magnitude of acceleration (MAG) defined as:

$$MAG_{kt} = \sqrt{x_{kt}^2 + y_{kt}^2 + z_{kt}^2}, \qquad k = 1, \ldots, K, \quad t = t_{k1}, \ldots, t_{kT}\ , \tag{4.1}$$

[3] computed as $n.days \times n.(epochs/day)$, where $n.(epochs/day) = 5400$



(a) **(b)**

Figure 4.3. (a): Observed locations over the Westwood area. (b): Observed physical activity level over the Westwood area on a subset of 10 individuals

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* (EE) measures (Freedson et al., 2012; Taraldsen et al., 2012). In particular, the *Metabolic Equivalent of Task* (MET) is currently the gold-standard measure of *rate of activity intensity* (Crouter et al., 2006; Hall et al., 2013; Lyden et al., 2014). When dealing with the filtered accelerometer outcomes (i.e. axes counts obtained after the transformation of the raw acceleration measurements by the proprietary software in the downloading process) the conversion into physical activity measurements and the corresponding accuracy vary by accelerometer model and brand, but mostly by the number of axis considered (Karantonis et al., 2006; Rothney et al., 2008). In particular, the introduction of tri-axial accelerometers raised questions about whether the vector magnitude of the three axes would provide more accurate assessment of physical activity intensity as compared to the magnitude of the vertical axis alone (Howe et al., 2009; Hänggi et al., 2013). In particular, Migueles et al. (2017) offer an extensive review of proposed accelerometer measurement cut-points and transformation into physical activity metrics, including single axis or MAG values.

More specifically, Sasaki et al. (2011), Santos-Lozano et al. (2013) and Kamada et al. (2016) investigated axis counts and vector magnitude resulting from the GT3X+ accelerometer in both controlled and free-living environments, while Aguilar-Farias et al. (2019) investigated the accuracy of relationships between MAG and MET in comparison to those based on the vertical axis counts only and validated the results with the EE and MET 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:

$$\text{MET}_{kt} = (0.000863 \times 6) \cdot MAG_{kt} + 0.668876, \tag{4.2}$$

and perform the same to the corresponding cut points for different PA intensity level (Table 4.1).

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, ∞)

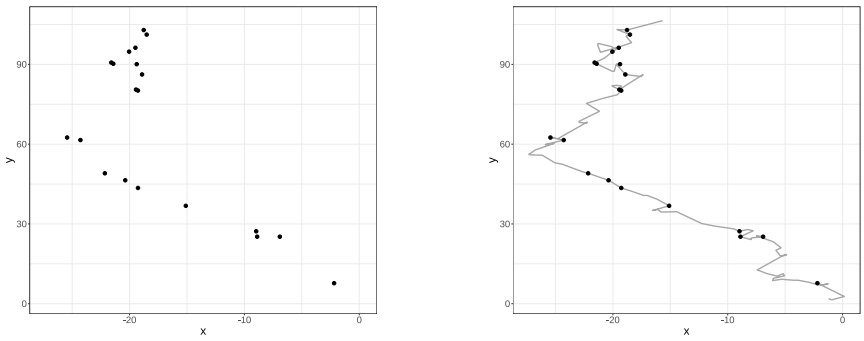
Table 4.1. MAG activity count cut-points for different PA intensity levels.

Based upon the aforementioned literature, the MAG is the outcome we fit and predict. We consider the closest outcome to the original source of information, without introducing additional and unnecessary noise or bias in the fitting process. However, for inferential purposes, MAG is later transformed into MET through Equation (4.2) 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).

4.3 Methodology

The outcomes corresponding to the K subjects are referenced with respect to the time at which they are recorded and the position in the trajectory. While it is tempting to work with a spatio-temporal process, dependence introduced by such processes may not be appropriate. An individual can visit the same location numerous times in his/her trajectory. These revisits need not occur at regular intervals and can be at distant time points. This suggests that proximity of two spatial locations in a trajectory need not result in strongly dependent MAGs recorded there. It appears more reasonable to model dependence among MAG measurements 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 4.4, the process actually belongs to a one-dimensional space, for which we define a proper distance measure $d(t_{ki}, t_{kj}) = \|\gamma_k(t_{kj}) - \gamma_k(t_{ki})\|$, where t_{ki} is the i-th recorded time point from 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 approximate such distances as the elapsed time between the two points $d(t_{ki}, t_{kj}) = |t_{kj} - t_{ki}|$, which would result in a good approximation of the spatial distance (especially if the subject is moving at constant speed). More generally, the elapsed separation across time will reflect dependence better than the spatial distance. The faster an individual is moving from one point to the other, the shorter the time elapsed, and higher the correlation between the two measurements. Hence, we model our measurements as $Y_k(\cdot) \equiv Z_k \circ \gamma_k(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ which, by construction, is a valid stochastic process.



(a) **(b)**

Figure 4.4. Example of observed points (a) and trajectory (b): black dots are realizations, grey line is domain of the process.

This will form the edifice of the model in Section 4.3.1, where we are modeling the dependence by solely considering stochastic evolution through time. How should spatial information be introduced in the model?

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 trajectory's temporal evolution. An added complication is that trajectories intersect and overlap and, in practice, can have multiple observations at the same location. Even more flexible spatio-temporal covariance kernels (e.g., non-separable or non-stationarity) will struggle to recognize the above features. Hence, we introduce the spatial effect in the mean using *spline regression* (see Section 4.3.4).

4.3.1 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. We model $\boldsymbol{Y}(\mathcal{T})$ as the finite realization of a K-variate process $\boldsymbol{Y}(\cdot)$ over $\mathbb{R}^+$:

$$\boldsymbol{Y}(t) = \boldsymbol{X}(t, \boldsymbol{\gamma}(t))^\top \boldsymbol{\beta} + \boldsymbol{w}(t) + \boldsymbol{\varepsilon}(t), \quad t \in \mathbb{R}^+, \tag{4.3}$$

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 $\boldsymbol{\varepsilon}(t) \sim \mathcal{N}_K(0, \tau^2 \boldsymbol{I}_K)$, $\tau^2 \in \mathbb{R}^+$, is a white noise process for measurement error. Assuming independence among all the components of the temporal process, i.e. among individuals, each element of $\boldsymbol{w}(t)$ is specified as

$$w_k(t) \overset{ind}{\sim} \mathcal{GP}\left(0, c_{\theta_k}(\cdot, \cdot)\right), \tag{4.4}$$

where $c_{\theta_k}(\cdot, \cdot)$ is a covariance function with parameters $\theta_k \in \Theta$.

Let y_{ki} and $\boldsymbol{x}_{ki}$ be the outcome and covariates for individual k at 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 individual k. We will then refer to the joint $n \times 1$ vector of the outcomes and $n \times p$

matrix of the predictors as:

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

where $\boldsymbol{X}_k$ 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. Then, let us denote with $\{\boldsymbol{w}_k\}_{k=1}^K$ the $T_k \times 1$ vectors comprising all the random effects on individual k, forming the $n \times 1$ vector $\boldsymbol{w} = \begin{bmatrix}\boldsymbol{w}_{1\cdot}^\top & \boldsymbol{w}_{2\cdot}^\top & \cdots & \boldsymbol{w}_{K\cdot}^\top\end{bmatrix}^\top$. We extend Equation (4.3) to a hierarchical model with posterior distribution

$$\pi(\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\theta}, \tau^2 \,|\, \boldsymbol{y}) \propto \pi(\boldsymbol{\theta}, \tau^2) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta) \times N(\boldsymbol{w} \,|\, \boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{\theta}}) \times N(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w}, \tau^2 \boldsymbol{I}_n)\,. \tag{4.5}$$

The independence assumption of Equation (4.4) is not strictly necessary, however applying Equation (4.5) involves the determinant and inverse of $\boldsymbol{C}_{\boldsymbol{\theta}}$, which require $\mathcal{O}(n^2)$ storage space and $\mathcal{O}(n^3)$ floating point operations (flops). This operations are cumbersome and become already unfeasible for $n \approx 10^3$, as the covariance matrix $\boldsymbol{C}_{\boldsymbol{\theta}}$ is dense (see Figure 4.5a). Under Equation (4.4) instead, the covariance matrix $\boldsymbol{C}_{\boldsymbol{\theta}} = \operatorname{diag}(\boldsymbol{C}_{\boldsymbol{\theta}_1,1}, \boldsymbol{C}_{\boldsymbol{\theta}_2,2}, \dots, \boldsymbol{C}_{\boldsymbol{\theta}_K,K}) = \oplus_{k=1}^K \boldsymbol{C}_{\boldsymbol{\theta}_k,k}$ 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, and where the covariance between observations belonging to different individuals is set to 0 (see Figure 4.5b). Each individual is allowed its own covariance parameters, $\boldsymbol{\theta}_k$, and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_K\}$ in Equation (4.5) is the collection of all the covariance kernel parameters. The block-diagonal structure of $\boldsymbol{C}_{\boldsymbol{\theta}}$ considerably alleviates this burden since

$$\det(\boldsymbol{C}_{\boldsymbol{\theta}}) = \prod_{k=1}^K \det(\boldsymbol{C}_{\boldsymbol{\theta}_k,k}), \qquad \boldsymbol{C}_{\boldsymbol{\theta}}^{-1} = \operatorname{diag}\left(\boldsymbol{C}_{\boldsymbol{\theta}_1,1}^{-1}, \boldsymbol{C}_{\boldsymbol{\theta}_2,2}^{-1}, \dots, \boldsymbol{C}_{\boldsymbol{\theta}_K,K}^{-1}\right).$$

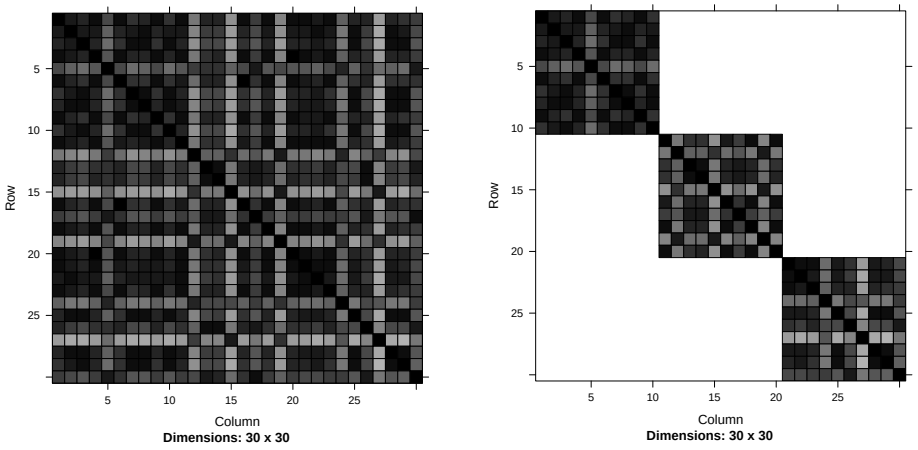
This reduces the flop count from $\mathcal{O}(n^3) = \mathcal{O}((\sum_{k=1}^K T_k)^3)$ to $\mathcal{O}(K \sum_{k=1}^K (T_k)^3)$, with a significant saving of calculations especially when the T_k's are reasonably small ($< 10^4$). Furthermore, each $\boldsymbol{C}_{\boldsymbol{\theta}_k,k}$ can be computed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

However, analyzing the Actigraph data in Section 4.2 will involve $T_k > 10^5$ measurements from some individuals. Full inference will be impractical without any exploitable structure for each $\boldsymbol{C}_{\boldsymbol{\theta}_k,k}$. Analyzing massive spatio-temporal data has witnessed burgeoning interest and a comprehensive review is beyond the scope of this work (see, e.g., Heaton et al., 2019, and references therein). We will pursue an approximation due to Vecchia (Vecchia, 1988), based on the directed acyclic graph (DAG), that has generated substantial recent interest (Datta et al., 2016a,b; Katzfuss et al., 2020; Katzfuss and Guinness, 2021; Peruzzi et al., 2020) in scalable Bayesian modeling.

4.3.2 Independent DAG models over individuals

We adapt *Vecchia's* likelihood approximation (Vecchia, 1988) to the random effects $\boldsymbol{w}_k$ for each $k = 1, 2, \dots, K$. Beginning with the observed time points $\{t_{k1} < t_{k2} < \cdots < t_{kT_k}\}$ for individual k and the DAG representation $\pi(\boldsymbol{w}_k) = \pi(w_{k1}) \prod_{i=2}^{T_k} \pi(w_{ki} \,|\, w_{k1}, \dots, w_{k(i-1)})$, we define

$$\pi(\boldsymbol{w}_k) \approx \tilde{\pi}(\boldsymbol{w}_k) = \pi(w_{k1}) \prod_{i=2}^{T_k} \pi(w_{ki} | \boldsymbol{w}_{k,N(i)})\,, \tag{4.6}$$



(a) $\boldsymbol{C}_{\boldsymbol{\theta}}$ **(b)** $\oplus_{k=1}^{K} \boldsymbol{C}_{\boldsymbol{\theta}_k,k}$

Figure 4.5. Covariance matrix of the process before (a) and after (b) assuming independence among individuals.

where $\tilde{\pi}(\cdot)$ is the joint density derived from $\pi(\boldsymbol{w}_k)$ by restricting the parents (conditional sets) of each w_{ki} in the DAG to a set $w_{kN(i)} = \{w_{kj} : j \in N(i)\}$, where $N(i)$ is a set of prefixed size m comprising the m nearest neighbors of t_{ki} from the past. Thus, $N(i) = \{t_{k(i-m)}, \ldots, < t_{k(i-1)}\}$ for $i > m$ and $N(i) = \{t_{k1}, \ldots, t_{k(i-1)}\}$ for $i \leq m$.

Such approximations yield valid[4] probability likelihoods (Lauritzen, 1996; Stein et al., 2004; Murphy, 2012) and can be extended to stochastic processes (Datta et al., 2016a) for inference on arbitrary time points.

The connection between sparsity and conditional independence follows by writing Equation (4.6) as a linear model $\boldsymbol{w}_k = \boldsymbol{A}_k \boldsymbol{w}_k + \boldsymbol{\eta}_k$,

$$\begin{aligned} \boldsymbol{w}_k &= \boldsymbol{A}_k \boldsymbol{w}_k + \boldsymbol{\eta}_k, \\ \boldsymbol{\eta}_k &\sim \mathcal{N}_{T_k}(0, \boldsymbol{D}_k) \end{aligned} \tag{4.7}$$

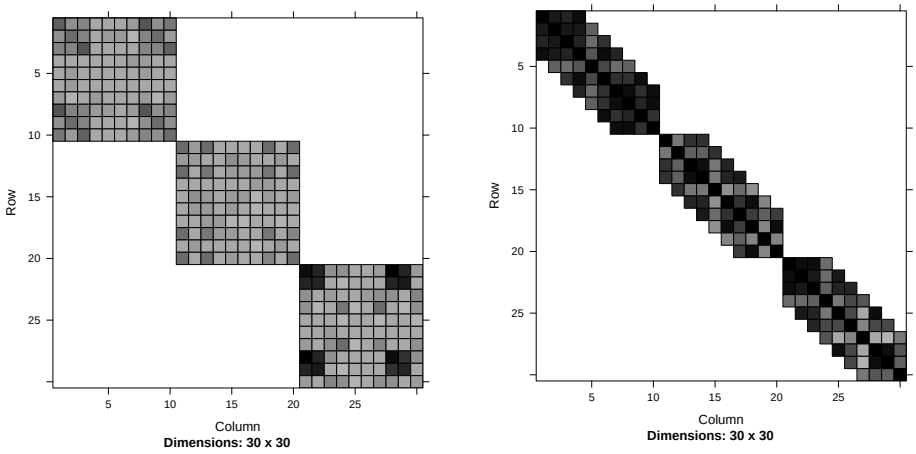
where $\boldsymbol{A}_k$ is a $T_k \times T_k$ strictly lower triangular matrix, $\boldsymbol{\eta}_k \sim \mathcal{N}_{T_k}(\boldsymbol{0}, \boldsymbol{D}_k)$ and $\boldsymbol{D}_k$ is the $T_k \times T_k$ diagonal matrix such that $[\boldsymbol{D}_k]_{ii} = d_{ii} = \mathrm{Var}\left(w_{ki} | \{w_{kj}, j < i\}\right)$ for $i = 1, \ldots, T_k$. The DAG imposes the lower-triangular structure on $\boldsymbol{A}_k$ and its (i, j)-th entry is allowed to be nonzero only for $j \in N(i)$. Therefore, each row of $\boldsymbol{A}_k$ has at most m nonzero entries so that

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

identifies with the *sparse Cholesky* decomposition of $\boldsymbol{C}_k^{-1}$, and where $\tilde{\boldsymbol{C}}_k^{-1}$ is the precision matrix corresponding to $\tilde{\pi}(\boldsymbol{w}_k)$. Replacing $\boldsymbol{C}$ with $\tilde{\boldsymbol{C}}$ in Equation (4.5) yields a computationally efficient hierarchical model with $\prod_{k=1}^{K} N(\boldsymbol{w}_k \,|\, \boldsymbol{0}, \tilde{\boldsymbol{C}}_k)$ as the prior on $\boldsymbol{w}$. An example of the structure of $\tilde{\boldsymbol{C}}$ and $\tilde{\boldsymbol{C}}^{-1}$ is given in Figures 4.6a and 4.6b.

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 $[\cdot, \cdot]$

[4] and consistent with respect to the parent process, as far as the size of the neighbour sets tends to the full size T_k.



(a) $\tilde{\boldsymbol{C}}_{\boldsymbol{\theta}}$ **(b)** $\oplus_{k=1}^{K} \tilde{\boldsymbol{C}}_{\theta_k,k}^{-1}$

Figure 4.6. Approximated covariance (a) and precision (b) matrix of the process.

indicates sub-matrices 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 $\boldsymbol{w}_k^\top \tilde{\boldsymbol{C}}_k^{-1} \boldsymbol{w}_k$ in terms of $\boldsymbol{A}_k$ and $\boldsymbol{D}_k$ and the determinant $\det(\tilde{\boldsymbol{C}}_k) = \prod_{i=1}^{T_k} d_{ii}$ at almost no additional cost.

Algorithm 4 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 $\tilde{\boldsymbol{w}}_k$ and replace the density of the latent process in Equation (4.5) with $\mathcal{N}_{T_k}(\tilde{\boldsymbol{w}}_k \,|\, \boldsymbol{0}, \tilde{\boldsymbol{C}}_k)$, 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 Equation (4.6) and Equation (4.7) depend upon the 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 $t_{k1} < t_{k2} < \cdots < t_{kT_k}$, 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:

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

As a result, the lower triangular matrix $\boldsymbol{A}_k$ is not just sparse but also banded, with a lower bandwidth equal to m. Consequently, $\tilde{\boldsymbol{C}}_k^{-1}$ is also banded with lower and upper bandwidth equal to m. This leads to further accrual of computational benefits. The overall cost is $\mathcal{O}(\sum_{k=1}^{K} T_k m^3) = \mathcal{O}(nm^3)$ (linear in n) for computing the posterior for any given values of the parameters.

4.3.3 Implementation using collapsed models

The Bayesian hierarchical model in Equation (4.5), either with $\boldsymbol{C}_{\boldsymbol{\theta}}$ or with $\tilde{\boldsymbol{C}}_{\boldsymbol{\theta}}$ in the prior for $\boldsymbol{w}$, allows full posterior inference for $\{\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\theta}, \tau^2\}$ using Markov chain Monte Carlo (MCMC). In particular, its standard implementation relies on a

Algorithm 4: 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**
    $\boldsymbol{L}_k[i+1,\, i+1] = 1$
    $\boldsymbol{L}_k[i+1,\, N(i+1)] = -\boldsymbol{C}_k[N(i+1),\, N(i+1)]^{-1} \cdot \boldsymbol{C}_k[N(i+1),\, i+1]$
    $\boldsymbol{d}_k[i+1] = C[i+1,\, i+1] - \boldsymbol{C}_k[i+1,\, N(i+1)] \cdot \boldsymbol{L}_k[i+1,\, N(i+1)]^\top$
    $\boldsymbol{R}_k[i+1,\, i+1] = 1/\boldsymbol{d}_k[i+1]$
    $\boldsymbol{R}_k[N(i+1),\, i+1] = \boldsymbol{L}_k[i+1,\, N(i+1)]^\top/\boldsymbol{d}_k[i+1]$
  end
  $\boldsymbol{L}[(T_{k-1}+1 : T_k),\, (T_{k-1}+1 : T_k)] = \boldsymbol{L}_k$
  $\boldsymbol{R}[(T_{k-1}+1 : T_k),\, (T_{k-1}+1 : T_k)] = \boldsymbol{R}_k$
  $\boldsymbol{d}[T_{k-1}+1 : T_k] = \boldsymbol{d}_k$
end

sequential sampler (a.k.a. *Sequential NNGP*) 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 convenience is 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 $\boldsymbol{\theta}$. In this respect, samplers based on 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 authors compared three alternatives to the original sequential sampler, which attempt at improving its performances through the exploitation of high-performance computing libraries for expensive numerical linear algebra computations. These have been named as the *Collapsed NNGP*, the *NNGP for the response* and the *Conjugate NNGP*. In the sequel, we only describe the implementation with collapsed likelihoods in the specific context of temporal processes, as they represent the only appropriate choice if the objective is to provide full inference on the latent component. 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 Equation (4.5), 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:

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

where $\tilde{\boldsymbol{\Lambda}} = \tilde{\boldsymbol{C}} + \tau^2 \cdot \boldsymbol{I}_n$. Hence, instead of Equation (4.5), we sample from:

$$\pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2 \,|\, \boldsymbol{y}) \propto \pi(\boldsymbol{\theta}, \tau^2) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta) \times N(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}, \tilde{\boldsymbol{C}}_{\boldsymbol{\theta}} + \tau^2 \boldsymbol{I}_n)\,, \tag{4.8}$$

Algorithm 5: Sampling from the posterior of the collapsed temporal model

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 neighbor sets $\{N_k(i)\}_{i=1}^{T_k}$
  end
end

1: Metropolis-Hastings update for $\{\boldsymbol{\theta}, \tau^2\}$
$\pi(\boldsymbol{\theta}, \tau^2|\cdot) \propto \pi(\boldsymbol{\theta}, \tau^2) \times \frac{1}{\sqrt{\det \tilde{\boldsymbol{\Lambda}}}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \tilde{\boldsymbol{\Lambda}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right)$
begin
  for $k = 1, \dots, K$ **do**
    a: Compute $\boldsymbol{L}_k = \left(\boldsymbol{I}_{T_k} - \boldsymbol{A}_k\right)^\top$, $\boldsymbol{d}_k = \text{diag}(\boldsymbol{D}_k)$ and $\boldsymbol{R}_k = \boldsymbol{D}_k^{-1}\left(\boldsymbol{I}_{T_k} - \boldsymbol{A}_k\right)$ 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.
  end
  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 $\pi(\boldsymbol{\theta}, \tau^2|\cdot) \propto \frac{\exp(-q_2/2)}{\sqrt{q_1}} \cdot \pi(\boldsymbol{\theta}, \tau^2)$
end

2: Gibbs' sampler update for $\boldsymbol{\beta}$
$\boldsymbol{\beta}|\cdot \sim \mathcal{N}_p(\boldsymbol{B}^{-1}\boldsymbol{b}, \boldsymbol{B}^{-1})$, where $\boldsymbol{B} = \boldsymbol{X}^\top\tilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X} + \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}$ and $\boldsymbol{b} = \boldsymbol{X}^\top\tilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{y} + \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}$
begin
  a: Compute $\boldsymbol{F} = \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}$ and $= \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}$
  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 $\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{B}^{-1}\boldsymbol{b}, \boldsymbol{B}^{-1})$
end

Repeat steps **1** and **2** to obtain M MCMC samples for $\{\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2\}$

considerably improving mixing and convergence.

We will need to compute the inverse and determinant of $\tilde{\boldsymbol{\Lambda}} = \tilde{\boldsymbol{C}}_{\boldsymbol{\theta}} + \tau^2\boldsymbol{I}_n$, which is $n \times n$. While $\tilde{\boldsymbol{\Lambda}}^{-1}$ does not share the same convenient factorization of $\tilde{\boldsymbol{C}}^{-1}$ and is also not guaranteed to be sparse, the Sherman-Woodbury-Morrison formulas reveal

$$\tilde{\boldsymbol{\Lambda}}^{-1} = \tau^{-2}\boldsymbol{I} - \tau^{-4}\boldsymbol{\Omega}^{-1}, \quad \text{with} \quad \boldsymbol{\Omega} = \tilde{\boldsymbol{C}}^{-1} + \tau^{-2}\boldsymbol{I}\,, \tag{4.9}$$

where $\boldsymbol{\Omega}$ enjoys the same sparsity as $\boldsymbol{C}^{-1}$. Moreover, $\det(\tilde{\boldsymbol{\Lambda}}) = \tau^{2n}\det(\tilde{\boldsymbol{C}})\det(\boldsymbol{\Omega})$. The core of the algorithm is therefore to compute $\tilde{\boldsymbol{\Lambda}}^{-1}$ through $\boldsymbol{\Omega}$. In our application, the random effect is assumed to be the realization of K independent temporal processes. As discussed in Section 4.3.2, this implies a block-diagonal structure for $\tilde{\boldsymbol{C}}$ that can be shown to be shared also by $\boldsymbol{\Omega}$ (see Equation (4.9)). Each block $\boldsymbol{\Omega}_k$ of $\boldsymbol{\Omega}$ can be computed independently for each individual and 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; see Algorithm 5. 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 Equation (4.8), where $\boldsymbol{\beta}$ is updated from its full conditional distribution, while $\{\boldsymbol{\theta}, \tau^2\}$ 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 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 and Roberts, 2006; Roberts and Rosenthal, 2009). The algorithm has been coded using the `R 4.0.1` statistical environment and `C++`, exploiting the interface provided by the `Rcpp` package (Eddelbuettel et al., 2011). All expensive computations are managed by the `Eigen` library (version 3.3.7, Guennebaud et al. (2010)), which provides efficient routines for numerical linear algebra with an emphasis on sparse matrices. Our implementation of Equation (4.8) outperforms the algorithms that update $\boldsymbol{w}$ in terms of computational speed as it is implemented in the `spNNGP` package (Finley et al., 2017). We present these comparisons in Appendix B.

4.3.4 Including the spatial effect

Accounting for spatial information in our Actigraph dataset requires some new considerations. As mentioned in Section 4.1, spatial information is available in terms of the physical location along the trajectory as well as through covariates that are functions of space. Considering the discussion in Section 4.3, the analytical goals of this dataset suggest accounting for *spatial heterogeneity*. Here, modeling $\boldsymbol{w}(\cdot)$ in Equation (4.3) as a spatio-temporal process (also considering scalable versions) is challenging for three main reasons: (i) the trajectory's domain does not have a positive area, (ii) associations among the measurements are more amenable to the temporal scale, and (iii) potentially retrievable spatially-referenced features may be recorded at different resolutions, either among them or with respect to the observed process. Therefore, we introduce spatial effects into the mean employing a smooth function of space, $f_S(\cdot) : \mathbb{R}^2 \to \mathbb{R}$, approximated by a spline basis representation (see, e.g., Goodman and Hardin, 2006; Ramsay and Silverman, 2007). For instance, if J_x and J_y are the dimensions of independently defined B-spline basis expansions on the x and y coordinates, respectively, then $f_S((x,y)) \approx \tilde{f}_S((x,y)) = \sum_{j=1}^{J_x}\sum_{h=1}^{J_y}\beta_{S,(j,h)}B_{x,j}(x)B_{y,h}(y)$, where $B_{x,j} = [\boldsymbol{B}_x]_j$ and $B_{y,h} = [\boldsymbol{B}_y]_h$ are the j-th and h-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 now modify Equation (4.3) to include the spline,

$$\boldsymbol{Y}(t) = \boldsymbol{X}(t)\boldsymbol{\beta} + \boldsymbol{B}_S(\boldsymbol{\gamma}(t))\boldsymbol{\beta}_S + \boldsymbol{w}(t) + \boldsymbol{\varepsilon}(t), \qquad t \in \mathbb{R}^+, \tag{4.10}$$

where $\boldsymbol{\gamma}(t) = \{\boldsymbol{\gamma}_1(t), \boldsymbol{\gamma}_2(t), \ldots, \boldsymbol{\gamma}_K(t)\}$, $\boldsymbol{\gamma}_k(\cdot) = (\gamma_{k,x}(t), \gamma_{k,y}(t)) : \mathbb{R}^+ \to \mathbb{R}^2$ 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. A proper choice 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 over-fitting. 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 $\boldsymbol{\gamma}(\mathcal{T}) = \{\gamma_1(t_{11}), \gamma_1(t_{12}), \ldots, \gamma_K(t_{KT_k})\}$.

Algorithm 6: $\boldsymbol{\psi}$ 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 $\boldsymbol{G} = \boldsymbol{X}^{*\top}\tilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^* + \boldsymbol{V}_{\boldsymbol{\psi}}^{-1}$ and $\boldsymbol{g} = \boldsymbol{X}^{*\top}\tilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{y} + \boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\psi}}$
begin
- **a:** Compute $\boldsymbol{F} = \boldsymbol{V}_{\boldsymbol{\psi}}^{-1}$ and $= \boldsymbol{V}_{\boldsymbol{\psi}}^{-\mathbf{1}}\boldsymbol{\mu}_{\boldsymbol{\psi}}$
- **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 $\boldsymbol{\psi} \sim \mathcal{N}_{p^*}(\boldsymbol{G}^{-1}\boldsymbol{g}, \boldsymbol{G}^{-1})$

end

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 $\lambda \sim \mathcal{G}\,(\alpha^*_\lambda,\, \beta^*_\lambda)$

end

Following Equation (4.8), we sample from the posterior,

$$\pi(\boldsymbol{\beta}, \boldsymbol{\beta}_S, \boldsymbol{\theta}, \tau^2 \,|\, \boldsymbol{y}) \propto \pi(\boldsymbol{\theta}, \tau^2) \times \pi_S(\boldsymbol{\beta}_S) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta) \times N(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{\beta}_S, \tilde{\boldsymbol{C}}_{\boldsymbol{\theta}} + \tau^2\boldsymbol{I}_n)\,, \tag{4.11}$$

where the prior $\pi_S(\cdot)$ shall be accurately chosen. 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) so over-fitting will not be an issue. However, some areas present sparsely observed points (trajectories are not uniformly distributed, as shown in Figure 4.3). This may cause coefficients corresponding to those regions to be weakly identified. To control for the balance of all these 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 (Hastie et al., 2000; Lang and Brezger, 2004). While keeping the Gaussian priors, 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, $\boldsymbol{\beta}_S \,|\, \lambda \propto \exp\left\{-\frac{\lambda}{2} \cdot \boldsymbol{\beta}_S\boldsymbol{P}\boldsymbol{\beta}_S^\top\right\}$ and $\lambda \sim \mathcal{G}(\alpha_\lambda, \beta_\lambda)$. We consider two possible forms for $\boldsymbol{P}$, which imply different penalization for 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 we collect the B-Spline basis elements with the other covariates as $\boldsymbol{X}^* = [\boldsymbol{X}, \boldsymbol{B}]$

and stack the corresponding coefficients 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 $\boldsymbol{\psi} \,|\cdot \propto \mathcal{N}_J\left(\boldsymbol{\psi} \,\middle|\, \boldsymbol{G}^{-1}\boldsymbol{g},\, \boldsymbol{G}^{-1}\right)$, where $\boldsymbol{G} = \boldsymbol{X}^{*\top}\tilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{*} + \boldsymbol{V}_{\boldsymbol{\psi}}^{-1}$ and $\boldsymbol{g} = \boldsymbol{X}^{*\top}\tilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{y} + \boldsymbol{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\psi}}$ with $\boldsymbol{V}_{\boldsymbol{\psi}}^{-1} = \text{diag}\left(\boldsymbol{V}_{\boldsymbol{\beta}}^{-1},\, \lambda \cdot \boldsymbol{P}\right)$ and $\boldsymbol{g} = \left[\boldsymbol{\mu}_{\boldsymbol{\beta}},\, \boldsymbol{\mu}_{\boldsymbol{\beta}_S}\right]^{\top} = \boldsymbol{0}^{\top}$. Moreover, the Gamma prior on λ implies a Gamma full-conditional distribution $\lambda \,|\cdot \propto \mathcal{G}\left(\lambda \,\middle|\, \alpha_\lambda + 1/2,\, \beta_\lambda + \boldsymbol{\beta}_S^{\top}\boldsymbol{P}\boldsymbol{\beta}_S\right)$.

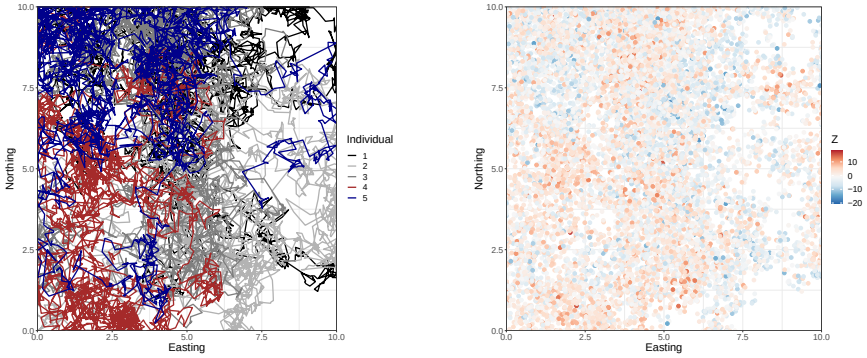
Estimating the model in (4.10) is achieved through a straightforward extension of Algorithm 5. We jointly update $\boldsymbol{\psi}$ and λ from their full conditional distributions. Algorithm 6 shows how the Gibbs' sampling step of Algorithm 5 can be modified to get full inference also on the spline 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 w.r.t. $p^* \to \mathcal{O}(np^{*2})$. Steps 1a and 1b (i.e. the most expensive in p^*) are executed in the first iteration and subsequently, only in those iterations where new values of $\boldsymbol{\theta}$ are accepted. When $\boldsymbol{\theta}$ is rejected, we retain in memory the previously computed value (which would stay unchanged). Thus, if we attain an optimal acceptance rate of $\approx 20\% - 30\%$ in the Metropolis Hastings step on $\boldsymbol{\theta}$, the computation is avoided in the majority of cases with a sensible improvement in computation time and speed.

4.3.5 Simulations

We conducted simulation experiments to evaluate the model described in Section 4.3.4 and compared the performance of our algorithm in terms of fitting, prediction error and computational speed with other routines available from the `spNNGP` package. Additional comparative experiments are provided in Section B.1 and B.2 of Appendix B. We executed our MCMC algorithms on a computing environment 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. Each of the following jobs, and the ones from Section 4.4, have been executed on a single node exploiting all 16 cores. The results presented here and in Section 4.4 are based upon posterior samples that were retained after diagnosing convergence using visual tools (e.g., traceplots, autocorrelation), effective sample sizes, Monte Carlo standard errors (MCSE) and other diagnostics offered by the `coda`, `mcse` and `bayesplot` packages in the `R` computing environment.

We first generated $T_k = 2 \times 10^5$ time points for $K = 5$ individuals, where each time point t_{ki} followed exponential waiting times between observations, i.e. $t_{ki} = \sum_{h=1}^{i-1} \delta_h$, and $\delta_h \overset{iid}{\sim} \text{Exp}(5)$. Given the time points, we constructed spatial trajectories $\boldsymbol{\gamma}_k(\cdot)$, $k = 1, \ldots, K$, by simulating $\boldsymbol{s}_k = [\gamma_k(t_{k1}), \ldots, \gamma_k(t_{kT_k})]^{\top}$, where subsequent components were 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. If the trajectory left the square, it was projected onto the border and the next step would resume from there. The simulated trajectories are shown in Figure 4.7a.

Given the time points and positions, we generated the latent temporal Gaussian processes $w_k(\cdot) \overset{ind}{\sim} \mathcal{GP}(0, c_{\boldsymbol{\theta}}(\cdot, \cdot))$ with an exponential covariance $c_{\boldsymbol{\theta}}(t, t') = \sigma^2 \exp\{-\phi \cdot |t - t'|\}$, where $\sigma^2 > 0$ represents the variance of the process, $\phi > 0$ is the decay in temporal correlation (range) and $\tau^2 > 0$ the residual variance (nugget). The spatial effects are then introduced through $f_S(\cdot) : \mathcal{S} \to \mathbb{R}$ by considering a



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

Figure 4.7. Observed trajectories (a) and observed points (b) for the simulated dataset.

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	

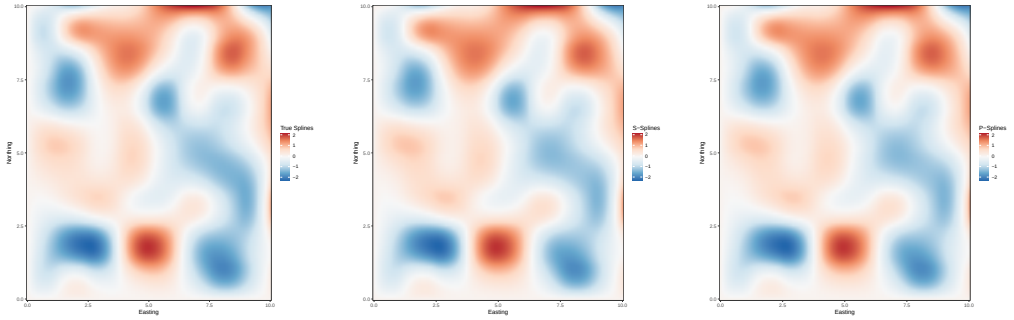
Table 4.2. Parameter estimates, predictive validation and fitting times (hours) on the simulated dataset for all the considered models.

tensor product spline basis of degree 2 and with 9 knots over the square domain (including boundary knots), where 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 individual-specific intercepts $\{\beta_{0k}\}_{k=1}^{5}$ 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 common across individuals, and set to be determined by slopes $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3]^\top$.

We generated values of the outcome for individual k at time t_{ki} and location $\boldsymbol{s}_{ki} = \gamma_k(t_{ki})$ according to the generative process defined by Equation (4.10) with parameters fixed as above. This yielded a simulated dataset $D_{sim} = \left\{(\text{Ind}_j, t_j, \boldsymbol{s}_j, y_j, \boldsymbol{x}_j^\top)\right\}_{j=1}^{n}$ with $n = 10^5$ observations, where Ind_j denotes the individual corresponding to row j. Then, we fit the model in Equation (4.11) on 70% of the total observations in D_{sim} using Algorithm 5 with the Gibbs' sampling modified as in Algorithm 6. The remaining 30% were held out to assess out-of-sample predictive performances in terms of *Relative and Root Mean Squared Prediction Error* (RMSPE), and *Coverage*, *Predictive Interval Width* (PIW). Intercept and slope regression parameters were

assigned $\mathcal{N}(0, 10^6)$ priors; the variance components, σ^2 and τ^2, were both assigned inverse Gamma $\mathcal{IG}(2, 2)$ priors; and the decay parameter ϕ received a Gamma prior $\mathcal{G}(1, 1)$. For the spline coefficients, we considered both the penalized versions in Section 4.3.4. The first is referred to as an S-Spline (shrinking splines), and the second as P-Spline (penalized splines).

Table 4.2 presents the posterior estimates. We also included the Deviance Information Criterion (DIC) for both models. Performances in the two settings are almost identical, but the DIC favors the S-Spline model. This is not surprising as the data were generated using an analogous shrinkage prior for the $\boldsymbol{\beta}_S$'s. Further details, including the estimates of the spline coefficients are provided in Appendix B. Figure 4.8 presents the posterior estimate of the spatial surface. We compare the true latent surface with the two (practically identical) estimates.



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

Figure 4.8. 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).

4.4 Application

We analyze activity levels throughout the "active time"—when the Actigraph device records the individual as being physically active—excluding epochs when the device was not worn or when there was no physical activity (e.g., the individual was sitting or lying down). The data processing and merging of actigraph data with GPS locations resulted in two final datasets (Section 4.2). These are treated separately. In both applications, 70% of the total observations were used for training the model, while the remaining were excluded to assess the out-of-sample predictive performances.

4.4.1 Temporal model

We first analyze D_1. Our predictors include a binary variable indicating if the measures refer to the period before or after a Bruin Bike Share (BBS) program was launched in Westwood, Los Angeles to account for the effect of a 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 linear combination of J_H spline basis functions $\phi_j(\cdot)$ with unknown coefficients $\beta_{H,j}$'s, $f_H(h) \approx \tilde{f}_H(h) =$

$\sum_{j=1}^{J_H} \beta_{H,j} B_{H,j}(h) = \boldsymbol{B}_H(h)\boldsymbol{\beta}_H$. The full process specification yields

$$\boldsymbol{Y}(t) = \boldsymbol{X}(t)\boldsymbol{\beta} + \boldsymbol{B}_H\left(h(t)\right)\boldsymbol{\beta}_H + \boldsymbol{w}(t) + \boldsymbol{\varepsilon}(t), \qquad t \in \mathbb{R}^+ \tag{4.12}$$

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 splines at that point.

We employ a second order approximation with 4 internal knots spread uniformly over the domain. Collecting the basis elements in the design matrix and stacking the coefficients as for the spatial model in Section 4.3.4 introduces 6 additional columns (hence, only 6 additional slope parameters) in the design matrix (i.e. the spline basis functions evaluated at the observed time-points). The large number of observations in each epoch and the reduced number of knots subdue any concerns surrounding over-fitting and unrobust inference.

We use a logarithmic transformation, $lMAG_k(t) = \log(MAG_k(t))$ for $k = 1, 2, \ldots, K$ and $t = t_{k1}, \ldots, t_{kT_k}$. We denote the parameter associated with variable "varname" as β_{varname} and the levels of each categorical covariate as $\text{varname}_{(j)}$ for $j = 1, \ldots, J_{\text{varname}}$. Hence,

$$\begin{aligned} lMAG_k(t) = \beta_0 + \beta_{\text{BMI}} \cdot \text{BMI}_k + \\ + \sum_{j=2}^{J_{\text{Eth}}} \beta_{\text{Eth},j} \cdot \mathbb{I}\left(\text{Ethnicity}_k = \text{Eth}_{(j)}\right) + \sum_{j=2}^{J_{\text{Age}}} \beta_{\text{Age},j} \cdot \mathbb{I}\left(\text{AgeClass}_k = \text{Age}_{(j)}\right) + \\ + \sum_{j=2}^{J_{\text{Sex}}} \beta_{\text{Sex},j} \cdot \mathbb{I}\left(\text{Sex}_k = \text{Sex}_{(j)}\right) + \sum_{j=1}^{J_H} \beta_{H,j} B_{H,j}\left(h(t)\right) + w_k(t) + \epsilon_k(t)\,, \end{aligned} \tag{4.13}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, $w_k(\cdot)$ is the DAG-based approximation (Section 4.3.2) for $\mathcal{GP}(0, c_{\boldsymbol{\theta}}(\cdot,\cdot))$ with $c_{\boldsymbol{\theta}}(t, t')$ the exponential function and $\epsilon_k(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2)$. For the categorical variables "Age" and "Sex", the levels start from 2 as $J = 1$ is taken as the baseline, which corresponds to an Asian female with age between 20 and 25 years. Other socioeconomic factors (e.g. education and income level) have been excluded from the analysis as they are strongly associated with ethnicity and age, while *Lux* (detecting light exposition) was excluded after some preliminary analysis revealed its low predictive power[5]. We have assigned priors such as $\boldsymbol{\beta} \sim \mathcal{N}_J(\mathbf{0}, 10^6 \cdot \boldsymbol{I}_J)$, $\sigma^2 \sim \mathcal{IG}(2, 2)$ and $\tau^2 \sim \mathcal{IG}(2, 2)$ with J being the total number of $\boldsymbol{\beta}$ coefficients. The presence of temporal dependence was investigated through an individual-specific exploratory analysis on the residuals from an ordinary least squares linear regression. Subsequently, it was decided that an exponential covariance function for temporal dependence (corresponding to an Ornstein-Uhlenbeck process) will be a parsimonious and effective choice to model the behavior of the underlying residual process.

We implemented Algorithm 5 for Equation (4.13). Inference was based on $10,000$ posterior samples retained after convergence was evinced (discarding an initial $5,000$ iterations as burn-in). Point estimates from the standard linear model were used as starting values for the regression coefficients. The run time of the collapsed sampler

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

(Section 4.3.3) on D_1 was ≈ 15 hours, achieving a desirable acceptance rate of $\approx 28\%$ at convergence.

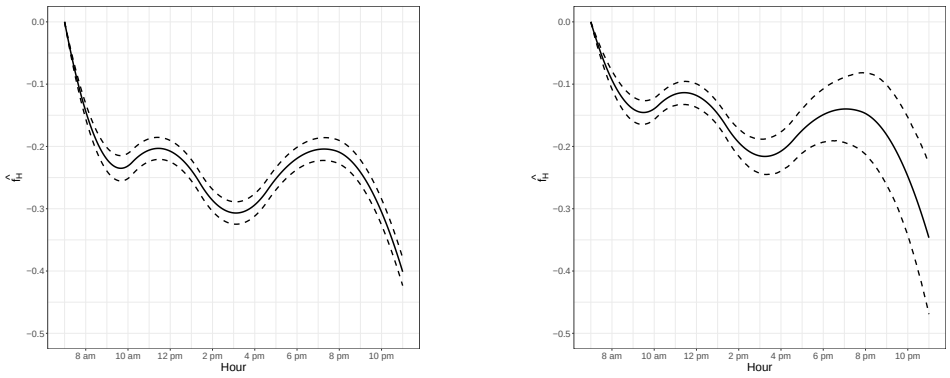
Table 4.3 presents the posterior estimates. The regression coefficients were slightly different from a Bayesian linear regression model, but presented very similar inference. African Americans, Latinos and Whites revealed higher values of $lMAG$ than Asian-Americans as did males over females. As expected, higher age-groups revealed lower $lMAG$. Unsurprisingly, the introduction of the temporal process produces 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 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. 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 4.1 and (4.2). This value, while large, is not surprising as we are modeling the epochs corresponding to active time.

Parameter	**Collapsed Model**		**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 4.3. Parameter credible intervals, 95%(2.5%, 97.5%) and predictive validation for 15×10^3 MCMC iterations on D_1.

The estimate of the temporal decay parameter ϕ, suggests a fairly sharp decline in temporal association, which drops below 0.05 after ≈ 1 minute (computed, with the exponential covariance function 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 4.4.2. Comparison between the estimates of the *hour of the day* spline term is shown in Figure 4.9. 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 the model in Equation (4.13) can deliver statistical estimates (with uncertainty quantification) of personalized daily PA profiles for any individual for any day. The splines and the temporal process combine to capture subject-specific diurnal variation. Figure 4.10 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.



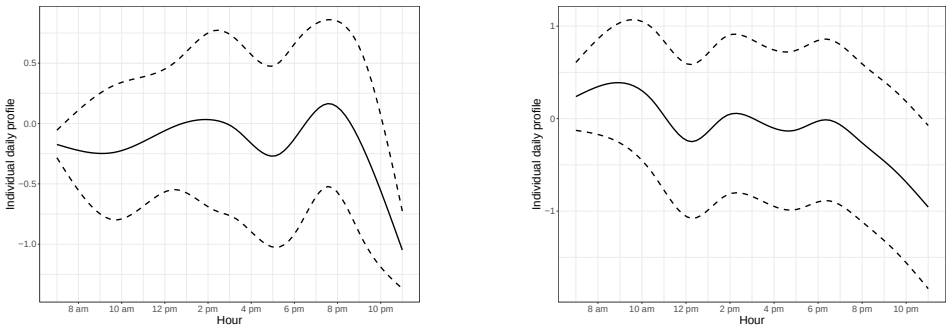
(a) Linear model **(b)** Collapsed NNGP

Figure 4.9. $\hat{f}_H$ for the linear model (left) and the NNGP model (right), with 95% credible intervals in dashed lines.

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 4.11, 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.

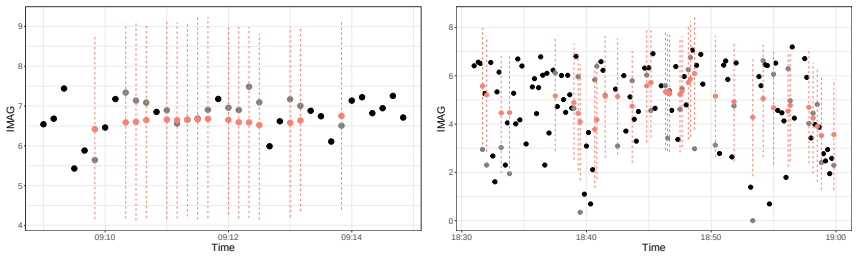
4.4.2 Including the spatial effect

We consider D_2 and fit the model in Equation (4.13), adding a spatial term (Section 4.3.4) to exploit GPS information. D_2 is restricted to those observations



(a) **(b)**

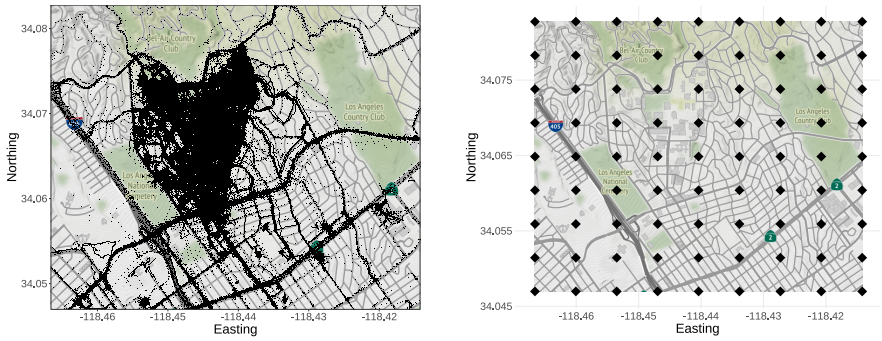
Figure 4.10. Personalized PA profiles for two individuals estimated with 95% credible intervals (dashed lines) using the spline daily effect and the temporal process in Equation (4.13).



(a) **(b)**

Figure 4.11. 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)

recorded in Westwood, Los Angeles. We introduce spatial splines obtained through the tensor product of two analogous univariate B-spline basis on longitude and latitude. We choose two bases of degree 2 with 9 equally spaced knots over a square encompassing Westwood. This sums up to $J_S = (7+2) \times (7+2) = 81$ terms for our complete spline basis, including the boundary knots.



(a) **(b)**

Figure 4.12. Observed locations (a) and knots (b) over the Westwood area.

In practice, since locations are functions of time through the trajectory function $\boldsymbol{\gamma}_k(\cdot), k = 1, \ldots, K$ of each individual, rewrite the time dependent component of the process mean as $\mu_k(t) = \beta_{\text{BBS}} \cdot \text{BruinBikeShare}(t) + \sum_{j=1}^{J_H} \beta_{H,j} B_{H,j}(h(t)) + \sum_{j=1}^{J_S} \beta_{S,j} B_{S,j}(\gamma_k(t))$, 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 4.12a) , over-fitting is not a concern. However, there are also areas in Westwood that 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, we consider the S-Spline (Ridge-like prior) for this application, where 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 4.4.1. Our posterior inference was based on $5,000$ samples retained after convergence out of $10,000$ MCMC iterations. Fitting the model to D_2 required $\approx$ 30 hours. The acceptance rate obtained is $\approx$ 28%,

Param.	Linear regression		S-Spline	
	Point	Interval	Point	Interval
Intercept	5.613	(5.536, 5.689)	5.31	(5.29, 5.33)
Eth. Latin-American	0.093	(0.079, 0.108)	0.114	(0.069, 0.159)
Eth. White	0.053	(0.040, 0.067)	0.053	(0.004, 0.102)
Eth. Black or other	0.066	(0.052, 0.080)	0.095	(0.054, 0.135)
Sex Male	0.019	(0.008, 0.029)	0.021	(-0.014, 0.055)
BMI	0.005	(0.003, 0.006)	0.006	(-0.007, 0.02)
Age [25-35)	-0.170	(-0.183, -0.156)	-0.191	(-0.227, -0.155)
Age [35-50)	-0.217	(-0.233, -0.201)	-0.249	(-0.298, -0.199)
Age [50-70]	-0.381	(-0.404, -0.359)	-0.456	(-0.528, -0.384)
BBS	-0.008	(-0.091, -0.071)	-0.107	(-0.140, -0.073)
σ^2			1.489	(1.461, 1.517)
ϕ			0.364	(0.351, 0.376)
τ^2			0.777	(0.768, 0.786)
Metric	Out-of-sample	In-sample	Out-of-sample	In-sample
Coverage	0.94	0.94	0.95	0.99
RMSPE (r)	0.95 (1.43)	0.95 (1.43)	0.53 (1.07)	0.26 (0.74)
PIW	5.62	5.62	4.78	4.74

Table 4.4. Parameter estimates and predictive validation on D_2.

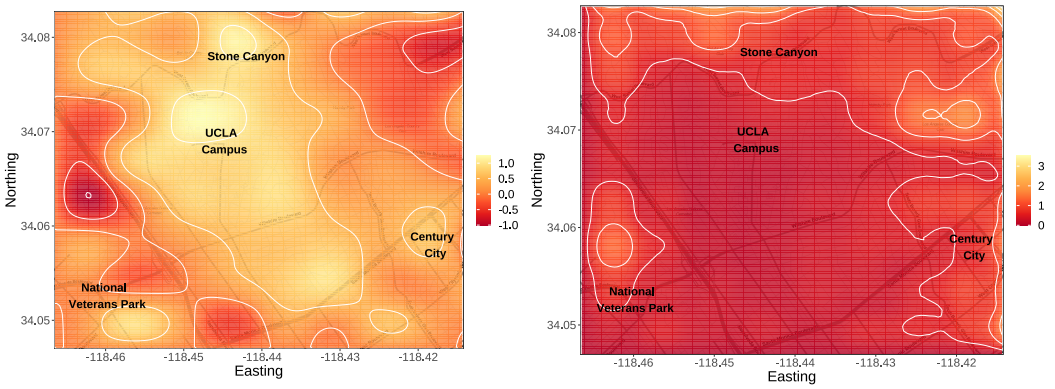
supporting the consistency of our adaptive strategy.

Table 4.4 presents parameter estimates and predictive performances of the model and compares with a standard linear regression model which includes the spatial spline terms, but neglects the temporal dependence structure. Conclusions on the regression coefficients are very similar to those from Section 4.4.1. However, accounting for the spatial effects allows for easier interpretation of the age-group regression coefficients: the older the person, the lower is the expected physical activity level. Also, both models estimate the effect of BBS as trending slightly negative. This somewhat surprising finding can be attributed to a few factors. 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, not all subjects were exposed to the BBS after its launch and, hence, could not take advantage of it.

In this application, the intercept is estimated to be ≈ 5.31 by our model and, hence, a MAG per minute count of 1214 (slightly lower than in the temporal application). This would again correspond to *vigorous* physical activity and a MET of ≈ 6.5. The estimate of ϕ implies that the dependence drops to 0 in less than a minute. Unsurprisingly, including the spatial effect and the temporal process improves predictive performances (RMSPE or PIW in Table 4.4) over a model including only spatial effects (using linear regression with splines). The spatial-temporal model delivers satisfactory coverage and outperforms its competitor in all of the other indices for the training and testing dataset.

Figure 4.13a shows the estimated spatial surface, while Figure 4.13b 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

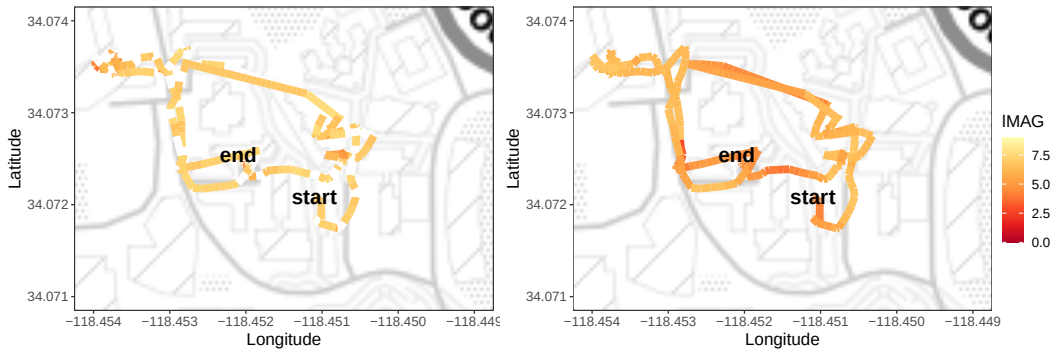
Century City shopping center (to the east). Lighter shades (orange) correspond to areas that are less developed (open space), such as the areas in the north east; or they are areas with a high degree of transportation infrastructure and traffic (e.g., toward the western boundary). These correspond to 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.



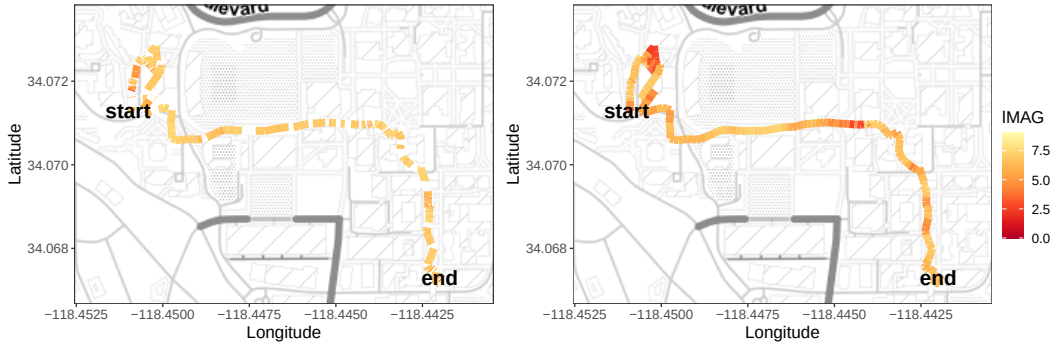
(a) **(b)**

Figure 4.13. (a) Spatially smoothed estimates from a shrinkage spline over Westwood, Los Angeles; (b) width of 95% posterior predictive intervals for the shrinkage spline.

Figure 4.14 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.



(a)



(b)

Figure 4.14. 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.

4.5 Discussion and further developments

We have devised a Bayesian modeling framework to conduct fully model-based inference for high-resolution accelerometer data over trajectories compiled from the PASTA-LA study. Our key data analytic developments included (i) modeling dependence over trajectories; (ii) accounting for subject-specific spatial-temporal variation for daily mobility; and (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). 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 spatio-temporal nature of outdoor PA research, our ability to predict in areas of data missingness drastically improve 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., 2020), 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 spatio-temporal 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 spatio-temporal 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 have 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 5

Modeling COVID-19 incident indicators

As soon as COVID-19 pandemic hit the globe, the whole scientific community buckled down and dedicated all energies in carrying out new research work to help fighting the battle against SARS-CoV-2. Perhaps, scientific research experienced unprecedented leaps forward in many fields, virology, epidemiology and statistics amongst all. The first vaccine was made available in roughly one year (Tanne, 2021), while new statistical and epidemiological models were developed to understand the dynamics of the virus spread, to predict its evolution and to build scenario evaluation following the implementation of economic, political and health interventions (Varotsos and Krapivin, 2020; Barbarossa et al., 2020; Car et al., 2020; Girardi et al., 2020a). This did not come without setbacks, mostly due to the inadequacy of national and international infrastructures to collect fully and promptly the required data. Bright side, all these research efforts toward the study of the pandemic evolution finally raised awareness about the true need of high-quality data, which are crucial for guiding decision making. Nevertheless, COVID-19 caught us unprepared and data quality frailties have been exposed worldwide during the pandemic (Idrovo and Manrique-Hernández, 2020; Costa-Santos et al., 2021; Vasudevan et al., 2021; Lloyd-Sherlock et al., 2021; Mingione and Alaimo Di Loro, 2021).

Since the beginning of April 2020, I also had the opportunity to give my contribution to the modeling of COVID-19 epidemic data. In particular, I had the pleasure of working jointly with Ph.D. Pierfrancesco Alaimo Di Loro and the StatGroup-19 research team, composed of (in alphabetic order): Prof. Fabio Divino, Prof. Alessio Farcomeni, Prof. Giovanna Jona Lasinio, Prof. Gianfranco Lovison and Prof. Antonello Maruotti. At first, we proposed a novel parametric regression model to fit *incidence* data typically collected during epidemics. This proposal was motivated by real time monitoring and short-term forecasting of the main epidemiological indicators within the first outbreak of COVID-19 in Italy (e.g. the number of daily positives and daily deceased). Model estimation was carried out via maximum-likelihood and forecasts proved to be reliable, close to the later observed true values (also for the second outbreak), and helped estimating accurately important characteristics of the epidemic, such as peak time and height, either at the national and regional level. Both the methodology and results are now published in Alaimo Di Loro et al. (2021a). The choice of including a modeling attempt based on the likelihood maximization in a thesis on Bayesian models is twofold: first, I wanted to provide a comprehensive picture of our whole research work up to date, by including all the steps that guided us through the understanding (yet ongoing) of COVID-19 epidemiological features; eventually, the inclusion of such attempt, makes even more clear the advantages of adopting a Bayesian hierarchical modeling approach to describe the complexities of COVID-19 dynamics more accurately.

Indeed, this first work presented some limitations, as regions were assumed to be independent among each other. Moreover, likelihood maximization led to results that happened to be unstable, possibly due to difficulties in finding a global optimum

for the likelihood of an inherently non-linear model. For these reasons, we developed the Bayesian counterpart of the proposal in Alaimo Di Loro et al. (2021a), including a network structure to deal with spatial dependence, and an auto-regressive process to take into account the time dependence. We analyzed data at the regional level and, interestingly enough, proved that substantial spatial and temporal dependence occurred in both epidemic outbreaks. Unsurprisingly, accurate predictions were obtained, improving those of the model where independence across regions was assumed, leading to the second publication by Mingione et al. (2021a).

In parallel, our research group devoted energies for the correct statistical communication of COVID-19 epidemic data to the general public. We aimed at informing properly the general public on the daily evolution of the epidemic. That was achieved by deploying an interactive tool (web application) freely accessible at `https://statgroup19.shinyapps.io/Covid19App/`. We wished to contribute in the process of risk literacy of the population to make them capable of distinguishing relevant information from harmful and dangerous misinterpretations. The app includes basic summaries of the epidemic indicators, easily readable interactive graphs and maps. All source codes are public and freely accessible at `https://github.com/minmar94/StatGroup19`, in the spirit of a completely *Open Data* community.

5.1 Nowcasting COVID-19 incidence indicators during the Italian first outbreak

5.1.1 Introduction

Italy has been the first European country to be severely hit by the first epidemic wave due to the spread of the SARS-CoV-2 virus. COVID-19 syndrome emerged in northern Italy in February 2020, with a basic reproduction number R_0 between 2.5 and 4 (Flaxman et al., 2020). In its most severe form, COVID-19 has two challenging characteristics (Peeri et al., 2020): it is highly infectious and, despite having a benign course in the vast majority of patients, it requires hospital admission and even intensive care for about 10% of infected (Hu et al., 2020). During the outbreak, it was crucial to set up appropriate data collection and modeling systems quickly. Both were necessary for monitoring infections evolution, evaluation of policy interventions, and prediction.

Generally speaking, the nature of epidemics' spread has nearly always followed the same scenario: first, the growth in the number of infected people is (close to) exponential; in a second moment, this growth gradually but consistently slows down as an effect, for instance, of various containment measures. This pattern can cyclically recur until the outbreak is tamed.

So far, a number of mathematical and statistical models of different complexity levels have been used to explain the spread of epidemics and predict their consequences. The starting point is often the Verhulst logistic equation (Liang, 2020), which can easily capture both the exponential increase in the number of infected people at the initial stage of the epidemic development and the tendency towards a constant value by its ending. In more complex models, people are divided into different groups: (S) the susceptible class, namely those individuals who are capable of contracting the disease and becoming infected; (I) the infected class, namely those individuals who are capable of transmitting the disease to others; (R) the removed class, namely infected individuals who are deceased or have recovered, who are either permanently immune or isolated. This group of mathematical models are called

SIR (or compartmental) models (Diekmann et al., 2013). References include Chen et al. (2020); Giordano et al. (2020); Gatto et al. (2020); Dehning et al. (2020), and several more. However, whilst being potentially very appropriate to model the true dynamics underlying any epidemic, the SIR-based models rely on accurate initial estimates of several quantities governing its spreading mechanism (which are unknown). Poor data input on key features of the pandemic can heavily bias these estimates, jeopardizing the reliability of any theory-based forecasting effort. SIR models are micro-simulation models and we believe that, generally speaking, they should be used mostly for "scenario evaluation" rather than predicting future outcomes. Indeed, they rely on several speculations and strict theoretical assumptions, not necessarily met by the analysed data and, especially during the first stage of the outbreak, failed in predicting various COVID-19 related outcomes (Ioannidis et al., 2020). Such specifics lead the choice of coefficients in the equations defining the SIR model and define its initial conditions. It is well known that even a slight change in those can lead to large differences in the final results. For instance, at the beginning of the epidemic, early data providing estimates for case fatality rate, infection fatality rate, basic reproductive number, and other key numbers that are essential for the modeling, are often inflated and may cause potentially large over-estimation of the epidemic severity. Similar criticism to using compartmental modeling for nowcasting can also be found in Baek et al. (2020), and references therein. Hence, we followed an alternative approach, which involved direct modeling of the observed counts (Salje et al., 2020). This encompasses the use of phenomenological models without detailed mechanistic foundations, but which have the advantage of allowing simple calibrations to the empirical reported data. Such approaches are particularly suitable when substantial uncertainty tarnishes the epidemiology of an infectious disease, including the potential contribution of multiple transmission pathways. In these situations, phenomenological models provide a starting point for obtaining early estimates of the transmission potential and short-term forecasts of the epidemic evolution (Chowell et al., 2016).

We propose a parametric regression for modeling the *incidence indicators* based on the use of the Richards' curve (a generalized logistic function) as a response function in place of the widely used exponential or polynomial trend. Furthermore, we replace the generally entrenched Gaussian assumption for the distribution of log-counts (Grasselli et al., 2020; Sebastiani et al., 2020) by the more appropriate Poisson or Negative Binomial distributions for counts. In this way, we avoid the implausible assumptions stemming from the more common alternatives: the former allows the underlying counts to potentially grow indefinitely; the latter neglects the proper specification of dependence between mean and variance under the log-normal distribution. We further propose different ways of including the effect exogenous information on the response function of counts in an extended generalized linear model framework. These models have been implemented during the outbreak with the aim of modeling the medium to long term evolution of the epidemic wave. The use of logistic-based curves is also widely discussed in the literature (Cabras, 2020; Girardi et al., 2020b; Ritz et al., 2015). Logistic growth curves can be seen as a flexible formulation for approximating a large variety of growth phenomena, especially in biology and in epidemiology (Hsu et al., 1984; Grossman and Bohren, 1985; Morris and Silk, 1992; Berkson, 1944; Wachenheim et al., 2003). In particular, highly flexible parametric models such as Gompertz curves and the unified Richards' family (Tjørve and Tjørve, 2010) have been proposed in the study of organisms' growth, for a review see Tjørve and Tjørve (2017).

5.1.2 Available data

The Italian Civil Protection Department (CPD), starting from February 24th, 2020, has been gathering data at the regional level every day and making these public in a `GitHub` repository. During most of the Italian epidemic, data were commented by the department head in an official press release at about 6 p.m. The daily updated data are currently stored at `https://github.com/pcm-dpc/COVID-19`. For public health service purposes, Italy is divided into 21 regions. There are 19 administrative regions, plus two autonomous provinces (Trento and Bolzano) that form the administrative region of Trentino-Alto-Adige. In the sequel, we focus on modeling the indicators aggregated at the national level. Nevertheless, we also tested our procedure for each Italian region to evaluate the graphical and quantitative performances of the proposed model. Results of this analysis are presented in Figures C.8 and C.9.

Incidence and prevalence indicators: different mathematical features

The epidemiological data provided by CPD can be distinguished into two basic types: incidence indicators (flows) and prevalence indicators (stocks).

Incidence indicators

Incidence indicators measure the number of individuals with a particular condition, related with the epidemic, recorded during a given period. They can be referred to different time periods; in particular, in the CPD data set, **daily incident counts** are available for the following indicators: positives, which are sub-classified into hospitalized (either in regular wards or in intensive care) and isolated-at-home; deceased; recovered/discharged.

These indicators can be considered, by analogy with the terminology used in econometrics, as **flow data**, quantifying the daily input (positives) and output (deceased and recovered/discharged) of the system. Figure 5.1 shows the time series of *daily positives* and *daily deceased*[1] aggregated at the national level.

From the viewpoint of the following modeling effort, one important feature of these indicators is that they can be referred to longer time intervals, simply cumulating them over time. The most interesting **cumulative incidence indicators** are those referring to the whole history of the pandemic, computed from a conventional date of "beginning of the pandemic" (typically, the day the systematic recording of daily positives began) to the current day: cumulative positives, cumulative deceased and cumulative recovered/discharged. In particular, given $Y_0 = 0$, we can build the whole series of cumulative counts conditionally on the value of the cumulative indicator at time $(t-1)$, and the incidence indicators at time t, for each $t = 1, \dots, T$:

$$Y_t^c = Y_{t-1}^c + I_t,$$

where Y_t^c represents the cumulative indicator and I_t represents the inputs in the system, e.g.: cumulative positives at time t are the cumulative positives at time $(t-1)$ plus the daily positives at day t. By their nature of cumulative counts, these data series are necessarily monotonically non-decreasing (see Figure 5.2).

[1] The red dot highlights a data entry error, i.e. the recording of a negative count. Same as Figure 5.2.

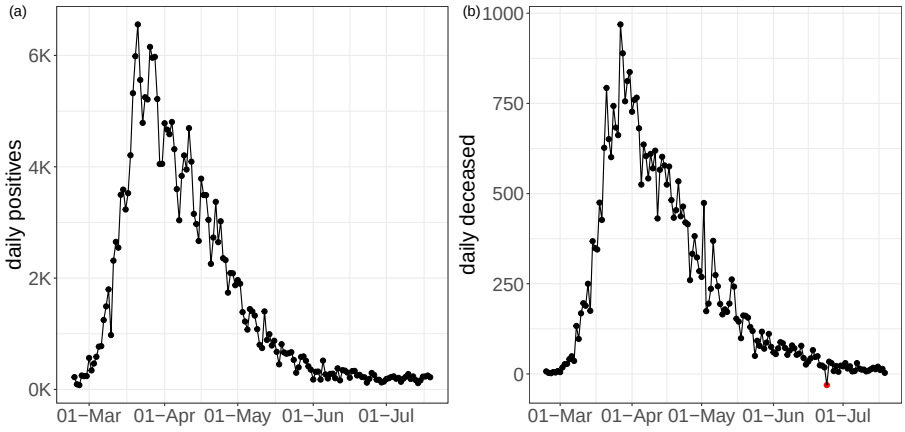


Figure 5.1. Time series of the Italian daily incidence indicators: *daily positives* (left panel) and *daily deceased* (right panel).

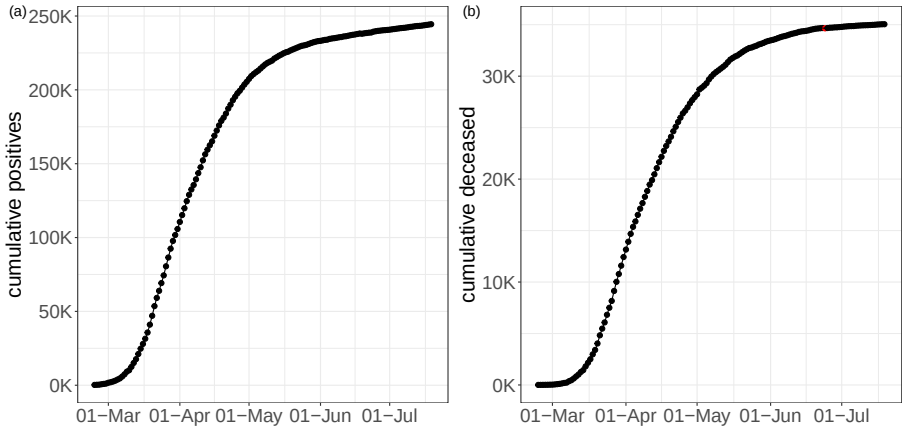


Figure 5.2. Time series of the Italian cumulative incidence indicators: *cumulative positives* (left panel) and *cumulative deceased* (right panel).

Prevalence indicators

Prevalence indicators measure the number of individuals with a particular condition, related with the epidemic, at a given instant in time (or at a given short interval of time, e.g. a day). They are typically obtained from simple algebra from other indicators; in particular, in the CPD data set, the following indicators are available daily: current positives, and current *Intensive Care Units* (ICU) occupancy. These indicators result from the balance between total inputs and outputs of the system, e.g.: current positives are the difference between cumulative positives and cumulative deceased plus recovered/discharged. Again, by analogy with the terminology used in econometrics, they can be considered as **stock data**. In particular, given $Y_0 = 0$, we can build the whole series conditionally on the value of the prevalence indicator at time $(t-1)$, and the incidence indicators at time t, for each $t = 1, \ldots, T$:

$$Y_t^p = Y_{t-1}^p + I_t - O_t,$$

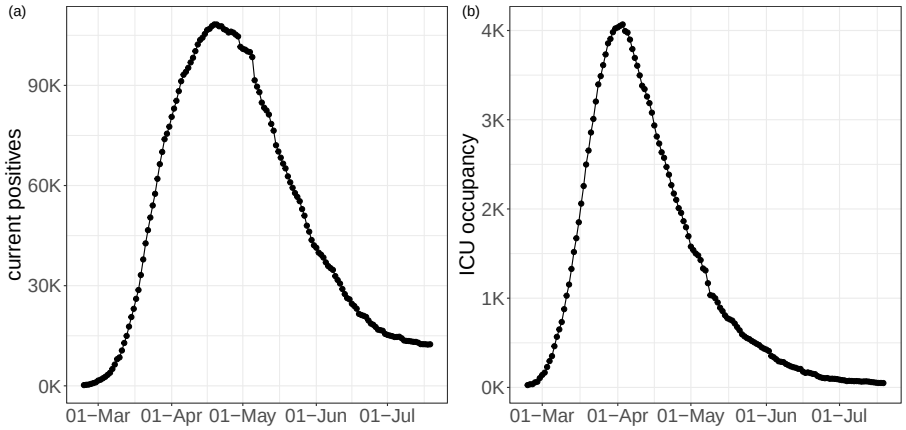


Figure 5.3. Time series of Italian daily prevalence indicators: *current positive* (left panel) and *ICU occupancy* (right panel).

where Y_t^p represents the prevalence indicator, I_t represents the inputs in the system and O_t represents the outputs, e.g.: current positives at time t are the current positives at time $(t-1)$ plus the daily positives at day t and minus the sum of deceased and discharged recovered at day t. However, given the different delay in reporting the various information by the regional agencies, there exists a relevant temporal misalignment among all the quantities reported at the daily scale. Therefore, the simultaneous consideration of all these flows may be significantly flawed and we rather prefer modeling the indicators individually.

Two important features of these indicators are that:

1. given their *stock* nature, they cannot be aggregated (e.g.: it does not make sense to compute "cumulative current positives");

2. by their own nature, these indicators are not monotone, since they can increase or decrease as a result of different trends of the component series. Typically, we expect the series of current positives and ICU occupancy to increase in the rising phase of an epidemic, reach a peak and then decrease to a lower asymptote (see Figure 5.3), although more complex patterns due to resurgence of the epidemic are also plausible.

Prevalence indicators are characterized by a strong and tangled dependence structure which is cumbersome to simplify into a manageable and useful statistical model on the short run. For this reason, the focus of this work concerns only incidence indicators. Our model proposal, from a strictly mathematical point of view, could potentially applied also on prevalence indicators. However, from the statistical point of view, the modeling assumptions which are assumed to hold (with good approximation) considering the incidence indicators, are likely to be strongly violated by prevalence indicators and the resulting outcome cannot be considered reliable. A brief discussion about some possible approaches for the analysis of prevalence indicators is given in Section 5.1.5.

Data issues

COVID-19 public Italian data present several issues that severely affect their quality. The information has been gathered and reported at a regional level, and

each regional healthcare organization has a different transmission and data collection system[2]. Measurement errors and errors in data entry are expected to be often present. Delays in reporting has been, sometimes, substantial. Some patients were transferred (e.g., from Lombardia to Puglia, and even to Germany) without notification, and they were counted as hospital patients of the receiving region (or not at all when sent abroad) and positive cases of the region of residence. Most importantly, counts were updated on the notification day rather than aligned to a more appropriate date. For example, death is counted on the day of the reporting, not on the day of the outcome, which could be even weeks before. Positive status is also counted on the day that test results are received, with swabs being processed from one day to weeks after symptoms' onset. No distinction between actively symptomatic and asymptomatic patients was made.

Swabs and positive cases are not time-aligned. For example, in countries like Singapore (https://www.moh.gov.sg/covid-19), daily data include information on total swabs tested, total unique persons swabbed as well as total swabs per 1,000,000 total population and total unique persons swabbed per 1,000,000 total population. In Italy, up to the 19-th of April 2020, only the total number of daily swabs is available, and no linkage between swabs and tested individuals was kept in the data repository. Hence, it is impossible to make statistically sound use of swabs' count to model the whole first pandemic wave.

Finally, it is crucial to recall that people diagnosed with COVID-19 disease are only a small fraction of the people infected by the virus. Moreover, since the tracking was highly symptoms driven, especially in the first phase of the outbreak, the detected number of positives cases can provide only a partial estimate of the *true* incidence of COVID-19 in the Italian population. Eventually, we expect this detected fraction to vary wildly over space and time.

In our opinion, the most reliable indicator is the count of ICU occupancy. The reason is that the Italian Society for Emergency Care issued national guidelines (that did not change substantially during the epidemic) for testing patients with a suspected infection by SARS-CoV-2, who also had top priority for swab access and reporting; and ICU admissions can be expected to depend on the proportion of infected population susceptible to severe infection, rather than on the regional strategy for testing and contact tracing. However, while probably reliable, this indicator also presents some drawbacks. First of all, it provides only a partial snapshot of the epidemic's current stage, which concerns the most severe cases of the disease. The latter is a critical issue, especially in the COVID-19 case, which is known to present severe symptoms only in a small percentage of the currently affected individuals. Second, this snapshot is affected by a constant delay (i.e., the time between catching the disease and manifesting severe symptoms). As mentioned above, its daily variation is obtained as a combination of new incoming patients (+) and the deceased or recovered ones (-), whose effects blend and are hard to disentangle. As a consequence, *incidence indicators*, such as *daily positives* and *daily deceased*, while being measured with some error and even more delay in the case of deaths, still represent the critical indicators for timely and appropriate monitoring of the pandemic.

[2] see `https://www.epiprev.it/materiali/2020/EP2-3/112_edit1.pdf` for further details.

5.1.3 Methodology

The time series of any of the observed indicators, denoted by $\mathbf{z} = \{z_t\}_{t=t_0}^{T}$, is modeled separately and considered as the realization of the stochastic process $\mathbf{Z} = \{Z_t\}_{t=t_0}^{T}$. The idea behind this paper is to model any of the mentioned indicators through a generalized model with a response function $\mathbb{E}[Z_t] = \mu(t) = g^{-1}(t;\boldsymbol{\theta})$, where $g(\cdot)$ is a known link function and $\boldsymbol{\theta}$ is a parameter vector, that is appropriate for the specific mathematical features of the epidemic process. This must be coupled with a response distribution $f(Z_t;\theta)$ coherent with the domain of such indicators, which are counts and therefore Natural numbers.

Response function for incidence indicators

Let us denote by $\{y_t^c\}_{t=0}^{T}$ the time-series of cumulative incidence indicators since the start of the epidemic ($t_0 = 0$, first day of systematic data recording). Visual inspection of these indicators in Figure 5.2 suggests that their expected values follow a logistic-type growth curve. Different example of logistic curves have been proposed in the literature, all representing solutions to specific differential equations that model the spread of epidemics (Banks, 1993; Hsieh et al., 2010; Ma et al., 2014). Differently from the more standard exponential models, these are able to describe the slow down of the outbreak associated with a decaying transmission rate just after the number of cases approaches its inflection point. They have been already widely used to describe the evolution of the COVID-19 pandemic in different states during its early to medium stage (Wu et al., 2020; Lee et al., 2020). Here, for all the incidence indicators, we consider the *Generalized Logistic Function*, also known as *Richards'* curve (see Figure 5.4 as an example), as response function for the mean of the process (Richards, 1959). This curve was widely used to describe various biological processes (Werker and Jaggard, 1997), but has been recently adapted also in epidemiology for real-time prediction of outbreak of diseases (Hsieh, 2009; Hsieh and Chen, 2009; Hsieh, 2010). The specialty of the Richards' curve lies in its ability to describe a great variety of growing processes, endowed with strong flexibility, that includes as special cases the standard logistic growth curve (Tsoularis and Wallace, 2002), the Gompertz growth curve (Gompertz, 1825) and others. It can be expressed in different forms (Causton, 1969; Birch, 1999; Kahm et al., 2010; Cao et al., 2019). One of its most general formulation depends on the vector of 5 parameters $\boldsymbol{\gamma}^\top = [b, r, h, p, s]$ and can be expressed as:

$$\mathbb{E}[Y_t^c] = g^{-1}(t;\boldsymbol{\gamma}) = \lambda_{\boldsymbol{\gamma}}(t) = b + \frac{r}{(1 + 10^{h(p-t)})^s}. \tag{5.1}$$

$b \in \mathbb{R}^+$ represents a lower asymptote and $r > 0$ is the distance between the upper and the lower asymptote, hence $b + r$ would be the final epidemic size; h is known as the *hill*, and represents the infection/growth rate; $p \in \mathbb{R}$ represents a lag-phase of the trajectory and determines the peak position (it tells when the curve growth speed slows down); $s \in \mathbb{R}$ is an asymmetry parameter regulating differences in the behaviour of the ascending and descending phase of the outbreak. In our context, since cumulative incidences are always monotone increasing indicators, it is reasonable to assume $h, s > 0$[3].

An extensive review of the Richards' curve and other logistic growth models, together with discussion on the proper interpretation of the parameters, is given in Tjørve and Tjørve (2010).

[3]Conversely, we may assume $h, s < 0$

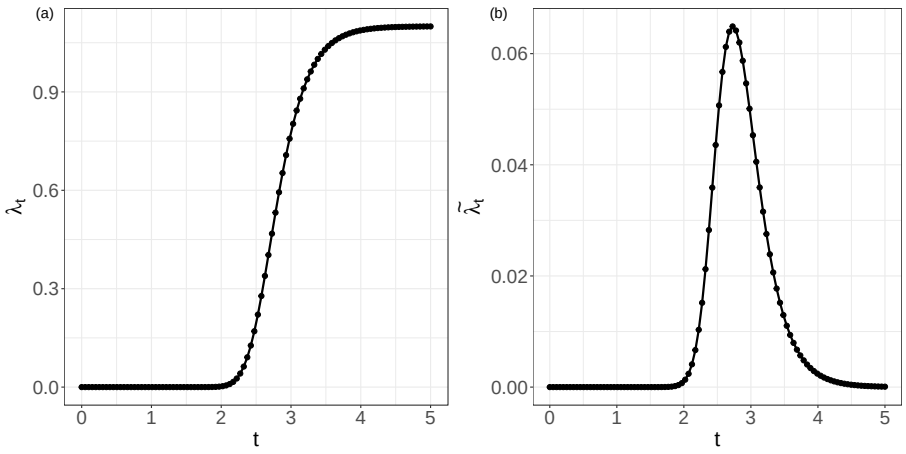


Figure 5.4. Example of Richards' curve (left panel) and derivative of the Richards' curve (right panel).

An *Extended Generalized Linear Model* using Equation (5.1) as response function seems to be a natural choice for modeling time series of *cumulative counts*, whose monotonically non-decreasing average behaves as the Richards' curve. Unfortunately, there is a significant drawback to this choice. As it will be better clarified thereafter, a very useful working assumption would be that all these counts were stochastically independent, given their mean function $\lambda_{\boldsymbol{\gamma}}(t)$. However, we cannot consider this assumption as realistic in the case of cumulative counts, since the constraint on the domain of definition on subsequent counts (i.e., $y_t^c \geq y_\tau^c, \forall \tau < t$) is not guaranteed to be satisfied. On the other hand, the stochastic independence assumption sounds more reasonable, albeit not necessarily true, for the *daily incident counts* $\{y_t\}_{t=1}^T$, i.e., the addenda of the cumulative counts excluding the starting point y_0, which can be defined as:

$$y_t^c = \sum_{\tau=0}^{t} y_\tau \quad \Rightarrow \quad y_t = y_t^c - y_{t-1}^c, \qquad t = 1, \ldots, T$$

where $y_0 = 0$ by definition.
Using Equation 5.1, and exploiting the additive properties of the expected value, we have:

$$\begin{aligned}\tilde{\mu}(t) = \mathbb{E}[Y_t] = \mathbb{E}[Y_t^c] - \mathbb{E}[Y_{t-1}^c] = \lambda_{\boldsymbol{\gamma}}(t) - \lambda_{\boldsymbol{\gamma}}(t-1) = \\ = r \cdot \left[(1 + 10^{h(p-t)})^{-s} - (1 + 10^{h[p-(t-1)]})^{-s}\right] = \tilde{\lambda}_{\boldsymbol{\gamma}}(t)\end{aligned}$$

which, in particular, does not depend on the baseline b. Therefore, we shall adopt an extended generalized model with response function given by the first differences of the Richards' curve $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{\gamma}} = \left\{\tilde{\lambda}_{\boldsymbol{\gamma}}(t)\right\}_{t=1}^T$ to model the daily expected values $\boldsymbol{\mu} = \{\mu(t)\}_{t=1}^T$ of the observed incident counts $\mathbf{y} = \{y_t\}_{t=1}^T$ (see example in Figure 5.4).

In addition, we may also consider adding a kink effect/baseline α to the first differences $\tilde{\lambda}_{\boldsymbol{\gamma}}(\cdot)$, which is to say assuming the following functional form for the mean of the daily counts:

$$\tilde{\mu}_{\boldsymbol{\theta}}(t) = \alpha + \tilde{\lambda}_{\boldsymbol{\gamma}}(t), \quad \alpha \geq 0, \tag{5.2}$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\gamma})$. This would correspond to the following mean function for the cumulative counts:

$$\mu_{\boldsymbol{\theta}}(t) = \alpha \cdot (t-1) + \lambda_{\boldsymbol{\gamma}}(t).$$

In practice, the parameter α includes the possibility of having a strictly positive baseline rate, which can be interpreted as the *endemic steady state incidence rate*. This is in line with the current perspective that SARS-CoV-2 might not be completely eradicated within the next few years (Shaman and Galanti, 2020). On the other hand, the first differences of the Richards' curve $\tilde{\lambda}_{\boldsymbol{\gamma}}(t)$ are (by construction) forced to decrease asymptotically to the value of 0. However, this asymptotic result is not necessarily observed in real data. In particular, Figure 5.1 highlights that both time-series do not attain the 0 value, but settle to a low, constant level. This situation may, potentially, continue indefinitely: new cases will be found as long as people will be tested. Consequently, the model without a baseline lacks the ability to catch this tail and, because of the curve parametric form, this may indirectly affect the fit on the whole series.

In the first instance, one solution would be to fit the model, including the kink effect α. Afterward, if it is estimated not to be sensibly different from 0, the model without α can be fitted again to stabilize the estimation procedure and decrease the uncertainty on the other parameters.

Response distribution for incidence indicators

Before introducing the distributions for the daily incident counts, we must make some assumptions about the time dependence structure. In particular, we assume that given the mean function $\tilde{\mu}_{\boldsymbol{\theta}}(t)$, the daily incident counts Y_t are stochastically independent from the previous cumulative counts: $Y_t \perp Y_\tau^c \; \forall \, \tau < t$. We denote this hypothesis of independence by **HI**. We also assume the value of the first cumulative count $Y_0^c = y_0^c$ to be known and fixed. Exploiting **HI**, we can express the joint density of all the subsequent cumulative counts conditional on $Y_0^c = y_0^c$ as the product of the univariate densities of the corresponding daily counts $\{Y_t\}_{t=1}^T$. The equivalence follows from the following conditional argument:

$$\begin{aligned} f_{Y_1^c,\ldots,Y_T^c}(y_1^c,\ldots,y_T^c|y_0^c;\theta) &= \prod_{t=1}^{T} f_{Y_t^c}(y_t^c|y_0^c,\ldots,y_{t-1}^c;\theta) = \\ &= \prod_{t=1}^{T} f_{Y_t^c}(y_t + y_{t-1}^c|y_0^c,\ldots,y_{t-1}^c;\theta) = \\ &= \prod_{t=1}^{T} f_{Y_t}(y_t|y_0^c,\ldots,y_{t-1}^c;\theta) \overset{\textbf{HI}}{=} \prod_{t=1}^{T} f_{Y_t}(y_t|\theta), \end{aligned}$$

where the second identity is justified in the light of $Y_t^c = Y_t + Y_{t-1}^c,\ t = 1,\ldots,T$, which is true by definition. From a practical point of view, this also implies a 1-st order Markov property for the cumulative counts:

$$Y_t^c|Y_{t-1}^c \perp Y_1^c,\ldots,Y_{t-2}^c, \quad t = 1,\ldots,T$$

and mutual independence between the daily counts:

$$Y_t \perp Y_\tau, \quad \forall \, t,\tau, t \neq \tau.$$

We remark that although these independence structure is just an approximation in the present case, this kind of approach has provided valid inference for all the

available Italian incidence indicators.
For communication purposes, it can be of interest to report the results of analyses and predictions in terms of cumulative, rather than daily, incidence indicators. Clearly, it is possible to model and predict the daily incidence indicators and, from these estimates and predictions, obtain the relevant cumulative incidence indicators.

Poisson distribution

Let us assume that the vector of daily incident counts, $\mathbf{y} = \{y_1, \dots, y_t\}$, is composed of independent Poisson realizations with expected value $\tilde{\mu}_{\boldsymbol{\theta}}(t)$:

$$Y_t|\boldsymbol{\theta} \sim Pois(\tilde{\mu}_{\boldsymbol{\theta}}(t)), \quad t = 1, \dots, T$$

Hence, the likelihood can be written as:

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) = \prod_{t=1}^{T} Pois(y_t|\tilde{\mu}_{\boldsymbol{\theta}}(t)) \propto \tilde{\mu}_{\boldsymbol{\theta}}(t)^{\sum_{t=1}^{T} y_t} \cdot \exp\left\{-\sum_{t=1}^{T} \tilde{\mu}_{\boldsymbol{\theta}}(t)\right\}$$

and the log-likelihood is given by:

$$l(\boldsymbol{\theta}|\mathbf{y}) = \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) \propto \sum_{t=1}^{T} y_t \log\left(\tilde{\mu}_{\boldsymbol{\theta}}(t)\right) - \sum_{t=1}^{T} \tilde{\mu}_{\boldsymbol{\theta}}(t).$$

Remark that, under the assumption of Poisson distribution and the baseline $\alpha = 0$ (i.e. $\tilde{\mu}_{(\alpha,\boldsymbol{\gamma})} = \tilde{\lambda}_{\boldsymbol{\gamma}}(\cdot)$), we can exploit the well-known Poisson's additive property[4] to conclude that each cumulative count Y_t^c is still marginally distributed according to a Poisson, parametrized by the original Richards' curve function $\lambda_{\boldsymbol{\gamma}}(\cdot)$:

$$Y_t^c|\boldsymbol{\gamma} \sim Pois\left(\sum_{\tau=1}^{t} \tilde{\lambda}_{\boldsymbol{\gamma}}(\tau)\right) = Pois(\lambda_{\boldsymbol{\gamma}}(t)).$$

Negative Binomial distribution

When counts are over-dispersed the Poisson distribution is not a suitable choice. We can model the observed daily incident counts $\mathbf{y} = \{y_1, \dots, y_t\}$ as independent realizations from a Negative Binomial with mean $\tilde{\lambda}_{\boldsymbol{\gamma}}(t)$ and dispersion parameter $\nu \in \mathbb{R}^+$:

$$Y_t|\boldsymbol{\theta} \sim NB(\tilde{\mu}_{\boldsymbol{\theta}}(t), \nu), \quad t = 1, \dots, T$$

Hence, the likelihood can be written as:

$$\mathcal{L}(\boldsymbol{\theta}, \nu|\mathbf{d}) = \prod_{t=1}^{T} NB(y_t|\tilde{\mu}_{\boldsymbol{\theta}}(t), \nu) \propto \prod_{t=1}^{T} \left[\frac{\Gamma(\nu + y_t)}{\Gamma(\nu)} \left(\frac{\nu}{\nu + \tilde{\mu}_{\boldsymbol{\theta}}(t)}\right)^{\nu} \left(\frac{\tilde{\mu}_{\boldsymbol{\theta}}(t)}{\nu + \tilde{\mu}_{\boldsymbol{\theta}}(t)}\right)^{y_t}\right]$$

and the log-likelihood is:

$$l(\boldsymbol{\theta}, \nu|\mathbf{y}) = \log \mathcal{L}(\boldsymbol{\theta}, \nu|\mathbf{y}) \propto \sum_{t=1}^{T} \log\left(\frac{\Gamma(\nu + y_t)}{\Gamma(\nu)}\right) + \nu \sum_{t=1}^{T} \log\left(\frac{\nu}{\nu + \tilde{\mu}_{\boldsymbol{\theta}}(t)}\right) + \sum_{t=1}^{T} y_t \log\left(\frac{\tilde{\mu}_{\boldsymbol{\theta}}(t)}{\tilde{\mu}_{\boldsymbol{\theta}}(t) + \nu}\right).$$

[4] the sum of independent Poissons is still a Poisson with parameter the sum of the parameters

The Negative Binomial does not satisfy the same additive property as the Poisson, hence we cannot draw the same conclusion reached in the Poisson case about the marginal distribution of the cumulative count Y_t^c when $\alpha = 0$. In general, the cumulative count in the Negative Binomial case will follow the distribution stemming from the sum of independent Negative Binomial random variable with common dispersion parameter ν but different means $\tilde{\boldsymbol{\mu}} = \left\{\tilde{\lambda}_{\boldsymbol{\gamma}}(t)\right\}_{t=1}^{T}$.

Response function depending on covariates

The trend of any of the considered indicators may also depend on additional exogenous information, which we may assume to be known *a priori* either because it is immutable (i.e., the day of the week), or because policymakers fixed it (daily number of *tested cases/swabs* set by the government). For instance: one might want to correct for possible weekly seasonality, which is known to affect the *daily positives* series since many laboratories are closed during the weekend and cannot process swabs. The latter can be used to disentangle the underlying trend of the epidemic from the obvious positive correlation between *tested cases* and *daily positives*. In general, we may want to include the effect of any set of k time-varying covariates $\underset{T\times(k+1)}{\mathbf{X}} = [\mathbf{x}(t)]_{t=1}^{T}$ in the Richards' framework through the usual linear predictor $\eta(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a $k+1$-dimensional vector of real valued parameters (including intercept). Let us denote the mean function of the considered indicator as $\tilde{\mu}_{\boldsymbol{\theta}}(t) = \mathbb{E}[Y_t]$, where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\gamma}, \boldsymbol{\beta})$. In order to respect the positivity of the mean parameter (which is necessary both in the Poisson and in the Negative Binomial case), we consider the link function $g(\cdot) = \log(\cdot)$, so that the effect on the mean is expressed as:

$$\mu_{\boldsymbol{\beta}}(\mathbf{X}) = \exp\{\eta(\mathbf{X})\} = \exp\{\mathbf{X}\boldsymbol{\beta}\}.$$

Considering a single time point t, we would get the following functional form:

$$\mu_{\boldsymbol{\beta}}(\mathbf{x}(t)) = \exp\{\mathbf{x}(t)\boldsymbol{\beta}\}.$$

The mean term of our model shall take into account both the effect of the covariates through $\mu_{\boldsymbol{\beta}}(\cdot)$ and the temporal behaviour induced by the Richards' curve $\lambda_{\boldsymbol{\gamma}}(\cdot)$. As a matter of fact, these two components may be combined in different ways. We considered two alternative specifications denoted in the sequel as: *additive* and *multiplicative*.

Additive inclusion of covariates

The inclusion of an additive effect of covariates implies that the effect of every covariate is constant through-out the pandemic, notwithstanding the current contagion level: for instance, one may think that an increase of daily *tested cases* will always produce the same increase of *daily positives*. If that is the case, we may just express the baseline parameter α at each time-point t as the *linked* linear combination of covariates $\mu_{\boldsymbol{\beta}}(\mathbf{x}(t)) = \exp\{\mathbf{x}(t)\boldsymbol{\beta}\}$, which would produce the following mean function:

$$\tilde{\mu}_{\boldsymbol{\theta}}(t) = \mu_{\boldsymbol{\beta}}(\mathbf{x}(t)) + \tilde{\lambda}_{\boldsymbol{\gamma}}(t).$$

On the whole vector of observations, this can be expressed as $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = \mu_{\boldsymbol{\beta}}(\mathbf{X}) + \tilde{\boldsymbol{\lambda}}_{\boldsymbol{\gamma}}$.

Multiplicative inclusion of covariates

The inclusion of a multiplicative effect of covariates would imply that the more serious the pandemic situation, the more severe the impact of any covariate on the indicators' daily rate.

First, let us recall that the first differences of the Richards' curve function can be computed as:

$$\tilde{\lambda}_{\boldsymbol{\gamma}}(t) = r \cdot \left[(1 + 10^{h(p-t)})^{-s} - (1 + 10^{h[p-(t-1)]})^{-s}\right] = r \cdot \tilde{\lambda}_{\boldsymbol{\gamma},-r}(t),$$

where $\tilde{\lambda}_{\boldsymbol{\gamma},-r}(t) = \left[(1 + 10^{h(p-t)})^{-s} - (1 + 10^{h[p-(t-1)]})^{-s}\right]$, namely the difference between the Richards' curve function at time t and $t-1$, which does not depend on r. On the log-scale, it would return the more familiar:

$$\log(\tilde{\lambda}_{\boldsymbol{\gamma}}(t)) = \log(r) + \log\left(\tilde{\lambda}_{\boldsymbol{\gamma},-r}(t)\right). \tag{5.3}$$

From Equation 5.3, it comes natural the idea of expressing $\log(r)$ at each time-point t as the linear combination of covariates $\eta(\mathbf{x}(t))$ as in a *Generalized Poisson* model with log link function. Indeed this provides a multiplicative effect of the covariates, where the parameter r can be expressed as $\mu_{\boldsymbol{\beta}}(\cdot)$ in the following way:

$$r_{\boldsymbol{\beta}}(\mathbf{x}(t)) = \mu_{\boldsymbol{\beta}}(\mathbf{x}(t)) = \exp\{\mathbf{x}(t)\boldsymbol{\beta}\}. \tag{5.4}$$

Note that the constant r is still present and included in Equation 5.4 through the intercept β_0. Therefore, the mean at time t is expressed as:

$$\tilde{\mu}_{\theta}(t) = \alpha + r_{\boldsymbol{\beta}}(\mathbf{x}(t)) \cdot \tilde{\lambda}_{\boldsymbol{\gamma},-r}(t).$$

Considering the whole vector of observations, we would have the following vector of means $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = \boldsymbol{\alpha} + r_{\beta}(\mathbf{X}) \cdot \tilde{\boldsymbol{\lambda}}_{\boldsymbol{\gamma},-r}$, where $\boldsymbol{\alpha} = \alpha \cdot \mathbf{1}_T$.

Model estimation

Parameters can be estimated by maximizing the log-likelihood $l(\boldsymbol{\theta}|\mathbf{y})$, where $\boldsymbol{\theta}$ in this case includes all the parameters the likelihood depends on (e.g. including ν in the Negative Binomial case). This optimization problem does not have an analytical solution, and numerical maximization must be used. To improve computation, we derived analytical expressions for the gradient and Hessian of the two possible log-likelihoods (i.e. Poisson or Negative Binomial counts), making Fisher-scoring iteration very fast. The expressions are reported in Appendix C. Given the non-smooth shape of the objective function, we are at risk of being trapped by local maxima of the log-likelihood, depending on the initial conditions. Therefore, in order to strengthen the optimization procedure, a multi-start procedure based on a combination of genetic and gradient descent algorithms has been used (Scrucca, 2013; Nash et al., 2020).

Once an approximate point of maximum $\hat{\boldsymbol{\theta}}$ has been obtained, we could theoretically obtain an estimate of the asymptotic variance-covariance matrix of the estimated parameters through inverse of the negative log-likelihood Hessian in $\hat{\boldsymbol{\theta}}$ (which corresponds to the *Observed Fisher Information*):

$$\hat{V}_{\boldsymbol{\theta}} = -\mathbf{H}\left(l(\hat{\boldsymbol{\theta}}|\mathbf{y})\right)^{-1},$$

where **H** denotes the Hessian matrix. Nevertheless, we may want to account for the potential misspecification of our model potentially arising from the independence assumption **HI**. Therefore, we resort to a robust approach for estimating the standard errors and covariance structure associated with the parameter vector $\boldsymbol{\theta}$. In particular, we consider the *Huber Sandwich Estimator* of the variance-covariance matrix (Hardin, 2003; Freedman, 2006), that can be computed as:

$$\hat{V}_{\boldsymbol{\theta}}^{R} = \left(-\mathbf{H}\left(l(\hat{\boldsymbol{\theta}}|\mathbf{y})\right)^{-1}\right) \nabla l(\hat{\boldsymbol{\theta}}|\mathbf{y}) \nabla l(\hat{\boldsymbol{\theta}}|\mathbf{y})^{\top} \left(-\mathbf{H}\left(l(\hat{\boldsymbol{\theta}}|\mathbf{y})\right)^{-1}\right),$$

where $\nabla l(\hat{\boldsymbol{\theta}}|\mathbf{y})$ represents the gradient of the log-likelihood in the point of maximum. Interval estimates for the parameters are directly derived through the asymptotic distribution of the *Maximum Likelihood Estimator*, with the corresponding robust covariance matrix $\hat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\boldsymbol{\theta}, \hat{V}_{\boldsymbol{\theta}}^{R}\right)$. A similar theoretical result for predictions is not as straightforward. Therefore, these are derived through a parametric double bootstrap procedure (Efron, 2004; Hall and Maiti, 2006; Efron, 2012), which accounts for both the uncertainty of parameter estimation and the randomness of the observations. In practice, re-sampled trajectories $\{\mathbf{Y}_i\}_{i=1}^{B}$ are obtained by simulating B sets of parameters from their asymptotic distribution and computing B mean functions trajectories $\{\mu_{\boldsymbol{\theta}_i}(\cdot)\}_{i=1}^{B}$. An artificial time series of counts is then simulated for each of the B trajectories and 95% confidence intervals are obtained by computing the point-wise 2.5% and 97.5% quantiles. The dispersion parameter ν, being a poorly identifiable nuisance parameter of no impact on the mean curve behaviour, has been excluded from the bootstrapping procedure and kept fixed at its estimated value $\hat{\nu}$.

Model validation

Diagnostic check on the model has been performed through the *Pearson residuals* and the *Deviance residuals*. Computation of the former is trivial, where we recall their definition as:

$$\hat{\rho}_t = \frac{y_t - \hat{y}_t}{\widehat{\mathrm{Var}}\left[Y_t\right]}, \qquad t = 1, \ldots, T.$$

Under the *Poisson* and *Negative Binomial* assumptions we have:

$$\widehat{\mathrm{Var}}_{\mathrm{Poi}}\left[Y_t\right] = \mu_{\hat{\boldsymbol{\theta}}}(t), \qquad \widehat{\mathrm{Var}}_{\mathrm{NB}}\left[Y_t\right] = \mu_{\hat{\boldsymbol{\theta}}}(t) + \frac{\mu_{\hat{\boldsymbol{\theta}}}(t)^2}{\hat{\nu}}, \tag{5.5}$$

respectively. The *Deviance Residuals* are instead defined as the individual contributions of each observation to the *Deviance* of the model, i.e. the discrepancy between the proposed model and the full model (perfect fit) fits in terms of *log-likelihood*:

$$\hat{d}_t = 2 \cdot \left[\log\left(f(y_t|\hat{\boldsymbol{\theta}}_s\right) - \log\left(f(y_t|\hat{\boldsymbol{\theta}}\right)\right],$$

where $f(\cdot|\cdot)$ is the chosen distribution function and $\hat{\boldsymbol{\theta}}_s$ is the parameter vector of the saturated model. For the *Poisson* and *Negative Binomial* this can be computed as:

$$\hat{d}_t^{\mathrm{Poi}} = \mathrm{sgn}\left(y_t - \mu_{\hat{\boldsymbol{\theta}}}(t)\right) \cdot \sqrt{2y_t \log\left(\frac{y_t}{\mu_{\hat{\boldsymbol{\theta}}}}\right) - \left(y_t - \mu_{\hat{\boldsymbol{\theta}}}(t)\right)},$$

$$\hat{d}_t^{\mathrm{NB}} = \mathrm{sgn}\left(y_t - \mu_{\hat{\boldsymbol{\theta}}}(t)\right) \cdot \sqrt{2\left[y_t \log\left(\frac{y_t}{\mu_{\hat{\boldsymbol{\theta}}}(t)}\right) - (y_t + \nu) \cdot \log\left(\frac{y_t + \nu}{\mu_{\hat{\boldsymbol{\theta}}}(t) + \nu}\right)\right]},$$

respectively (Svetliza and Paula, 2003). If the model correctly describes the variability in the data, then both the *Pearson residuals* and the *Deviance residuals* are expected to be *Normally distributed* and *independent*, with the latter being generally more robust to outliers.

Fitting performances are further evaluated through numerical metrics such as the pseudo-R^2 and coverage of the 95% prediction intervals:

$$R^2 = 1 - \frac{\text{MSE}}{\sigma_y^2} = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2},$$
$$\overline{\text{Cov}}_{95\%} = \frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{I}_{(\widehat{y}_t^l; \widehat{y}_t^u)}(y_t),$$

where MSE is the *Mean Squared Error*, $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$, $\mathbb{I}_{\mathcal{Y}}(\cdot)$ denotes the indicator function over the set $\mathcal{Y}$, and $\widehat{y}_t^l$ and $\widehat{y}_t^u$ are the lower and upper bounds of the 95% confidence intervals, respectively.

Step-ahead predictions

We test our model's ability to predict the evolution of the epidemic (at least its first wave) from the short to the medium term. Indeed, while the choice of a rigid parametric form for the mean function is penalizing in terms of flexibility and fitting ability, it allows for extrapolation outside the observed domain and is supposed to provide robust forecasts (at least in the short/medium term). Therefore, using the best model for the two indicators we calculated the *out-of-sample* Root Mean Squared Prediction Error (RMSPE) for:

- different fitting windows $t = 1, \ldots, \tilde{t}$;
- different forecast horizons, say $K \in \{1, 5, 10, 15\}$.

We recall that, given the fitting window set $1, \ldots, \tilde{t}$:

$$\text{RMSPE}_{\tilde{t},K} = \sqrt{\frac{1}{K}\sum_{j=1}^{K}(y_{\tilde{t}+j} - \hat{y}_{\tilde{t}+j})^2}$$

5.1.4 Application

For the sake of brevity, here we present results referred to the proposed *Richards' growth model* only for *daily positives* aggregated at the national level. Further results of the model performances for *daily deceased* are included in Appendix C. We only present results obtained adopting the Negative Binomial distribution because of the substantial over-dispersion present at all levels for these indicators (spatially and temporally heterogeneous data collection process, varying containment measures, etc.).

Model on *daily positives*

To decide whether or not to include the kink effect, we fitted the model with and without the baseline α and compared the two fits in terms of *log-likelihood*, *AIC*, *BIC* and *Corrected AIC (AICc)*. The values are presented in Table 5.1 and provide clear evidence in favor of the model with baseline (i.e. with mean $\mu_{\boldsymbol{\theta}}(\cdot)$ as in Equation 5.2).

Table 5.1. *Log-likelihood, AIC, BIC and AICc for the model without baseline and the model with baseline, on daily positives.*

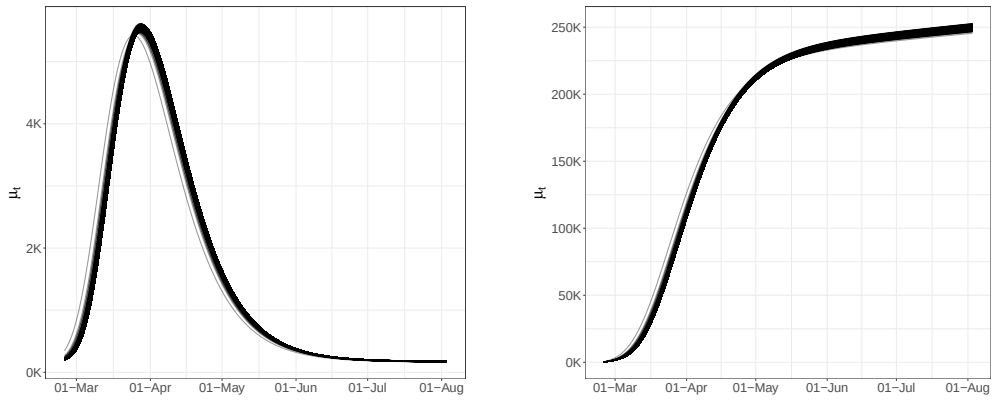
Index	Model without baseline	Model with baseline
log-likelihood	-1081.4	-982.8
AIC	2152.7	1953.6
BIC	2162.3	1965
AICc	2137.8	1935.7

Table 5.2. *Parameters' points estimates and* 95% *confidence intervals for the model with baseline on daily positives.*

Parameter	Point estimate	95% Interval
α	173.17	$(103.2, 290.54)$
r	222.95×10^3	$(220.56 \times 10^3, 225.36 \times 10^3)$
h	0.0288	$(0.0285, 0.0291)$
p	-31.18	$(-32.75, -29.62)$
s	72.54	$(48.29, 96.79)$
ν	18.73	$(17.77, 19.73)$

Parameters' estimates of the model $\hat{\boldsymbol{\theta}}$ and the respective 95% confidence intervals are shown in Table 5.2, where the baseline α is estimated to be $\hat{\alpha} = 173.17$, with interval $(103.2, 290.54)$, which confirms that the baseline is estimated to be significantly different from 0, and it should be included in the model. This parameter represents the long-term endemic incidence rate that may (possibly indefinitely) follow the end of the main outbreaks. This obviously would hold exactly with constant social interactions, containment measures, control of cases etc. Hence, in the considered time horizon, we expect this endemic level to be of ≈ 173 daily positives per day. When the baseline is included, the parameter r does not indicate anymore the final epidemic size, but only the *final outbreak size*. This is the number of positive cases due to the uncontrolled outbreak, additional to what would have been observed in the steady endemic state. This amount is estimated to be $\approx 222,950$, an amount that would have been reached in $\approx 1,288$ days at the endemic state level. The parameters h, p and s do not have an easily quantifiable and absolute interpretation, but are useful for comparison. We recall that h indicates how fast the infection spreads, p how soon it starts descending (lag-phase) and s represents the asymmetry between the ascending and descending phase ($s < 1$ the ascending is slower than the descending and vice-versa). Finally, ν is an over-dispersion parameter and does not present any evident communicable interpretation. The larger it is and the lower the over-dispersion, according to the formula in Equation (5.5).

We here want to stress the fact that the uncertainty characterizing some of the parameters (like s) is not alarming. In particular, variations of s at values distant from 1 have very little effect on the curve shape. Furthermore, the parameter vector presents a covariance structure that highlights how different combination of parameters can yield similar curves. Indeed, simulating $B = 5,000$ set of parameters from the Normal distribution with variance corresponding to the covariance underlying the Huber Sandwich covariance matrix, we get the set of difference and cumulative curves represented in Figure 5.5.



(a) First differences **(b)** Cumulative

Figure 5.5. Bootstrapped trajectories corresponding to the Huber Sandwich covariance matrix in the point of maximum for the model with baseline on *daily positives*.

We can also directly obtain point predictions $\{\hat{y}_t\}_{t=1}^{T}$ as:

$$\hat{y}_t = \mu_{\hat{\boldsymbol{\theta}}}(t), \quad t = 1, \ldots, T, \tag{5.6}$$

and prediction intervals $\left\{(\widehat{y}_t^l; \widehat{y}_t^u)\right\}_{t=1}^{T}$ through the same set of bootstrapped trajectories, whose statistical validity relies on the asymptotic properties introduced in Section 5.1.3.

Figure 5.6 shows the model fit on the whole available time series of counts: the left-side panel shows the fit for the daily time-series, the right-side panel for the cumulative time-series. We can see how the estimated curve does catch the observed general behaviour, providing a smooth approximation, only marginally influenced by extreme values. Our model produces a pseudo-$R^2 = 0.941$ and coverage $\overline{\text{Cov}}_{95\%} = 0.945$, meaning that the percentage of observed daily counts falling inside the estimated bounds is perfectly coherent with the specified confidence level. Looking at Figure 5.6, we notice how daily counts boundaries get smaller as time passes, due to the implicit relationship between mean and variance that characterizes *count* distributions. At the same time, the opposite happens to the bounds on the cumulative counts. The latter is not surprising: indeed, they are built marginally on all the epidemic's possible scenarios. Therefore, they give us a clear sight of what we could have currently observed, keeping into account and aggregating the uncertainty at each stage of the epidemic. We performed a diagnostic check on both the Pearson and the Deviance residuals, but only report the latter for the sake of brevity. The plots in Figure 5.7 show the Deviance residuals behaviour: histogram (a), including the p-value from the Shapiro test; Normal qq-plot (b); auto-correlation plot (c); plot of the residuals vs. fitted values (d). The first two check the (approximated) Normality assumption on the residuals, while the second two control for the correlation of the residuals (among them and with the observed values).

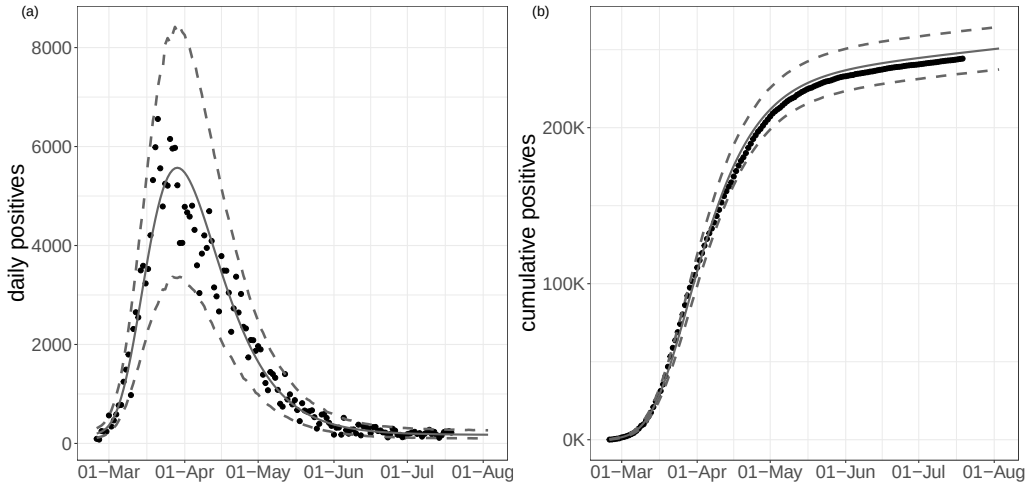


Figure 5.6. Observed (black dots) and fitted values (grey solid lines) with 95% confidence intervals (grey dashed lines) for the model with baseline on *daily positives*.

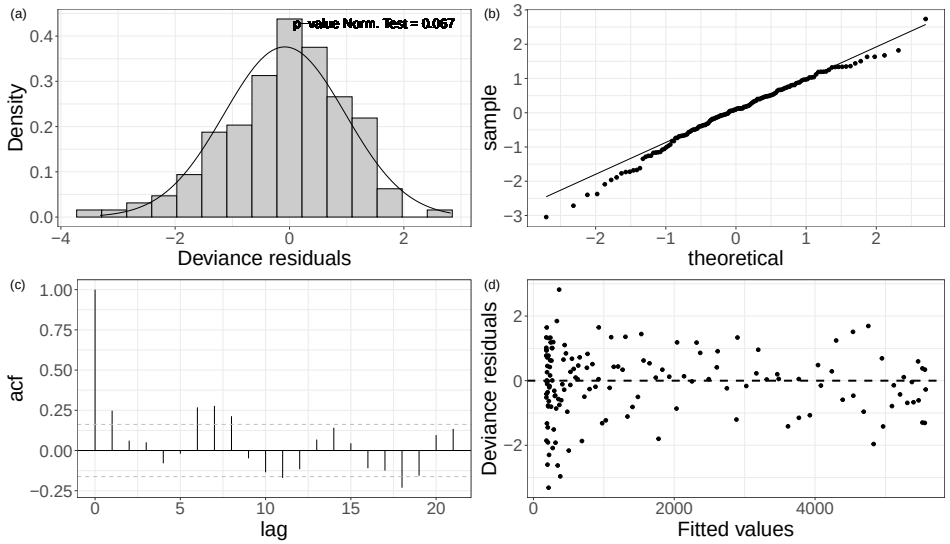


Figure 5.7. Deviance residuals for the model with baseline on *daily positives*.

Weekly seasonality

The diagnostic check on both type of residuals showed that the Normality assumption is not rejected, but the correlation plot manifests undesirable patterns (see Figure 5.7). In particular, the auto-correlation between errors is larger at lag 7 (and multiples of this). We can interpret this outcome as the presence of an intense weekly seasonality (especially during/after the weekend). This suggests people would rather not come forward for testing on the weekend or, alternatively, the system has less capacity at the weekend, meaning it is more challenging to get a test. This may be adjusted by simply adding a weekday effect in our model as a covariate, using the approach described in Section 5.1.3. Such effect may be included either in an additive or a multiplicative fashion. At first, we considered effects for each day of

Table 5.3. *Log-likelihood, AIC, BIC and AICc for the models with baseline including additive or multiplicative week-day effect on daily positives.*

Index	Additive effect	Multiplicative effect
log-likelihood	−971.74	−974.1
AIC	1929.48	1934.3
AICc	1942.67	1947.5
BIC	1908.60	1913.4

the week, taking *Monday* as a corner point. Preliminary results showed that not all week-days present a significant deviation from the common mean. On the other hand, the distribution of the Deviance residuals $\hat{d}_t$ of the standard model aggregated by week-day (see Figure 5.8) shows that an evident overestimation pattern (i.e., negative deviations) is taking place on Monday and Tuesday.

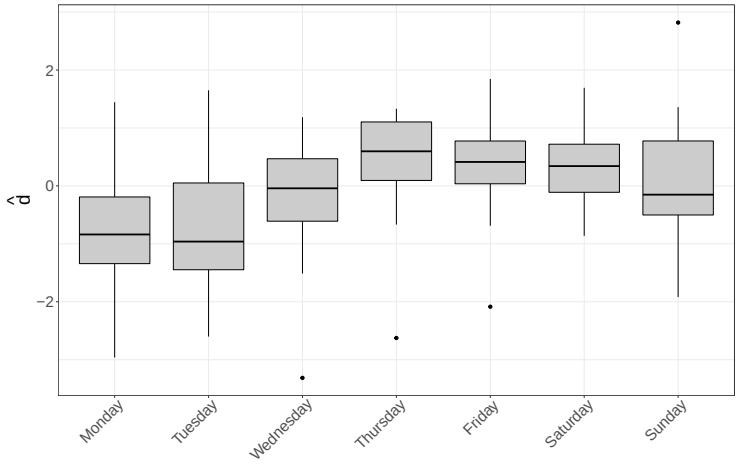


Figure 5.8. Deviance residuals distribution aggregated by day of the week for *daily positives*.

Therefore, in the sequel, we will present only results obtained with the dichotomous variable that is equal to 1 whenever the week-day is Monday or Tuesday (0 vice versa). Note that the lower testing effort during the weekend is reflected in the data on Monday and Tuesday, since daily reports involve mostly results received the day before, with swabs therefore dating back 48 hours on the day of publication. This confirms that working with daily data require special care as the cases reporting may suffer from week seasonality. The additive option is chosen over its alternative because of its lower/improved *AIC*, *BIC* and *AICc* score (see Table 5.3).

The resulting fit of the model with week seasonality on the observed data are shown in Figure 5.9.

Estimated parameters are shown in Table 5.4. This model estimates the baseline to be at $e^{\hat{\beta}_0} = 192.48$ on Wednesday to Sunday and at $e^{\hat{\beta}_0+\hat{\beta}_{wd}} = 121.51$ on Mondays and Tuesdays. Deriving the corresponding 95% intervals, these two baselines result significantly different from the estimate of the overall baseline $\hat{\alpha}$ in the model without covariates. The estimates of the outbreak size $\hat{r}$ and of the infection rate $\hat{h}$ of the two models are in agreement, while the point estimates of the asymmetry parameter $\hat{s}$ are different but both large and mutually included in the corresponding 95% intervals. This is reasonable since we would not expect the outbreak size, rate and symmetry

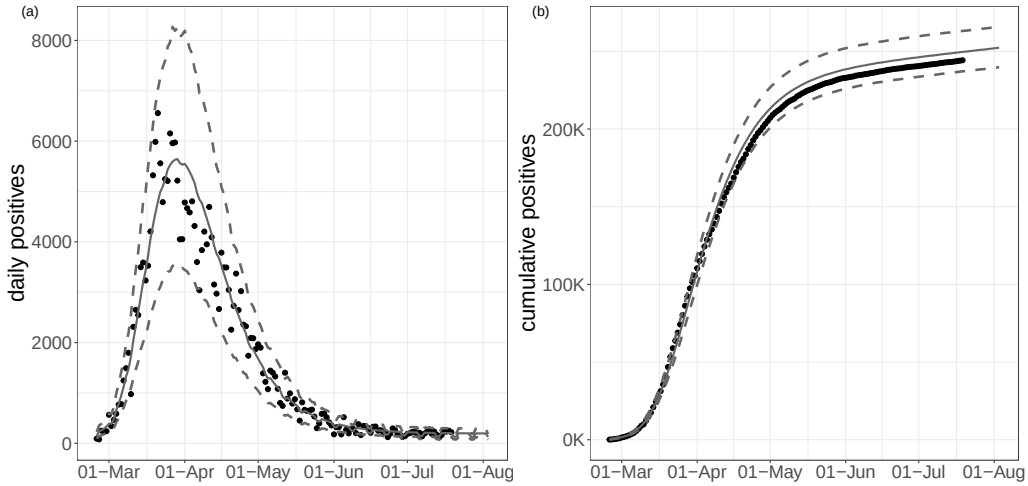


Figure 5.9. Observed (black dots) and fitted values (grey solid lines) with 95% confidence intervals (grey dashed lines) for the model with baseline and week-day additive effect, estimated on the *daily positives*.

Table 5.4. *Parameters' point estimates and* 95% *confidence intervals for the additive model on daily positives*

Parameter	Point estimate	95% Interval
β_0	5.26	$(5.18, 5.34)$
β_{wd}	-0.46	$(-0.53, -0.38)$
r	224.57×10^3	$(224.13 \times 10^3, 225.01 \times 10^3)$
h	0.0289	$(0.0287, 0.0291)$
p	-23.26	$(-29.64, -16.88)$
s	44.42	$(-35.67, 124.51)$
ν	22.01	$(21.35, 22.70)$

to vary after accounting for week-day heterogeneity. On the other hand, the new estimate $\hat{p}$ of p detects a shorter lag-phase and hence a slightly faster approach to the descending phase. Finally, the estimate of the dispersion parameter $\hat{\nu}$ is slightly larger than the one obtained using the model without covariates, denoting less over-dispersion with respect to the equivariance hypothesis. This is completely reasonable since the week-day effect is able to explain some of the previously unaccounted heterogeneity.

In terms of model validation, the inclusion of this effect improves sensibly the pseudo-R^2 (0.956), while the average coverage $\overline{\text{Cov}}_{95\%}$ remains stable around 0.950. The diagnostic check of the Deviance residuals showed that adherence to Gaussianity improved and the correlation pattern at lag 7 is still present but mitigated (see Figure 5.10).

Prediction of future cases and of the peak date

For the latter empirical model, the RMSPEs for each steps-ahead are presented in Figure 5.11. Results match the expectations as: (i) the error decreases with the length of the fitting window; (ii) the error trend is more stable on larger testing windows (10-15 steps ahead vs 1-5 steps ahead); (iii) larger errors are made around

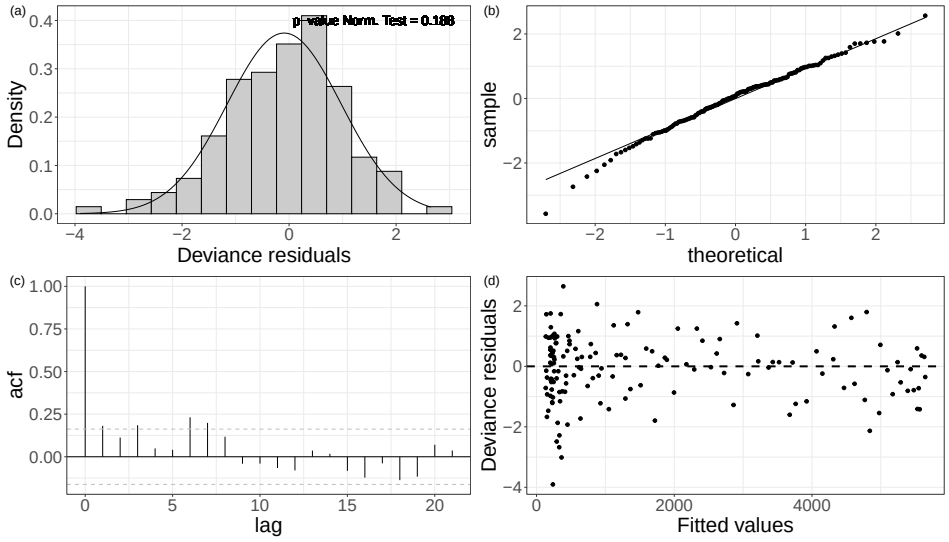


Figure 5.10. Deviance residuals for the model with baseline and week-day additive effect estimated on *daily positives*.

the day of the peak. It can be seen nevertheless that predictions are always reasonable at these time horizons. This is a good point for our theoretical framework, as it can be used as a guidance tool to plan non-pharmaceutical-interventions due to its capability to predict future scenarios with reasonable accuracy. Of course, with more detailed data and including confounding factors, the accuracy may be further improved. Unfortunately, the aggregated available data do not contain important information which may strongly improve the prediction, e.g. stratifications of cases by age, gender, comorbidities, etc.

Finally, we evaluate the model's ability to predict the date of the peak. The approximate dates and heights of the peak have important epidemiological implications. This becomes possible under the assumption that sensible modifications of the adopted epidemiological strategies do not emerge. However, if exogenous events, e.g. efficient treatments or vaccines, arise at a certain point in time, our framework allows to include it to predict the peak, in a similar manner as we did for the week seasonality effect. To do so, we estimate the model without covariates, using all available data until $K \in \{15, 10, 5, 3, 2, 1\}$ days before the observed peak. For the sake of conciseness, we only report results for $K \in \{10, 5, 2, 1\}$ as shown in Figure 5.12.

When $s = 1$, the peak $\hat{t}$ is directly expressed by the parameter p. When $s \neq 1$, after some algebra it can be seen that the peak can still be computed analytically as:

$$\widehat{t_{\hat{\gamma}}} = \hat{p} + \frac{\log_{10}(\hat{s})}{\hat{h}}.$$

Confidence intervals are obtained through the same bootstrap procedure introduced in Section 5.1.3. The dashed grey vertical lines represent the bounds of the confidence interval and the predicted date of the peak (confidence area is shaded with the same grey). The solid vertical black line represents the *"true"* date of the peak (i.e. obtained via smoothing of the observed counts through non-parametric polynomial approximations). The observed time-series is represented through point and lines, where the black section is referred to the training window while the grey section is referred to the testing (out-of-sample) window.

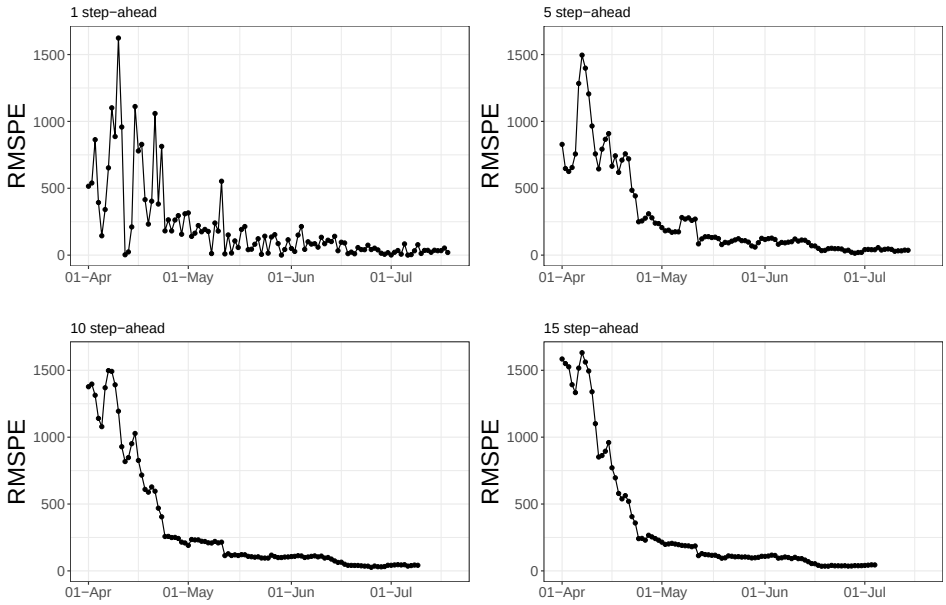


Figure 5.11. RMSPE for *daily positives* at different steps-ahead.

Table 5.5. *Delay (days) in point estimation of the peak.*

Days before	10		5		2		1	
	Delay	Width	Delay	Width	Delay	Width	Delay	Width
daily deceased	-1	37	-3	25	-4	22	-3	21
daily positives	20	106	17	69	1	37	2	37

As expected, as we approach the real date of the peak, we predict it more accurately. Point predictions are very accurate since 5 days before the actual peak. At the same time, interval bounds get tighter and tighter as the fitting interval approached the day of the peak and, in general, the day of the peak is always included in such bounds (see Table 5.5 for exact numerical evaluation).

All the aforementioned results have been calculated also for the national aggregated *daily deceased*. Exposition and discussion of these results, which are in fact very similar to the *daily positives* ones, are included in the supplementary material. Here, we just want to highlight how the peak is accurately predicted with a shorter delay and generally smaller uncertainty for the *daily deceased* than for the *daily positives* (see Table 5.5). This is probably related to the more regular behaviour of the series, due to a likely more homogeneous collection process of the records.

Finally, we here want to stress the point that we are introducing a framework with the highly desirable goal to formulate a model which would predict an evolution curve. To be more precise, a great variety of epidemiological models have been proposed in the literature, but most standard versions of SIR-like models typically yield an increase before the peak that is quite similar to the decrease after the peak. The proposed framework, based on more complex evolution dynamics, is robust enough to be fitted successfully on the (poor quality) available data while explaining and forecasting different increasing and decreasing behaviour before and after the peak. We emphasize that such increase-decrease quantitative behaviours appear to satisfactorily conform to reality.

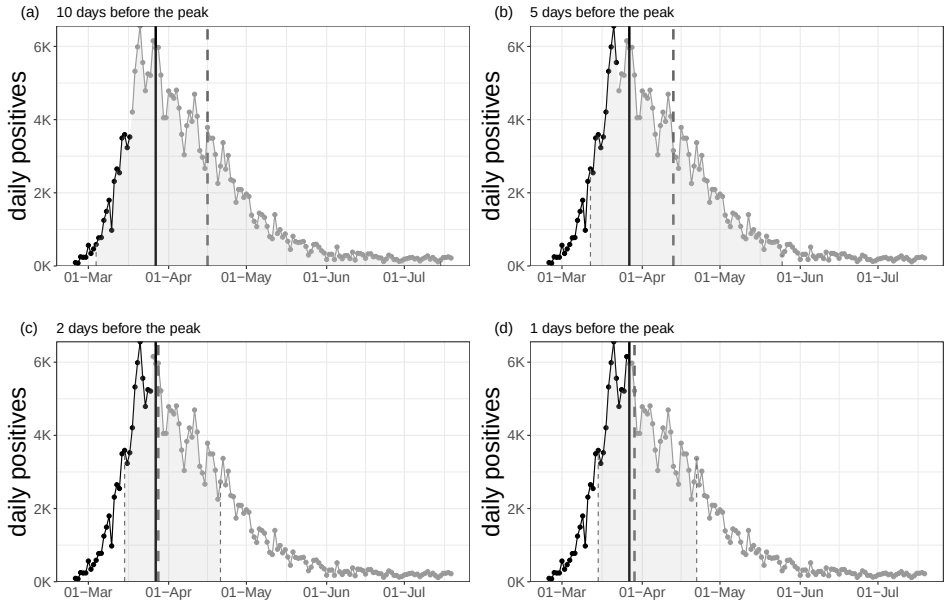


Figure 5.12. Estimation of the date of the peak for *daily positives* at different steps-before.

5.1.5 Discussion and further developments

We presented an approach to modeling and prediction of epidemic indicators that has proven useful during the first outbreak of COVID-19 in Italy. The model has been validated on publicly available data, and has proved flexible enough to adapt to different indicators.

It is important to underline up front that the available data are clearly biased. Incidence depends on testing and tracing efforts, whose indications have varied wildly over time and space. Comparability of indicators over time and space might in part be achieved by including the daily number of swabs as a predictor, which anyway would make predictions cumbersome to obtain. Different definitions of COVID-19 related death make it also very hard to compare mortality across countries. This problem does not apply to our data, that refer to Italy. However, while this definition has been constant over time in Italy, it shall be remarked though that also deaths might be underestimated, with the degree of undercount positively associated with incidence. Correcting for this bias is not trivial and require corrections based on individual-level data and/or reliable statistics about excess mortality.

Summarizing the results, we would like to emphasize that the proposed Richards' curve model describes properly the growth in the number of COVID-19 daily positives and daily deceased, despite its simplicity. Indeed, it is able to reflect properly the trend of the considered daily incidence indicators and also allows for the straightforward inclusion of exogenous information. Basic covariates such as the week-day effect proved to sensibly enhance model fitting and prediction accuracy. While we have illustrated results at the national level, the model can clearly be used also at the regional/local level (perhaps including specific local effects) and the resulting fits are included in the supplementary material. The maximum likelihood approach so far considered is rather stable, as long as reasonable starting values are passed to initialize the algorithm. Of course, different approaches could be investigated.

A limitation of our approach is that logistic growth curves are constrained so that only one wave at a time can be successfully modeled. This implies that initial (and possibly final) dates shall be set by the user to identify a wave. This is rather simple empirically (e.g., the initial date can be the last day with zero incidence, and the final date can be the first day with incidence above (or under) a pre-specified threshold). On the other hand, multiple waves could be modeled by modification of our non-linear model as a weighted average of multiple Richards' curves (one for each wave), in which weights of the non-current wave are forced to decay to zero with the distance from the wave-specific peak. We leave this as grounds for further work.

A Bayesian approach will also be experimented in order to overcome possible issues with the asymptotic properties of the maximum likelihood estimator. Notably, implementation of the *No-U-Turn Sampler* (NUTS) algorithm for the estimation of non-linear models might be a valid working solution. Additionally, a Bayesian approach may also be used to include spatial dependence into the modeling framework and also to relax the first-order Markov assumption for taking into account more complex temporal dependence. In particular, the latter may be key in order to adapt the introduced Richards' curve model for the nowcasting of prevalence indicators, e.g. *current positives* and *current intensive care units occupancy*. Indeed, any modeling effort shall account for the strong temporal dependence between subsequent counts stemming from the fact that daily counts at time t potentially include units which are in stock since times $\tau < t$. Furthermore, as specified in Section 5.1.2, prevalence indicators are non-monotonic and their value is the result of the combination of the incidence components building up each of those. These two last issues may be addressed by adapting the Richards' response function to accommodate non-monotonicity and/or by hierarchically specifying a model for the prevalence indicators through the combination of models for their incidence components. A successful attempt in accurately nowcasting the *ICU occupancy* is given in Farcomeni et al. (2021).

5.2 Spatio-temporal modelling of COVID-19 incident cases using Richards' curve: an application to the Italian regions

5.2.1 Introduction

One of the limitations of the approach presented in Section 5.1 is that indicators in each area are modelled independently. That is clearly only a working assumption, as mobility have occurred across Italian regions also during the hard lockdown of Spring 2020. Even sick people with COVID-19 have been sometimes transferred from one region to another. Furthermore, it is likely that regions close to each other culturally, economically, or geographically (e.g., sharing borders) present similar features as people experience similar climates, pollution and have similar lifestyles. For these reasons, this work aims to overcome this limitation by explicitly taking into account the spatial dependence *across* regions and the temporal dependence *within* regions. We make this extension for different specifications of the generalized growth model of Alaimo Di Loro et al. (2021a) in a Bayesian framework. The Bayesian formulation by itself is already a notably additional advancement, regardless of the model specification. We report here that posterior summaries, in our experience, seem to be more stable compared to the maximum likelihood estimates, possibly due to difficulties in finding a global optimum for the likelihood of an inherently non-linear

model. Furthermore, exploiting hierarchical models' flexibility in a Bayesian context, we replace the Negative Binomial assumption in Alaimo Di Loro et al. (2021a) with the Poisson distribution. The still present over-dispersion and unobserved heterogeneity are accounted for by including observation-specific random effects. If the random effects were assumed to be gamma-distributed, the corresponding marginal would indeed be a Negative Binomial and the method would be analogous to the original modelling framework. However, we rather consider normally distributed random effects on the log scale. Gaussianity allows for a more straightforward specification of prior information and inclusion of possible dependence structures in the process governing such effects. While temporal correlation is dealt with an *Auto-Regressive* (AR) structure, spatial dependence is included by specifying a suitable *Conditional Auto-Regressive* (CAR) prior, where the covariance matrix is identified using two possible networks: one based on geographic proximity and one built on historical data of transport exchanges between regions (taken from Della Rossa et al., 2020). The advantage of introducing this dependence structure is twofold. On the one hand, the resulting simultaneous model provably gives more accurate description of the true pandemic evolution than separate models for each region. On the other hand, it can be expected that parameter estimates of characteristics of interest (e.g., peak time and height) can benefit from the pooling information from multiple regions. We separately evaluate the first and second wave of Sars-CoV-2 in Italy. Similarly to Bartolucci and Farcomeni (2021), we consider weekly incidence, even if observed cases are made available daily. That is done in order to mitigate the issues with erratic daily fluctuations due to late reporting. Even though we are aware that this does not solve the data issues, it sufficiently alleviates them, as testified by the smoother time series obtained at the weekly level.

5.2.2 Methodology

Public data about COVID-19 in Italy are published every day by the Civil Protection Department, since February 24th, 2020[5]. For each of nineteen regions and two provinces (Trento and Bolzano, forming the region of Trentino Alto Adige), these include (i) prevalence indicators (currently positive, Intensive Care Unit (ICU) occupancy, hospital occupancy) and (ii) incidence indicators (e.g. newly diagnosed positives, deceased, new admissions to ICU, swabs, subjects tested). For a more technical description refer to Dicker et al. (2006); Alaimo Di Loro et al. (2021a). For any of the incidence indicators in a given area, the number of new cases at time $t = 1, \dots, T$ can be obtained as the first difference of its cumulative counterpart as $Y_t = Y_t^c - Y_{t-1}^c$ where Y_t and Y_t^c are the number of new and cumulative cases at time t, respectively, and where we may assume $Y_0^c = 0$ without loss of generality. Cumulative indicators present some peculiarities: they are monotone non-decreasing and their behaviour usually follows a logistic-type growth curve. These curves have been widely used to describe various biological processes (Werker and Jaggard, 1997). More recently, they have also been adapted in epidemiology and biostatistics for modelling the onset and the spreading of epidemics (Hsieh, 2009; Hsieh and Chen, 2009; Hsieh, 2010).

Alaimo Di Loro et al. (2021a) proposed a modified Richards' curve (Richards, 1959), also known as the *Generalised Logistic Function*, for modelling cumulative incidence indicators. The generalised logistic function can accurately model various monotone processes and include other widely-used logistic growth curves as special cases (Tsoularis and Wallace, 2002; Gompertz, 1825). In Alaimo Di Loro et al.

[5] `GitHub` repository: https://github.com/pcm-dpc/COVID-19.

(2021a) a parametric model is specified for region-specific incidence indicators (e.g., Poisson or Negative Binomial), where the cumulative indicators are assumed to follow a five-parameters Richards' curve.

Here, we propose a slightly different parametrization of the classic Richards' curve to allow the model to reach an endemic state in which there is a constant (hopefully, small) growth. In particular, without loss of generality, we modified Equation (5.1) by considering a linear trend on the baseline b and using the exponential as the power base:

$$\Lambda_{\boldsymbol{\gamma}}(t) = b \cdot t + \frac{r}{(1 + e^{h(p-t)})^s}. \tag{5.7}$$

This is similar to the use of an endemic parameter for the first differences as pursued in Alaimo Di Loro et al. (2021a).

Assuming that $\mathbb{E}[Y_t^c] = \Lambda_{\boldsymbol{\gamma}}(t)$, the expected value for the innovation Y_t can be straightforwardly obtained as:

$$\mathbb{E}[Y_t] = \mathbb{E}[Y_t^c] - \mathbb{E}[Y_{t-1}^c] = \Lambda_{\boldsymbol{\gamma}}(t) - \Lambda_{\boldsymbol{\gamma}}(t-1) = \lambda_{\boldsymbol{\gamma}}(t), \tag{5.8}$$

where:

$$\lambda_{\boldsymbol{\gamma}}(t) = b + r \cdot \left[(1 + \exp(h(p-t)))^{-s} - (1 + \exp(h(p-t+1)))^{-s} \right]. \tag{5.9}$$

Inclusion of space-time dependence in the Richards' based model

Let $\mathbf{Y}_g = [Y_{gt}]_{t=1}^T$ denote the time-series of number of *new* cases in area g, for $g = 1, \ldots, G$, such that

$$\boldsymbol{Y} = \left[\mathbf{Y}_1^\top, \mathbf{Y}_2^\top, \ldots, \mathbf{Y}_G^\top \right]^\top.$$

The main assumption of our model is that Y_{gt} arises from a Poisson distribution with mean $\mu_{gt} = E_g \cdot m_{gt}$. This can be expressed as:

$$\begin{aligned} &Y_{gt} | \mu_{gt} \sim Pois(\mu_{gt}) \\ &\log(\mu_{gt}) = \log(E_g) + \log(m_{gt}), \quad g = 1, \ldots, G, \quad t = 1, \ldots, T, \end{aligned}$$

where $\log(E_g)$ is an offset term that accounts for region-specific exposures levels.

When the offset is present, all other parameters impacting the overall rate become dimensionless, regardless of the scale of the corresponding region. In other words, the term m_{gt} can be interpreted as a relative measure of the risk of region g at time t with respect to the considered offset E_g.

Different specifications of m_{gt} lead to different models, each with its own characteristics. We decompose the log-risk in three main components:

$$\log(m_{gt}) = \phi_{gt} + \log\left(\lambda_{\gamma_g}(t)\right) + \boldsymbol{x}_{gt}^\top \boldsymbol{\beta}, \tag{5.10}$$

where ϕ_{gt} is a specific random effect for the g-th area at time t, $\lambda_{\gamma_g}(t)$ is a deterministic function denoting the general time trend, with possibly region specific parameters $\boldsymbol{\gamma}_g$ as for Equation (5.9), and $\boldsymbol{x}_{gt}^\top \boldsymbol{\beta}$ a linear predictor based on K covariates with associated regression coefficients $\boldsymbol{\beta}$.

The Spatio-Temporal CAR model

The observation-specific random effects $\{\phi_{gt} : g = 1, \dots, G, t = 1, \dots, T\}$ are included to account for unobserved heterogeneity in the data. At each time, possibly correlated random effects allow regional curves to deviate from their global average. Besides, their presence corrects for the evident over-dispersion (with respect to the Poisson assumption) present in the time-series.

These random effects can either be completely independent, present temporal dependence, spatial dependence, or spatio-temporal dependence. In a Bayesian framework, the covariance structure can be induced hierarchically by specifying a suitable prior on the complete set of random effects. In order to simplify the formulation of the random effects prior, we collect all of them in a set of time-varying vectors $\boldsymbol{\phi}_t = [\phi_{1t}, \dots, \phi_{Gt},]^\top$, $t = 1, \dots, T$.

Spatial dependence at each time point can be introduced by using a CAR prior (Besag, 1974) over some network, that under Gaussianity produces a so-called *Gaussian Markov Random Field* (GMRF, Rue and Held (2005)). This approach falls into the wide range of methods related to *disease mapping* (see Waller and Carlin (2010) and Lawson (2018) for a review). Such CAR prior specification allows incorporating the undeniable spatial correlation at the second level of the model hierarchy, avoiding analytical complications inherent in modelling spatial correlation within non-Gaussian distributions with inter-related mean and variance structures (Gelfand et al., 2010). This form of dependence is valid on discrete domains arranged over a network, where neighbouring relationships are determined by an adjacency matrix $\boldsymbol{W}$ (possibly weighted). The matrix $\boldsymbol{W} = [w_{ij}]$ is a $G \times G$ symmetric matrix with all diagonal elements equal to 0 (as no region/area/unit is its own neighbour), and where off-diagonal elements w_{ij} are greater than 0 if and only if areas i and j are connected ($i \sim j$): the larger the connection strength w_{ij}, the closer the two random effects are pulled together. The original expression of this prior starts from the consideration of the full conditional of each random effect given all the others. For the generic $t \in \{1, \dots, T\}$, the full conditional $\phi_{gt}|\boldsymbol{\phi}_{-gt}$, $g = 1, \dots, G$ has mean equal to the weighted combination of the random effects in its neighbourhood:

$$\phi_{gt}|\boldsymbol{\phi}_{-gt} \sim \mathcal{N}\left(\sum_{j=1}^{G} w_{gj}\phi_{jt}\,,\, \sigma^2\right), \quad \forall\, g = 1, \dots, G,$$

where $\boldsymbol{\phi}_{-gt} = \left[\phi_{1t}, \dots, \phi_{(g-1)t}, \phi_{(g+1)t}, \dots, \phi_{Gt},\right]^\top$ and σ^2 is the overall variance of the random effect. This induces smooth variations over close regions, as determined by $\boldsymbol{W}$. Following Brook's lemma, for a fully connected graph (i.e. with no "islands"), this local specification implies a very specific global multivariate prior on the vector $\boldsymbol{\phi}_t$, centered at $\mathbf{0}$ and with precision matrix $\boldsymbol{Q}$ which depends on the network structure. Under row-wise normalization of the weights in $\boldsymbol{W}$, and introducing a spatial smoothing parameter α, this global prior can be expressed as:

$$\boldsymbol{\phi}_t \sim \mathcal{N}_G\left(\mathbf{0},\, \sigma^2 \cdot \boldsymbol{Q}(\alpha, \boldsymbol{W})^{-1}\right), \quad \forall\, t = 1, \dots, T, \tag{5.11}$$

where $\boldsymbol{Q}(\alpha, \boldsymbol{W}) = (\boldsymbol{D} - \alpha\boldsymbol{W})$ and the matrix $\boldsymbol{D}$ is a diagonal matrix containing the row sums of the weights of each region on the diagonal. This simply ensures that the weights of each region are properly normalized over all its neighbours (i.e. the row-wise normalization). The spatial smoothing parameter α regulates the amount of spatial dependence: values close to 0 approximate independence (no impact of $\boldsymbol{W}$) and values close to 1 strong spatial dependence (full impact of $\boldsymbol{W}$).

This spatial CAR expression introduces of spatial dependence among random effects belonging to connected regions at the same time-point. Nevertheless, we cannot neglect the indisputable temporal correlation which characterizes such kind of data. Following the original work by Rushworth et al. (2014), we induce such dependence by imposing a temporal *Auto-Regressive* structure over the vectors $\{\boldsymbol{\phi}_t\}_{t=1}^{T}$. This yields a spatio-temporal CAR model (CAR-AR) whose only difference in this work from the original version is that, instead of the mixed specification of Leroux et al. (2000), we here consider the typical CAR of Besag (1974). In particular, we also extend the original AR(1) formulation introduced in Rushworth et al. (2014) to order $J \geq 1$. The extended specification amounts to the following prior for the collection of time-varying spatial vectors:

$$\begin{aligned} \boldsymbol{\phi}_t &\sim \mathcal{N}_G\left(\mathbf{0},\, \sigma_0^2 \cdot \boldsymbol{Q}(\alpha, \boldsymbol{W})^{-1}\right), && t = 1, \ldots, J, \\ \boldsymbol{\phi}_t &\sim \mathcal{N}_G\left(\sum_{j=1}^{J} \rho_j \cdot \boldsymbol{\phi}_{t-j},\, \sigma^2 \cdot \boldsymbol{Q}(\alpha, \boldsymbol{W})^{-1}\right), && t = J+1, \ldots, T, \end{aligned} \tag{5.12}$$

where $\{\rho_j\}_{j=1}^{J}$ is the set of coefficients governing the amount and direction of temporal dependence at different lags. Such a complex AR structure may be very useful to catch dependence and seasonality patterns at different temporal scales. For instance, we may expect to observe a strong weekly seasonality at the daily level, which may be well-captured by an AR(7) specification (potentially with some of the lower order coefficients set to zero). Remark that AR(J) processes are stationary only if the characteristic polynomial $\phi(z) = 1 - \sum_{j=1}^{J} \rho_j z^{J-j}$ has the reciprocal of its roots $\{\eta_j\}_{j=1}^{J}$ lying inside the unit circle. This property is desirable as it favors the identification of all the considered space-time components. This is not easily enforced for arbitrary values of J. A general prior choice in this sense is provided in Huerta and West (1999), where the process is reparametrized in terms of η_j's and other auxiliary latent components. Section 5.2.3 focuses on the analysis of weekly data that do not show any cyclic behaviour. Therefore, we only expand on the case $J = 1$, where stationarity is guaranteed by simply enforcing $\rho_1 = \rho \in (-1, 1)$. In the time series literature the first J time points are usually ignored, and just conditioned upon. Equation (5.12) makes a simplifying working assumption of independence of the first J time points to obtain a marginal parameterization. This clearly is irrelevant for the case $J = 1$, while in the other cases we recommend checking the goodness of fit of the initial joint distribution, and maybe adjust the assumptions. All the aforementioned hyperparameters are then ascribed standard hyperpriors, commonly found in the literature:

$$\alpha \sim \mathcal{B}e(0.5, 0.5) \qquad \rho \sim \mathcal{U}nif(-1, 1) \qquad \sigma^2 \sim \mathcal{IG}\,(2,\, 2)\,,$$

where the latter has been coded as $\sigma^2 = \frac{1}{\tau^2}$ with $\tau^2 \sim \mathcal{G}a\,(2,\, 2)$ (see Algorithm D.1 in Appendix D).

The logistic growth trend

In Section 5.2.2, we state that we want to model the general trend of COVID-19 counts (of positives) in a single outbreak by using the first differences of the Richards' curve as in Alaimo Di Loro et al. (2021a). The first differences in Equation (5.9) do not present an elegant expression and are slightly cumbersome to work with. Since data are collected at equally spaced time intervals, we propose to linearly

approximate $\lambda_{\boldsymbol{\gamma}}(t)$ with the derivative of the Richards' curve, as follows:

$$\lambda_{\boldsymbol{\gamma}}(t) \approx \tilde{\lambda}_{\boldsymbol{\gamma}}(t) = \frac{d}{dt}\Lambda_{\boldsymbol{\gamma}}(t) \cdot \Delta t = b + r \cdot s \cdot h \cdot \exp\{h \cdot (p-t)\} \cdot (1 + \exp\{h(p-t)\})^{-(s+1)}, \tag{5.13}$$

where $\Delta t = 1$. In our implementation, we initially considered both the exact and the linearised version of Equation (5.9) and Equation (5.13), respectively. In the final results, differences were negligible, but the latter provided improved numerical stability and convergence of the chains. Thus, we decided to stick to this version, which is also used to produce the final results included in Subsection 5.2.3.

The expression in Equation (5.10) implicitly considers a very highly parametrised model, where each region is allowed its on vector of parameters $\{\boldsymbol{\gamma}_g\}_{g=1}^{G}$, hence its own Richards' curve, to drive the trend of the regional outbreaks. In the sequel, we alternatively envision the existence of one common single Richards' curve governing the spread of the epidemic in all the Italian regions, which then deviate from this *global average* as an effect of specific characteristics (observed or unobserved). This is obtained as a particular case of the former, where $\boldsymbol{\gamma}_g = \boldsymbol{\gamma}$, $\forall\, g \in \{1, \ldots, G\}$.

There is an essential difference between these two specifications, especially in terms of the role of the space-time random effects. The first one is a local model, where the random effects represent temporal variations of the mean underlying each regional counts-series from the region-specific trend. In the second case, they instead represent the spatio-temporal deviations of each region's means, at each time, from the common curve. From a dependence interpretation standpoint, in the first case we are assuming that if region g and region j are connected, when region g deviates from its trend $\lambda_{\boldsymbol{\gamma}_g}(\cdot)$, then region j will likely have similar deviation from its own trend $\lambda_{\boldsymbol{\gamma}_j}(\cdot)$ as well. In the second case, we are assuming that if region g and region j are connected, when region g deviates from the general trend $\lambda_{\boldsymbol{\gamma}}(\cdot)$, then region j will also deviate from the general trend similarly.

Let us recall that the parameters $\{b, r, h, s\}$ governing the differences of the Richards' curve are constrained on the positive domain $\mathbb{R}^+$. In order to favour the elicitation of diffuse priors and the Bayesian estimation process, these have been parametrised on the log-scale as $\{\log(b), \log(r), \log(h), \log(s)\}$. The first two have been assigned a $\mathcal{N}(0, 100)$ prior, while the last two a $\mathcal{N}(0, 1)$ one. These correspond to very vague priors on the log-scale, where the second are assigned a lower variance given their double-exponentiated nature in Equation (5.13). One may argue that the same prior specification for $\log(b)$ and $\log(r)$ does not reflect the natural intuition about b being some order of magnitude lower than r. However, while this may seem obvious in the case of COVID-19 pandemic waves, we here want to point out that this is not necessarily true in general. For instance, in the case of an endemic disease, we may observe a relatively high baseline (endemic rate) with only small seasonal waves of infections that could be rapidly contained. However, we performed some preliminary runs embedding such prior belief before proceeding to the final analysis of Section 5.2.3. The results were indistinguishable to the ones obtained using vague priors and therefore we opted for the latter in order to let the data drive the final estimates. The parameter p, unlike the others, belongs to the whole real line $\mathbb{R}$ and, more importantly, is not dimensionless. Its magnitude is indeed related to the dimension of the analysed time window. It can be loosely interpreted as the lag-phase of the outbreak (for $s = 1$ it represents precisely the point of maximum of the curve), which is the point in time when the exponent $h \cdot (p - t)$ becomes negative. It is not well-identified for varying s, and hence it has been given a $\mathcal{N}\left(T/2,\, T/(2 \cdot 1.96)\right)$ prior to help it move inside the observed time interval (included in $[0, T]$ with 95% probability).

The linear predictor

The linear predictor $\boldsymbol{x}_{gt}^{\top}\boldsymbol{\beta}$ describes the effect of covariates on the log-risks. Since the *dimension* of the region is already accounted for in the offset term, such covariates shall account for exogenous factors that affect the spread of the virus, or the ability to detect the infected people in each area at different times. In practice, this term shall represent all the meaningful observed heterogeneity between regions and within regions over time.

For instance, the population density is a region-specific and constant over time feature that can likely impact on the rate of infection. This covariate has been considered in the recent work by Jalilian and Mateu (2021) and proved to be valid for both explanation and prediction purposes. Another interesting variable to study may be the number of daily swabs. Its inclusion accounts for the effort in detecting positive cases carried out by a region at a specific time.

If K covariates are considered, then the vector $\boldsymbol{x}_{gt}^{\top}$ is associated to a $(K \times 1)$ vector of coefficients $\boldsymbol{\beta}$. In our Bayesian machinery, this vector is assigned a multivariate Normal prior with independent components $\mathcal{N}_K\left(\mathbf{0}, 100 \cdot \boldsymbol{I}_K\right)$, which corresponds to a fairly diffuse prior considering the log-linear link.

It is here important to highlight that we are not including the intercept in the linear predictor. In the case of region-specific Richards' curves, the intercept is implicitly defined by the parameters b_g and r_g already, and its inclusion would introduce a non identifiable parameter and jeopardize proper convergence of the estimation algorithms. In the case of a common single Richards' curve, one may want to include region-specific intercepts $\{\beta_{0g}\}_{g=1}^{G}$. These would have the effect of moving the whole region-specific curve up or down with respect to the *global average*, again accounting for unobserved heterogeneity among regions. However, the goal is to have this heterogeneity explained by the spatio-temporal random effects ϕ_{gt}: the inclusion of such individual intercepts would add an unwanted player in the game and make the interpretation of the final results intricate.

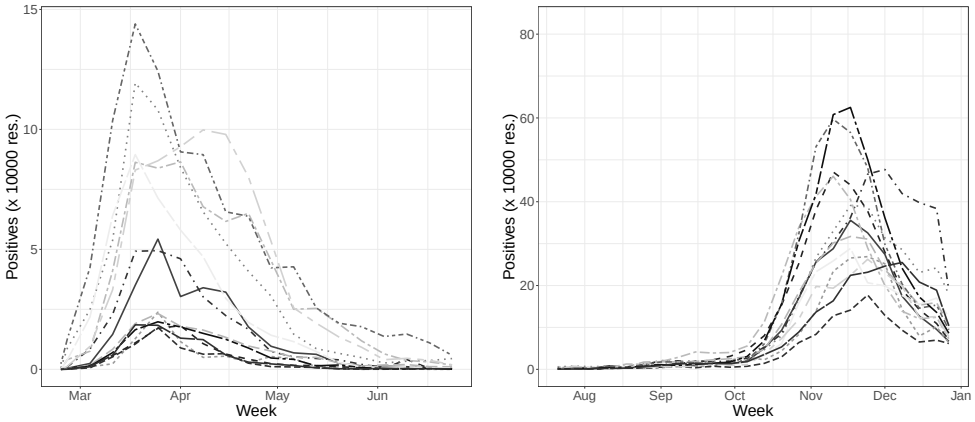
Estimation

Estimation has been carried out using `Stan`[6], which is a probabilistic programming language for statistical modelling and high-performance statistical computation (Carpenter et al., 2017; Stan Development Team, 2021). It interacts with `R` and can be called directly from `RStudio` (Allaire, 2012) through the `rstan` package (Stan Development Team, 2020). Among its many capabilities, it allows to get full Bayesian inference by drawing from the posterior density by a specific Markov Chain Monte Carlo (MCMC) sampling method known as Hamiltonian Monte Carlo (HMC, Betancourt (2017)). The HMC techniques provide an efficient sampling scheme based on the simulation of Hamiltonian dynamics for approximating the target distribution (Neal et al., 2011). Its functioning relies on the analogy between the parameter value and the trajectory of a fictitious particle subject to a potential energy field, preserving its total energy (the Hamiltonian), obtained as the sum of potential and kinetic energy. In practical terms, given the chain's current value and the corresponding log-density, it picks the new value of the chain by proposing a random shift in the log-density value and then moving arbitrarily far away from that point along the corresponding contour line. The latter allows for fast and complete exploration of the whole density that does not negatively affect the acceptance rate, minimizing the risk of wasting time (or even getting stuck) in local high-density

[6] Webpage at `https://mc-stan.org/`

areas. Unlike the Metropolis-Hastings (Metropolis et al., 1953) or Gibbs sampler (Geman and Geman, 1984), it provides robust performances and easily reaches convergence even for very complex models, e.g.: posterior density characterized by complex geometries, multi-layered hierarchical models with many parameters, models depending on large sets of latent variables, etc. One of the advantages of `Stan` is that it allows for an easy implementation of the No U-Turn Sampler (NUTS, Hoffman et al. (2014)). NUTS proved to perform at least as efficiently as the standard HMC but, generally, does not require any tuning of the hyperparameters governing the proposals. Hence, it sensibly reduces the computational burden and averts any user intervention or wasteful runs. Another advantage of the NUTS algorithm is that situations in which the sampling cannot thoroughly explore the whole posterior distribution are easily detectable. Indeed, when the approximation of the Hamiltonian dynamic fails to reach specific areas without departing from the original Hamiltonian value, the so-called *divergent* transitions arise. For more details we refer to Betancourt (2016a). The `Stan` interface reports divergences as warnings and provide ways to access which iterations encountered them. The `bayesplot` package (Gabry et al., 2019) can be used to visualize them and locate the areas in which the exploration failed. If no divergences occur, we can be confident that the chain was able to explore the whole domain of interest of the log-posterior density.

After few warm-up iterations, during which the NUTS automatically adapts its future behaviour to the shape of the posterior density, chain convergence and desirable accuracy are usually reached even in few iterations ($\approx 10^3$). Nevertheless, doing more iterations does not harm and longer chains lead to more robust result. When the log-posterior density is computationally intensive to compute or the geometry of the posterior is particularly complex, the approximation of the Hamiltonian dynamics can be significantly slowed down and negatively impact on the total run-time. For instance, for the model presented in Section 5.2.2, there are T spatial vectors of random effects that contribute to the overall density. The evaluation of the contribution of each of these requires the computation of the corresponding prior, which in turn involves the computation of inverse and determinant of the $G \times G$ matrix $Q(\alpha, \boldsymbol{W})$. It is clear how the naive implementation of such a model is all but efficient. Nevertheless, we can exploit two facts in order to ease computations. First, the spatial covariance structure does not vary over time, especially along iterations; this implies that we can compute the inverse and determinant only once in advance without doing the same calculations repeatedly. Second, each region has only a few neighbours, and the matrix $Q(\alpha, \boldsymbol{W})$ is not full, paving the way for efficient algebraic solutions. In practice, many efficient strategies can be adopted in order to alleviate the computational burden by speeding up linear algebra operations. Here, we based ours on the *Exact-sparse CAR* elaborated in Joseph (2016), which accrues significant computational efficiency by more than halving the needed run-time[7]. The original code has been slightly modified in order to include the temporal AR structure of our spatio-temporal CAR. The core of the `Stan` program needed to update the CAR-AR random effects is presented in Algorithm D.1 of Appendix D. The full codes to reproduce the results presented in Section 5.2.3 are available in a public `GitHub` repository accessible at `https://github.com/minmar94/Covid19-Spatial`.



(a) **(b)**

Figure 5.13. Regional weekly time series for *positives* during the first (a) and second (b) wave.

5.2.3 Application

We test and compare our proposals on the regional weekly positives time series made available by CPD. We consider the time series of the *weekly positives* at the regional ($G = 20$) level, for the first and the second wave of the epidemic. From an epidemiological perspective, there is no strict definition for what is or is not an epidemic wave (or phase). However, the scientific community agrees on the fact that the word *wave* implies a natural pattern of peaks and valleys, suggesting that even during a lull, future outbreaks of the disease are possible. Our proposal is able to model only one epidemic wave at a time, since the Richards' curve entails a single peak time and height. The latter implies that the start and end date of each wave must be set by the researcher. Albeit this is a drawback of our approach, it can be easily seen through sensitivity analyses that results do not drastically depend on this choice. We mention here the work by Bartolucci and Farcomeni (2021), which can flexibly model more than one wave at a time, but at the price of not being able to explicitly estimate important characteristics of each wave (e.g., peak time, onset time, etc.). Similarly, Farcomeni et al. (2021) does not require to identify a time frame for waves, but it is restricted to short term predictions.

In our application, we set the 24th of February 2020, as the start date of the first wave, namely when systematic data recording started, while the 19th of July 2020 is set as the end date. That is the day in which discos and pubs were re-opened after the lockdown period (a total of 22 weeks). For the second wave, the 20th of July 2020 was set as the start date, while the 27th of December 2020 was set as the end date (for a total of 24 weeks), which corresponds to the end of the last week of the year and, more importantly, is the day the vaccine campaign began in all Europe (a.k.a. *V-day*). The regional time series of the *weekly positives* for both waves are reported in Figure 5.13a and Figure 5.13b, respectively. It can be seen that the second wave had a slower onset (due to the seasonality of infections in early Summer) but a much higher peak for most regions. That is not only due to a larger number of infected with respect to the first wave (which has mostly hit only the northern part of Italy) but also to the much larger proportion of identified cases.

[7]The computational gain is inversely proportional to the degree of the network defining neighbourhoods

We recall that we used the logarithm of the number of residents scaled by a factor of 10^4 as an offset, essentially studying the number of positives per $10,000$ residents rather than crude incidence. This is necessary in order to be able to compare different regions, which can have very different number of infected only due to a very different number of residents. We also included the number of total weekly swabs (standardised) as covariate, to take into account different contact tracing efforts. The number of positive swabs can be assumed to be negatively associated to the proportion of undetected cases, and is one of the official indicators of the World Health Organization.

We also compared two different specifications for the adjacency matrix in the CAR-AR model. The first matrix, which we refer to as $\boldsymbol{W}_1$, specifies a neighbourhood structure based on proximity flows and the availability of direct train, flights, and ferry connections. This matrix has been also used in Della Rossa et al. (2020), and can lead to distant regions to be neighbours because of, for instance, frequent internal flight connections. The original matrix is a weighted measure of commuters' flow and is not symmetric since exchanges may have different magnitudes in the two directions. As a fast and viable solution to symmetrise the matrix in this application, we decided to dichotomise it. We set $w^*_{ij} = 1$ if there exists a positive flow in at least one of the two directions. The second adjacency matrix, which we refer to as $\boldsymbol{W}_2$, is the most typically adopted network defined on regions' mutual geographical position. In our application, we considered a first-order structure, where only pairs of regions sharing at least one land border are considered as neighbours.

The two different neighbourhood structures are shown in Figure 5.14a and Figure 5.14b, respectively. In particular, we report the number of edges (connections) and the (scaled) degree of each region. We notice that using $\boldsymbol{W}_1$ we end up with 18 out of 20 regions that have at least one connection (Molise and Valle d'Aosta have none), three of them having 12 neighbours (which is the mode), and where Sicilia is the most connected area with 15 neighbours. On the other hand, using $\boldsymbol{W}_2$ we end up with seven regions that have 3 neighbours; two regions that have 6 neighbours, while Sardegna, which is an island, has no connections. For Sicilia, which is also an island, we selected Calabria as the only neighbour. The two regions are separated by very few kilometres of sea (the Strait of Messina), with extremely frequent ferry connections.

As a baseline model for comparison, we also considered the possibility of a completely disconnected graph $\boldsymbol{W}_0 = [0]_{ij}\,,\ \forall\, i, j$, hence assuming complete spatial independence between regions. Nevertheless, being temporal dependence undeniably present in the observed series, we always retain the temporal AR structure between subsequent vectors $\boldsymbol{\phi}_{t-1}, \boldsymbol{\phi}_t,\ t = 2, \ldots, T$. As a matter of fact, preliminary runs that neglected this feature of the date produced way worse results (especially in terms of out-of-sample performances) that will not be reported in the sequel.

For the sake of brevity, we will refer to the model ignoring spatial dependence with the fully disconnected graph $\boldsymbol{W}_0$ as M_0, and as M_1 and M_2 to the models including spatial dependence using $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ as adjacency matrices, respectively. We considered these three dependence structures for the model with one common Richards' curve, which we name *common*, and the model with region-specific Richards' curves, which we name *regional*.

For all models considered, we ran two separate chains for $10,000$ iterations, allowing `Stan` to perform $5,000$ warm-up iterations each, which were discarded for inferential purposes.

We computed several metrics in order to compare the goodness-of-fit and predictive performance of the model's alternatives. The large flexibility of the space-time random effects specification easily makes the model fit the observed set of data

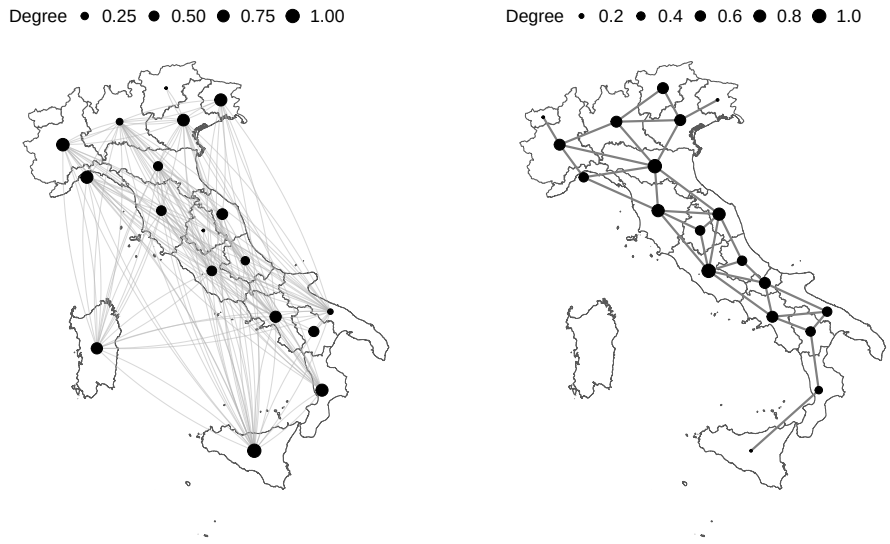


Figure 5.14. Network structure of the two adjacency matrices considered: $\boldsymbol{W}_1$ (a) and $\boldsymbol{W}_2$ (b).

almost perfectly. This feature exposes the typical in-sample metrics to over-fitting, flawing any sensible interpretation of the results, and would inevitably favour the highly parametrized *regional* model. Therefore, we decided to avert the over-fitting issues by artificially subtracting 15% randomly selected points from each region's time series. These are treated as missing data in the estimation process, and the ability to reconstruct the missing pieces properly is then verified in terms of various metrics: *Coverage*, *Root Mean Squared Error* (RMSE) and *Predictive Interval Width* (PIW). Comparison of these three metrics for the three dependence structures, with common and regional Richards, for the two waves, are presented in Table 5.6.

		Common			Regional		
Wave	**Metric**	M_0	M_1	M_2	M_0	M_1	M_2
I	Coverage	0.98	0.98	0.98	0.96	1	0.96
	PIW	1535	1178	1144	3311	846	1017
	RMSE	423	184	272	399	331	314
II	Coverage	0.96	0.97	0.92	0.97	0.96	0.92
	PIW	33393	4497	4046	16900	3131	3121
	RMSE	12841	910	995	3669	1008	1038

Table 5.6. Out-of-sample predictive performances of our proposals with a common or region-specific Richards' curve in the first and the second wave.

We can clearly observe how the out-of-sample performances are comparable across the *common* and *regional* specifications. The coverage is close (actually larger in most cases) to the 95% nominal level in both cases, for all the dependence structures. M_1 and M_2, under both the common and regional specification of the logistic trend, show equivalent coverage and similar PIWs. However, the common specification provides more accurate out-of-sample predictions in terms of RMSE, whenever the random-effects account for the spatial dependence (M_1 and M_2). Therefore, given the comparable out-of-sample performances and the more parsimonious specification of the *common* model, we chose this one as the preferred option. All future results

will then be referred to this specification.

Parameter estimates for the spatial (α) and temporal (ρ) auto-correlation, together with the swabs' effect (β) are reported in Table 5.7. Here, we want to first highlight that there is a clear evidence of strong dependence both spatial and temporal, with values of $\hat{\rho} > 0.8$ in all models for both waves. It is instead notable how the transport based graph detects low spatial dependence ($\hat{\alpha} \approx 0.14$) during the first wave, and a large spatial dependence during the second wave ($\hat{\alpha} \approx 0.93$). This change in the spatial correlation parameter between the first and the second wave for M_1, highlights the different type of non-pharmaceutical measures that were adopted to contain the spread of the contagion and the estimated values are completely reasonable, given the harder block to inter-regional movements that characterized the first wave as compared to more liberal mobility policies that accompanied the second one. On the contrary, the geographic vicinity effect is stable across the two waves, probably capturing similarities between close regions that depend on shared unobserved characteristics more than on people exchange.

The parameter β represents the effect of additional swabs on the number of detected positives. Being the swabs variable standardized, this does not allow for a trivial interpretation. However, we can observe a positive effect which was more evident during the first wave than the second wave. This happens unsurprisingly, since the testing efforts were not yet at full capacity during the first outbreak, with many undetected cases, detected as soon as additional testing hubs were made available.

Wave	**Param.**	M_0	M_1	M_2
I	α	–	0.14 (0.02, 0.21)	0.76 (0.71, 0.81)
	ρ	0.89 (0.87, 0.91)	0.88 (0.90, 0.93)	0.86 (0.85, 0.89)
	β	0.36 (0.26, 0.44)	0.34 (0.25, 0.42)	0.21 (0.14, 0.29)
II	α	–	0.93 (0.92, 0.95)	0.87 (0.85, 0.90)
	ρ	0.88 (0.86, 0.90)	0.87 (0.85, 0.89)	0.82 (0.80, 0.85)
	β	0.42 (0.38, 0.46)	0.27 (0.24, 0.30)	0.13 (0.09, 0.16)

Table 5.7. Comparison of parameters' estimates for the spatial (α) and temporal (ρ) auto-correlation, and for the swabs' effect in the first and the second wave.

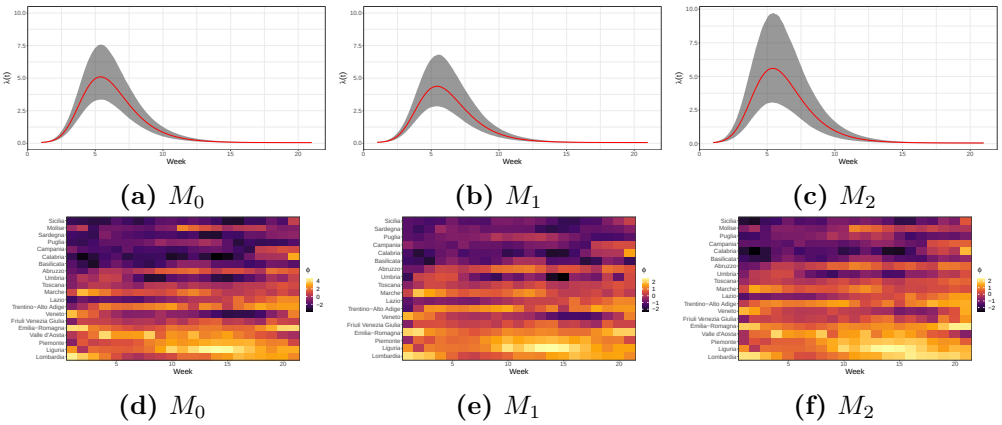
Wave	**Model**	b	r	h	p	s
I	M_0	0.05 (0.04, 0.06)	23 (20, 27)	0.62 (0.60, 0.64)	2.0 (1.5, 2.5)	7.8 (6.3, 9.9)
	M_1	0.06 (0.05, 0.07)	20 (17, 22)	0.62 (0.59, 0.65)	2.2 (1.7, 2.8)	7.9 (5.5, 9.3)
	M_2	0.05 (0.04, 0.06)	26 (21, 31)	0.61 (0.58, 0.65)	2.2 (1.5, 2.9)	7.8 (5.2, 9.3)
II	M_0	$7 \cdot 10^{-5}$ $(1 \cdot 10^{-6}, 1 \cdot 10^{-3})$	158 (143, 172)	3.46 (3.26, 3.63)	23.2 (23.1, 23.3)	0.06 (0.05,0.07)
	M_1	$2 \cdot 10^{-4}$ $(3 \cdot 10^{-5}, 7 \cdot 10^{-3})$	178 (127, 215)	2.72 (2.33, 3.08)	22.9 (22.8, 23.2)	0.09 (0.07,0.10)
	M_2	$4 \cdot 10^{-4}$ $(3 \cdot 10^{-6}, 1 \cdot 10^{-2})$	194 (163, 220)	3.50 (3.20, 3.70)	23.1 (22.9, 23.2)	0.06 (0.05, 0.07)

Table 5.8. Parameters' estimates of the Richards' curve for the first and the second wave.

Table 5.8 shows the estimated parameters of the common Richards in all settings. We here recall that b represents the baseline (endemic rate), r the final size of the outbreak (in terms of cases every $10,000$ residents), h the contagion speed, p the lag-phase and s the asymmetry. We can clearly observe how the second wave is characterized by a larger final outbreak size, a larger endemic rate and a longer lag-phase (meaning the curve approximates exponential growth for a longer time window) in all cases. Furthermore, while the first outbreak was characterized by positive asymmetric behaviour (with a fast and sudden growth followed by a long descending phase) the second wave presented a negative asymmetric evolution,

probably because of the softer lockdown measures undertaken. Indeed, the positive asymmetry characterising the first wave reflects the hard containment measures implemented by the Italian government at the beginning of the epidemic (March 2020), which were gradually loosened. On the contrary, the second wave experienced a negative asymmetry as the prevention policies were mild at the beginning of the second outbreak (mid Summer 2020) and were suddenly strengthened following the abrupt increase of positive cases in late November 2020.

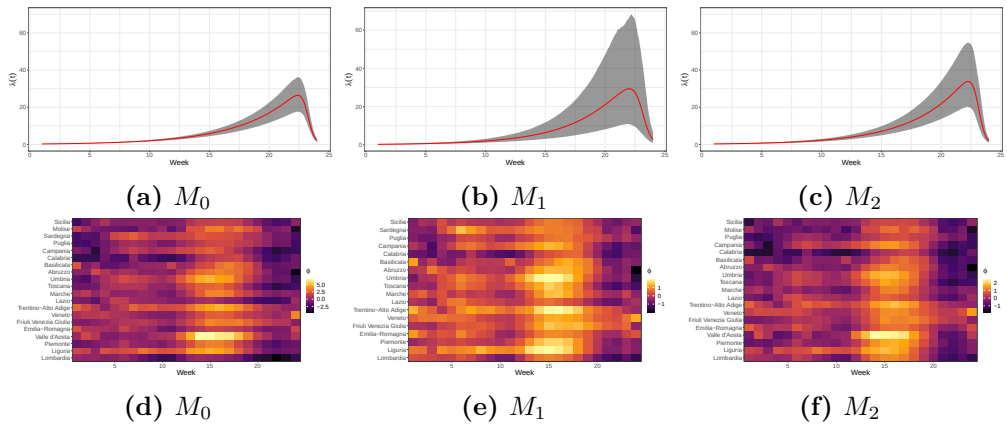
Figure 5.15a-5.15c and Figure 5.16a-5.16c show the estimated common Richards' curves (red solid line) by the proposed models with the associated uncertainty (grey areas represent the 95% credible intervals) for the first and the second wave, respectively. In Figure 5.15d-5.15f and Figure 5.16d-5.16f we instead report the heatmaps of the estimated spatio-temporal effect for each model specification for the first and the second wave, respectively. The estimated common Richards' curve and random effects during the first wave highlight how deviations from the global average presented a strong geographic *clustering effect*, as the number of positive cases increased from the South to the North of Italy. On the contrary, there is relative homogeneity in the deviations of each region from the national epidemic at each time point during the second wave. This means that all regions experienced a similar epidemic trend in terms of shape but different in terms of relative magnitude. Notably, some peculiar regional behaviours are highlighted very clearly. For example, a sudden surge in the contagion between October and November 2020 experienced by Trentino Alto Adige and Umbria. In general, we notice that a larger uncertainty characterizes the common Richards' estimated by M_1 for the second wave compared to the other two models. Considering that the PIWs (see Table 5.9) do not vary much across the proposed dependence structures, this implies a stronger identification of the random effects, i.e. less variability of the random effects. We can then assume that this model provides a better description of the regional heterogeneity in the data.



(a) M_0 **(b)** M_1 **(c)** M_2

(d) M_0 **(e)** M_1 **(f)** M_2

Figure 5.15. Common Richards' curve for the first wave for the different specifications of the random effect (top panels); Posterior mean of the random-effect (bottom panels).



(a) M_0 (b) M_1 (c) M_2

(d) M_0 (e) M_1 (f) M_2

Figure 5.16. Common Richards' curve for the second wave for the different specifications of the random effect (top panels); Posterior mean of the random-effect (bottom panels).

In order to fully compare the different dependence structures on the space-time random effects for the *common* model, we also evaluated the overall fitting performances in terms of two wide-scope indicators: the *Watanabe-Akaike Information Criterion* (WAIC) (Watanabe, 2010), and the *Leave-One-Out* (LOO) score as in Vehtari et al. (2017). These two metrics shall be considered as a proxy of the out-of-sample prediction accuracy, but can be directly computed from the fitted Bayesian model by retaining the log-likelihood values at the different steps of the chain (Vehtari et al., 2017). Results of the estimated validation metrics for the three models for both the first and the second wave are reported in Table 5.9. We notice that results are comparable, with M_1 performing slightly better in terms of RMSE, WAIC and LOO for both waves, guaranteeing a greater or equal coverage in both scenarios. Furthermore, comparing the in-sample and out-of-sample RMSE, we can see how the independent M_0 model strongly overfits on the training set. On the other hand, limiting and driving the behaviour of the random effects through a spatial structure (such as M_0 and M_1) strongly improves the predictive power of the model and leads to way more reliable results. All things considered, M_1 is then chosen as the best dependence structure, also in light of the appealing interpretation of the varying spatial dependence strength as expressed by α in the two waves (see Table 5.7). Hence, the following results will be referred to this model.

Wave	Model	Coverage	RMSE	WAIC	LOO
I	M_0	0.98 (1)	423 (2.1)	2869	3087
	M_1	0.98 (1)	184 (2.3)	2650	2849
	M_2	0.98 (0.99)	272 (2.5)	2774	2982
II	M_0	0.96 (0.99)	12841 (2.8)	4112	4393
	M_1	0.97 (0.99)	910 (4.6)	3820	4080
	M_2	0.92 (0.99)	995 (4.1)	3971	4252

Table 5.9. Validation metrics for the estimated models for the first and the second wave: coverage and rmse out-of-sample (in-sample), WAIC and LOO.

Figures 5.17a-5.17b show the map of the temporal averages of the space time effects of each region: $\bar{\phi}_g = \sum_{t=1}^{T} \hat{\phi}_{gt}/T$. These values can be interpreted as the effect of the over-dispersion on the contagion's spread due to the interactions with the neighbourhood and the auto-regressive term. That allows to verify which regions generally presented an infection rate larger than the national average along the two

waves. As already pointed out, we can notice a stronger geographical clustering during the first wave, which is even more evident looking at the maps than at the heatmaps. Given the low value of the estimated α in the first wave, this effect is not really linked to the superimposed networking structure, but is an inherent characteristic of the data at regional level: the pandemic initially hit stronger the North of Italy and only later slowly spread to the South. On the contrary, the geographic clustering effect vanishes during the second wave and the coloring of the map looks smoother. Regions similarity is actually explained by people exchange and transportation between regions (larger value of α). It is crucial to notice that, differently from what was often reported by the news in Italy, Lombardia did not perform worse in terms of positive cases with respect to the the rest of the country along the second wave (net of the tracking effort and regional offset). We added the maps obtained using M_2 for comparison in Figure D.1a and D.1b, and we only report here that differences were negligible.

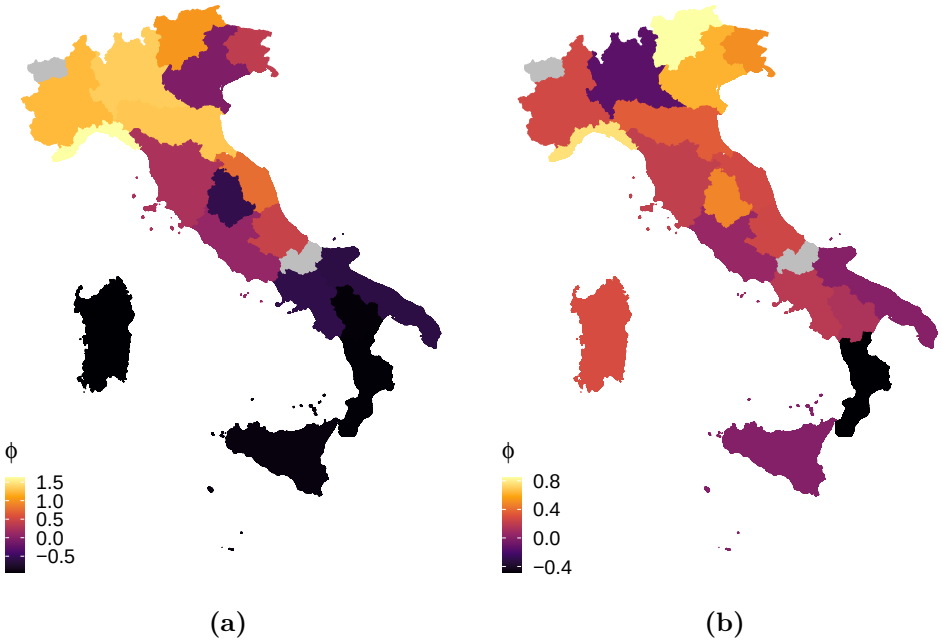
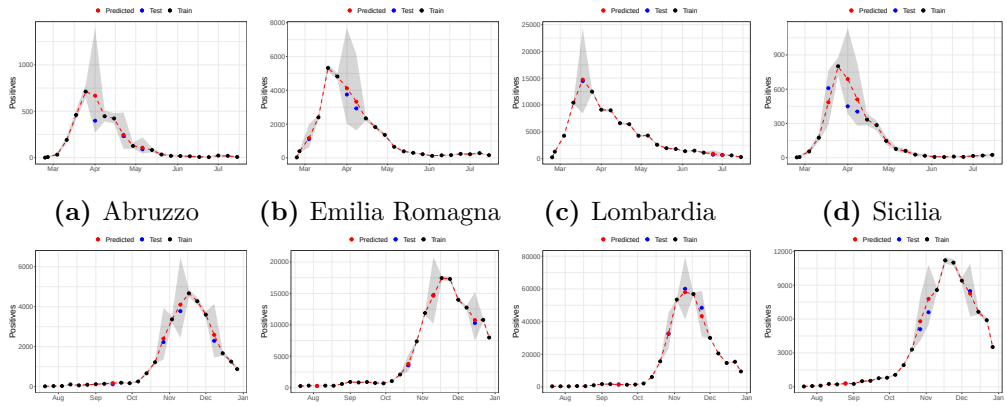


Figure 5.17. Average estimated spatial random effect $\bar{\phi}_g$ by M_1 for the first (a) and the second (b) wave.

We argue that the chosen model is able to reconstruct the *true evolution* of the incidence curve, also at missing points in the series. Some examples are in Figure 5.18a-5.18h. The fit along all time points (in sample and out-of-sample) are plotted together with the 95% posterior predictive intervals for four randomly selected regions (Abruzzo, Emilia Romagna, Lombardia and Sicilia), for the first and the second wave of the epidemic. We can clearly notice how the random effect allows the model to capture the wiggly behaviour of the observed data, and how almost all the observed data fall into the prediction intervals in spite of the large over-dispersion of the observed counts. More importantly, the predictive intervals obviously widen in correspondence of the missing observations, and practically always include the true value, even when this deviates from a typical, expected behaviour (see again Figures 5.18a-5.18h). Given the homogeneity assumptions for Richards' curve for all

regions, this must be mainly due to the dependence structure induced by M_1. Figure D.2 shows specifically the out-of-sample predictive performances, where values in the test set are plotted on the log-scale. Appendix D also includes an evaluation of the forecasting performances (i.e. prediction of future outcomes) of the model, yielding results very coherent to the ones included in this section.



(a) Abruzzo **(b)** Emilia Romagna **(c)** Lombardia **(d)** Sicilia

(e) Abruzzo **(f)** Emilia Romagna **(g)** Lombardia **(h)** Sicilia

Figure 5.18. Observed time series (red dots) and model fit (black solid) with 95% prediction intervals (black dashed) for 4 randomly chosen regions in the first (top panels) and the second wave (bottom panels).

5.2.4 Discussion and further developments

Modelling incident cases poses several issues, ranging from the discrete nature of the observations to the dependence structure of the data across times and neighbouring regions. The proposed generalised logistic growth curve accommodates the main data features, and provides a satisfactory solution for data analysis and prediction under a spatially heterogeneous framework. The proposal is applied to Italian regional data, but it can be applied to data from any other country. Similarly, if data were available at the province and/or municipality level, within-region spatial dependence could be explored, promptly identifying clusters of positive cases.

The spatio-temporal dependence is modeled through the inclusion of region-specific random effects. The inclusion of random effects relaxes the working assumption used in Alaimo Di Loro et al. (2021a), where independence was assumed. Failure of the independence-assumed model to fit the data could be due to misspecification of any of the elements defining the linear predictor. Here, not all possible covariates, such as population density or pollution exposure levels, were considered in model specification. Their joint effect is, however, summarized by latent variables, i.e. the random effects. On the one side, the additional computational burden can be dealt with `Stan` within a Bayesian framework with minor efforts. On the other side, the improvements in the goodness of fit and predictions are evident. As shown in Section 5.2.3, the dependence network plays a crucial role and gives useful insights. Here, we analysed two separate waves in Italy. During the first wave, strict restrictions were applied, strongly limiting mobility across the country, mainly allowing for transportation routes only. As a result, the spread of the contagion was mostly influenced by geographic proximity, with northern regions being more affected by the epidemic than those in the Center and South of Italy. During summer, instead, people took the chance of less restrictive travelling constraints and enjoyed the

summer season having holidays far from their region of residence. This led to a completely different spatial association. The country was not anymore divided into three geographical macro-regions, but a more uniform development of the epidemic was observed. Regions' coloring Figure 5.17b reflects both the type of regional policy in terms of screening and the type of non pharmaceutical measures taken. For example, Calabria did not develop a consistent screening activity, and hence was subject to more strict restrictions than other regions.

These results, together with the considerations about the data collection procedures implemented in Italy and their impacts on data quality in Section 5.1, have important consequences in terms of public health policies. In particular, they make even more clear that the ability to monitor, predict and hence govern a pandemic is directly linked to the availability of high quality data, fully harmonised at regional level. Although in a country like Italy, with a high level of regional autonomy, the data collection procedures are necessarily decentralised, the COVID-19 pandemic has shown the need of an integrated system of health data collection, transmission, storage and dissemination, which must be necessarily centralised in order to avoid the heterogeneity, time misalignment and lack of quality which have characterised the available data in this critical period, and have hampered the possibility to obtain from the data the best possible information to support public decisions. Moreover, all our results also point out that in a country with marked geographical, environmental, social and economic regional differences, like Italy, the management of a pandemic must be coordinated at national (and perhaps even at supranational) level, but implemented through specific measures that take into account regional heterogeneity. At the same time, inter-regional mobility should always be considered as crucial, since it can jeopardise the effectiveness of restrictions imposed at regional level, as shown by the different role played by the spatial component in the second wave compared with the first wave in Italy.

The latent dependence structure consistently aids interpretation. However, it may induce some bias in the predictions if the underling network is misspecified. To avoid such bias, the network might be explicitly modeled, and estimated together with all other parameters in the MCMC machinery. There are some examples of this approach in the recent literature (Rushworth et al., 2017; Ejigu and Wencheko, 2020; Corpas-Burgos and Martinez-Beneito, 2020), and it is a possible further development for our model.

In the future we will also consider weighted spatial structures as in Della Rossa et al. (2020), by specifying $\boldsymbol{W}_1$ as $\left(\boldsymbol{W}_1 + \boldsymbol{W}_1^\top\right)/2$. Indeed, using weighted adjacency matrices may better reflect the underlying similarity among geographical units and either boost or mitigate the neighborhood effect on the mean of each one of them. However, this was not pursued in this paper in order to directly compare unweighted versions of the two spatial graphs. We do not generally expect results to differ extremely when applying the weighted matrix. However, we are aware that the specification of the spatial weights matrix can both affect model fitting and parameter estimation. More importantly, in a recent paper by Duncan et al. (2017), where 17 different specifications of $\boldsymbol{W}$ have been compared to perform spatial smoothing, the model using binary, first-order adjacency weights proved to be an optimal choice for achieving a good model fit. Another important extension would be the development of a space-time model capable of capturing the entire evolution of an epidemic, fitting all waves within the same model specification. That could be done by developing a model based on a mixture of Richards' curves, each capable of describing individual epidemic waves. In particular, a change-point model, in the spirit of Girardi et al. (2021), could be specified under our framework. The unknown change point, which

could be in principle more than one, could be estimated along with all other model parameters. The resulting model is a (constrained) finite mixture model that could be implemented in future research, whose computational burden is not much different from the one considered here. Similarly, assessing the effectiveness of the Italian risk-zones policy during the different waves (Pelagatti and Maranzano, 2021) could also be implemented under the proposed framework, providing further insights to the decision-makers to govern the epidemic spread better. Eventually, to exploit the general idea of dependence in both space and time, we may imagine defining a space-time neighborhood structure linking neighboring regions at different time points. In all mentioned developments, we have to remember that swabs play a crucial role. Hence, we have to imagine a nested, hierarchical model structure where a proper predictive model is added if the prediction of cases becomes a crucial feature.

Further results (e.g. chain diagnostics) are available from the authors upon request and we point again the reader to the public `GitHub` repository available at `https://github.com/minmar94/Covid19-Spatial`.

Acknowledgements

The authors wish to thank an anonymous referee and the associate editor for their suggestions that considerably helped in improving the paper from its previous version. This work has been partially supported by Fondo integrativo speciale per la ricerca (FISR), grant number FISR2020IP_00156.

Chapter 6

Final discussion

> "You can never know everything, and part of what you know is always wrong. Perhaps even the most important part. A portion of wisdom lies in knowing that. A portion of courage lies in going on anyway."
>
> *Robert Jordan*

This dissertation covered the estimation of Bayesian hierarchical models with applications to phenomena in different domains. The presented solutions to each specific problem stemmed from the need of providing a good compromise between the interpretability of the results and the computational burden for prediction and estimation purposes.

Chapter 1 served as an introduction to allow the reader into the view of Bayesian hierarchical models as a comprehensive tool for model building and inference.

Chapter 2 focused on the model formulation and estimation under the Bayesian framework. Firstly, the general equation of a Bayesian hierarchical model was defined. Then, a brief introduction to the main MCMC estimation methods, among which the Gibbs sampler, the Metropolis-Hastings algorithm and the Hamiltonian Monte Carlo was provided. The goal was to introduce the basic ingredients for the understanding of each methodological choice that has been undertaken in the following applications.

Chapter 3 presented a Bayesian Beta regression model, including a variable selection step, for the modeling of food losses percentages of cereals and derived products at the country-commodity level, now published by Mingione et al. (2021b). The proposed distributional framework already represented a substantial improvement over the previous approaches. In addition, the scalability of the hierarchical modeling structure easily allows for the application of the same model to other food groups or at single steps of the supply chain, when more data will be available. More importantly, although being computationally demanding, the proposed spike and slab prior provides an interpretable measure of the importance of the variables explaining food losses dynamics worldwide. This could be utterly helpful to support decisions on interventions, investments and policy-making towards the achievement of SDG 12.3. For further discussion on how countries can use the FLI as a suitable tool for policy-making, the author points to Fabi and English (2019) and Koester and Galaktionova (2021). We are currently working in collaboration with FAO on the development of an interactive dashboard to inform the interested users on the global state of Food Loss and Waste. We would like to provide useful summary statistics at different levels of detail (e.g. country, SDG region, commodity, etc.), and allow scenario building based on the model in Mingione et al. (2021b) to help

decision-makers implementing timely prevention policies.

Chapter 4 included the fast estimation of Gaussian processes, when the dependence among observations occurs in the temporal domain. The methodology was applied to high-frequency sampled accelerometer data and it was used for the estimation of physical activity trajectories at the individual level. Although not directly modeled within the dependence structure, spatial effects were accounted using spline regression. The disentanglement of the spatial and temporal effects appears successful, providing excellent results both in terms of predictive performances and interpretability. The NNGP setting was profitably adapted to the univariate case and a specific algorithm was implemented, improving on the computational performances of the already existing tools.
The proposal is currently published as a preprint by Alaimo Di Loro et al. (2021b), and represents a novelty in the context of human activity tracking in a free-living environment, having the potential of being a game-changer for the new monitoring technologies. Indeed, it provides a decently efficient and precise tool to impute gaps and/or predict biometrical variables at unobserved time-points and locations. Low-cost tracking devices (e.g. smartphone apps) already present these issues and can suddenly benefit from such modeling tool. Future work on this topic would consider the inclusion of spatial varying covariates (e.g. normalized difference vegetation index (NDVI), distance from parks, elevation, etc.) which could likely affect physical activity levels of individuals in open spaces. The consideration of alternatives ways of including spatial dependence is of no less importance. In this first modeling attempt, we settled for spatial regression using splines, however spatio-temporal Gaussian processes (Datta et al., 2016a) can be exploited to model the latent component. Even more specific solutions could be envisioned by including dependence among trajectories (points near in space may belong to completely different paths and be distant in time) via convolved Gaussian processes (Alvarez and Lawrence, 2011). Eventually, we are currently working on a stable version of the collapsed algorithm to make available an `R` package for the estimation of NNGP models to actigraph data.

Chapter 5 described a growth model for the epidemiological incident cases using the Richards' curve. The methodology is published by Alaimo Di Loro et al. (2021a), and was firstly applied to COVID-19 incident cases (e.g. daily positives and daily deceased) at both the national and regional level during the Italian first epidemic wave. Although results were promising, and predictions proved to be trustworthy also for the second wave, the proposal presented two main limitations: (i) regions were assumed to be independent among each other; (ii) likelihood-based inference produced unstable results due to the complex parametrization of the mean term. Hence, we extended the proposed model to the Bayesian setting and included a latent component to deal with the spatial dependence among geographical units using a CAR prior. An efficient implementation of such model was provided to hasten the computational time, exploiting the sparsity of the adjacency matrix driving the spatial dependence. The proposal is published by Mingione et al. (2021a), and was applied to Italian regional weekly positive COVID-19 cases during the first and the second wave of the epidemic. Results were improved with respect to the previous approach, and several features of the epidemic wave, such as peak time and height, could be properly detected. Nevertheless, both the modeling proposals depend on the suitable choice of epidemic waves' time windows, and therefore can be applied to each one of them separately. In this respect, we are working on an extension of the model presented in Section 5.2 which is capable of capturing the entire evolution of an epidemic, considering all waves within the same model specification by means of a mixture of Richards' curves.

In general, all the results point out that in a country like Italy, with marked geographical, environmental, social and economic regional differences, the management of a pandemic must be coordinated at national level, but implemented through specific measures that take into account regional heterogeneity. This lesson can be summarised in the motto "centralize information, localize decisions", which should be taken seriously by decision-makers when dealing with the current COVID-19 pandemic, and to prepare for future, unwelcome, but not unlikely pandemics.

Summing up, all the proposed applications showed that the statistician has become altogether sufficiently knowledgeable in many subject matter to keep pace with the technological enhancements in several fields and the broad range of applications for which his/her skills may be required. By means of hierarchical modeling, the presented solutions proved to be reliable and provided a – partial but – accurate explanation of the phenomenon under consideration. The adoption of a Bayesian approach always helped the model formulation, and allowed for a proper quantification of the uncertainty surrounding the quantities of interest for each of the proposed applications. This was consistently highlighted in commenting the results, which were validated with common sense and public awareness of science. However, there are still open research questions and lot of potential for further developments, as pointed out in each specific discussion at the end of the previous chapters. In doing so, the good statistician must always keep in mind that doing statistics is like doing crosswords, except that he/she can never know for sure whether he/she found the solution.

Appendix A

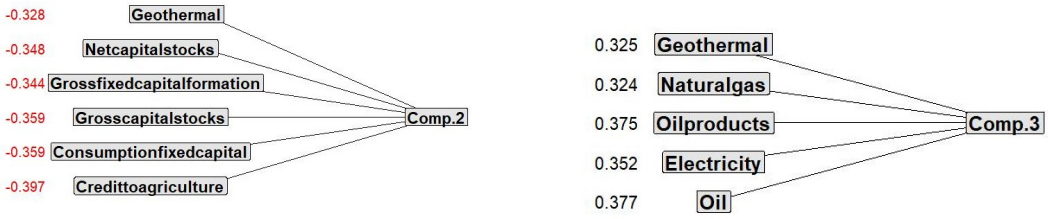
Measuring and modeling food losses

A.1 Explanatory variables

Variable	Source	Description
Lead	World Bank Pink Sheets	prices, annual average (nominal)
Coal	World Bank Pink Sheets	”
Copper	World Bank Pink Sheets	”
Nickel	World Bank Pink Sheets	”
Crude Oil	World Bank Pink Sheets	”
Crude Petrol	World Bank Pink Sheets	”
Aluminium	World Bank Pink Sheets	”
Zinc	World Bank Pink Sheets	”
Platinum	World Bank Pink Sheets	”
Silver	World Bank Pink Sheets	”
Gold	World Bank Pink Sheets	”
Iron	World Bank Pink Sheets	”
Tin	World Bank Pink Sheets	”
Potash	World Bank Pink Sheets	”
Urea	World Bank Pink Sheets	”
Phosrock	World Bank Pink Sheets	”
TSP	World Bank Pink Sheets	”
DAP	World Bank Pink Sheets	”
Gas	World Bank Pink Sheets	”
Natural Gas	International Energy Agency	”
Heat	International Energy Agency	”
Geothermal	International Energy Agency	”
Oil	International Energy Agency	”
Oil Product	International Energy Agency	”
Biofuels	International Energy Agency	”
Electricity	International Energy Agency	”
Credit to Agriculture	FAOSTAT	/
Net Capital Stocks	FAOSTAT	/
Gross Fixed Capital Formation	FAOSTAT	/
Gross Capital Stocks	FAOSTAT	/
Consumption Fixed Capital	FAOSTAT	/
Spending on Agriculture	IFPRI	Share of agricultural GDP
Logistic Performance Index	World Bank	Composite indicator evaluating trade logistics
Rainfall	World Bank	Yearly average in mm
Temperature	World Bank	Yearly average in Celsius

Table A.1. Explanatory variables, their source and a brief description. Each variable is coloured with respect to its category: building materials (red), energy prices (green), fertilizers (brown), economic factors (grey), transportation and logistics (blue), weather (orange).

A.2 Dimensional reduction of the design matrix



(a) (b)

Figure A.1. Path diagram representation of PCA, showing variables associated to the 2nd component (a) and the third component (b).

A.3 Model implementation using JAGS

```
model {
# Likelihood
for(i in 1:n){
 Y[i] ~ dbeta(alpha[i], delta[i])
 alpha[i] <- mu[i] * phi
 delta[i]  <- (1-mu[i]) * phi
 logLik[i] <- (alpha[i]-1)*log(Y[i]) + (delta[i]-1)*log(1-Y[i])
 logit(mu[i]) <- zbetaYear[CTRY[i]]*Year[i] + inprod(X[i,],zbeta[]) +
                 etacoucomm[CC[i]]
}
# Priors country-commodity
for(k in 1:neff){
 etacoucomm[k]~dnorm(0,tau)
}
# Prior for tau
 tau ~ dgamma(4, 0.1)
# Prior for phi
 phi ~ dunif(5,150)
### SPIKE AND SLAB ###
# Prior for beta and gamma
for(ctry in 1:nctry){
 ppYear[ctry] ~ dbeta(5,5)
 gammaYear[ctry] ~  dbern(ppYear[ctry])
 betaYear[ctry] ~ dnorm(0,0.001)
 zbetaYear[ctry] = betaYear[ctry]*gammaYear[ctry]
}
for(j in 1:p){
 beta[j] ~ dnorm(0,0.001)
 pp[j] ~ dbeta(5,5)
 gamma[j] ~  dbern(pp[j])
 zbeta[j] =  beta[j]*gamma[j]
}
### HORSESHOE ###
# Prior for beta and gamma
# for(ctry in 1:nctry){
# shrinkYear[ctry] ~ dt(0,1,1)T(0,)
# betaYear[ctry] ~ dnorm(0, 1/(shrinkYear[ctry]*global))
#}
# for(j in 1:p){
# shrink[j] ~ dt(0,1,1)T(0,)
# beta[j] ~ dnorm(0, 1/(shrink[j]*global))
#}
# Prior for global shrinkage
# global ~ dt(0,1,1)T(0,)
}
```

Algorithm A.1. Example code for food losses estimation using JAGS.

```
model {
# Train
for(i in 1:ntrain){
 Ytrain[i]  ~ dbeta(alphatrain[i], deltatrain[i])
 alphatrain[i] <- mutrain[i] * phi
 deltatrain[i]  <- (1-mutrain[i]) * phi
 logit(mutrain[i]) <- betaYear[CTRYtrain[i]]*Yeartrain[i] +
                      inprod(Xtrain[i,],beta[]) +
                      etacoucomm[CCtrain[i]]
 }
# Prediction - test
for(i in 1:ntest){
 Ytest[i]  ~ dbeta(alphatest[i], deltatest[i])
 alphatest[i] <- mutest[i] * phi
 deltatest[i]  <- (1-mutest[i]) * phi
 logit(mutest[i]) <- betaYear[CTRYtest[i]]*Yeartest[i] +
                     inprod(Xtest[i,],beta[]) +
                     etacoucomm[CCtest[i]]
 }
for(k in 1:neff){
  etacoucomm[k]~dnorm(0,tau)
 }
# Prior for tau
 tau ~ dgamma(4, 0.1)
# Prior for phi
 phi ~ dunif(5,150)
### SPIKE AND SLAB ###
# Prior for beta and gamma
for(ctry in 1:nctry){
  betaYear[ctry] ~ dnorm(0,0.001)
 }
for(j in 1:p){
  beta[j] ~ dnorm(0,0.001)
 }
### HORSESHOE ###
# Prior for beta and gamma
# for(ctry in 1:nctry){
# betaYear[ctry] ~ dnorm(0, 0.001)
#}
#for(j in 1:p){
# shrink[j] ~ dt(0,1,1)T(0,)
# beta[j] ~ dnorm(0, 1/(shrink[j]*global))
#}
# Prior for global shrinkage
# global ~ dt(0,1,1)T(0,)
}
```

Algorithm A.2. Example code for food losses prediction using JAGS.

A.4 Further results

Model	Variable	$\hat{\gamma}$	$q_{.025}$	Mean	$q_{.975}$
M1	*Biofuels*	1	0.345	0.449	0.558
	SpendingOnAgri	1	-0.231	-0.190	-0.149
	Comp.2	1	0.136	0.178	0.219
	Comp.3	0.56	0.037	0.081	0.125
M2	*Biofuels*		0.332	0.439	0.547
	SpendingOnAgri		-0.224	-0.180	-0.136
	Comp.2		0.132	0.173	0.213
	Comp.3		0.035	0.078	0.120
	Temperature		-0.0617	-0.021	-0.005

Table A.2. Selected variables with the two different priors and 95% posterior credibility intervals of the β_k^*.

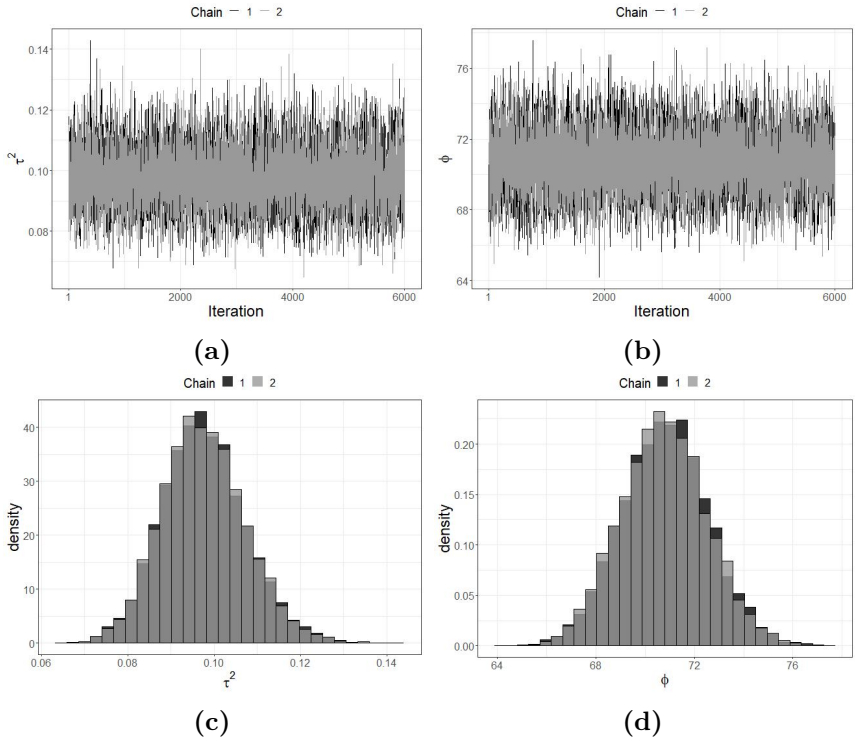


Figure A.2. Traceplot and posterior density of the estimated variance components by M1: variance of the random effects τ^2 (a) (c) and variance of the outcome ϕ (b) (d).

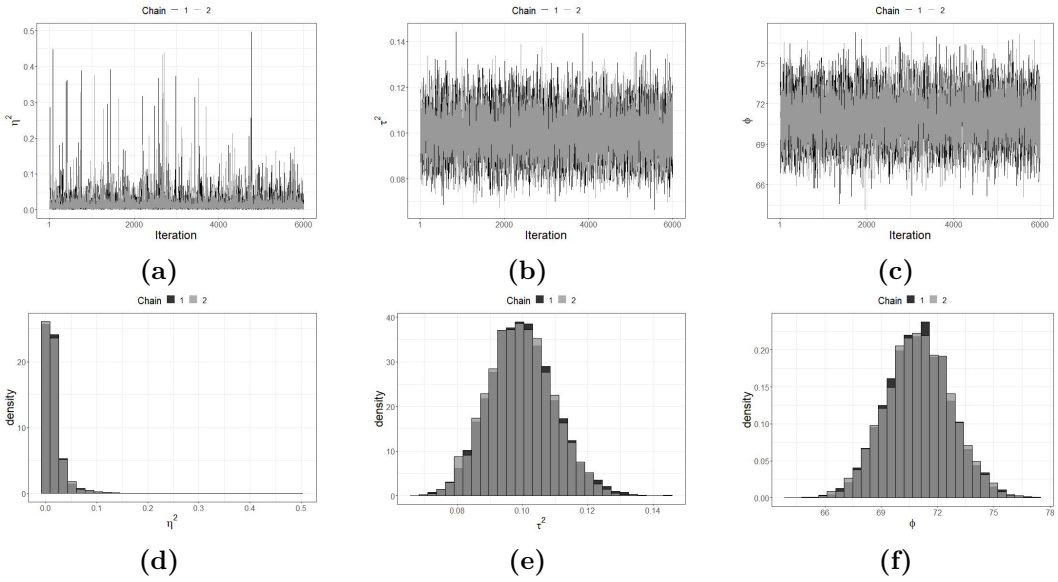
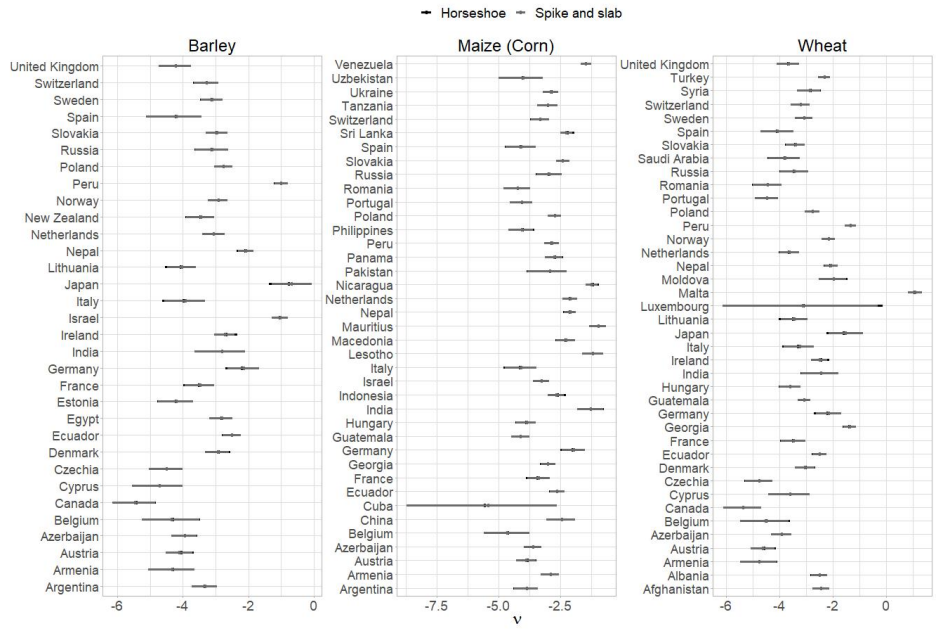
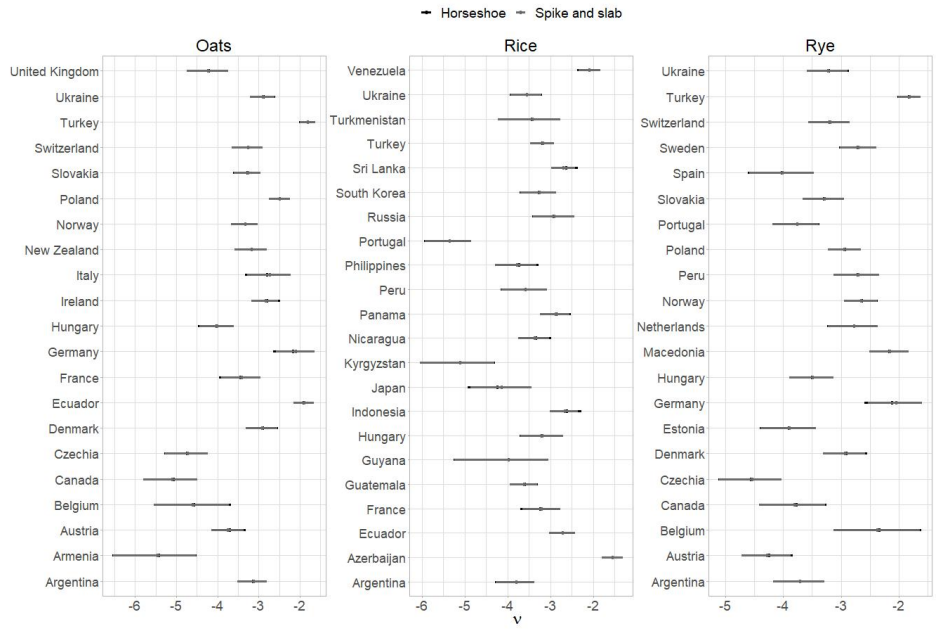


Figure A.3. Traceplot and posterior density of the estimated variance components by M2: global shrinkage parameter η^2 (a) (c), variance of the random effects τ^2 (b) (d) and variance of the outcome ϕ (c) (e).

Country-crop random effects

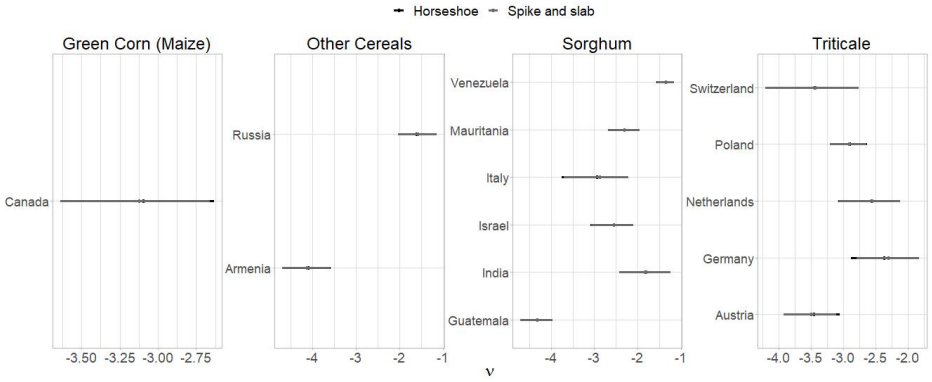


(a)

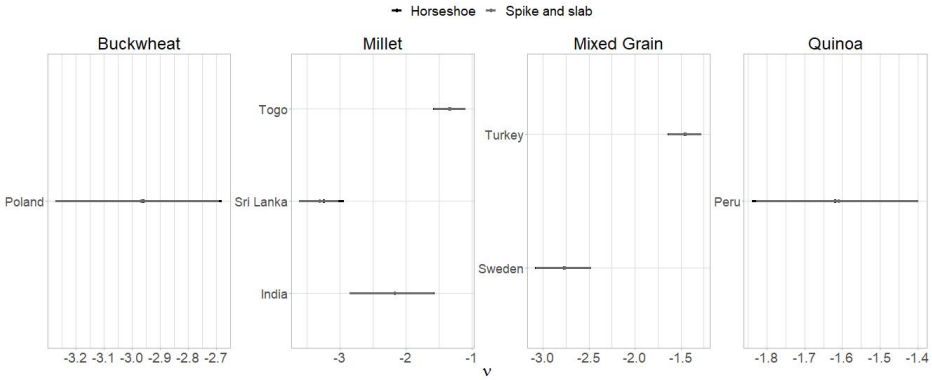


(b)

Figure A.4. M1 and M2 random effects point estimates and 95% credible intervals for each cereal commodity (a) and (b).



(c)



(d)

Figure A.4. M1 and M2 random effects point estimates and 95% credible intervals for each cereal commodity (a) and (b).

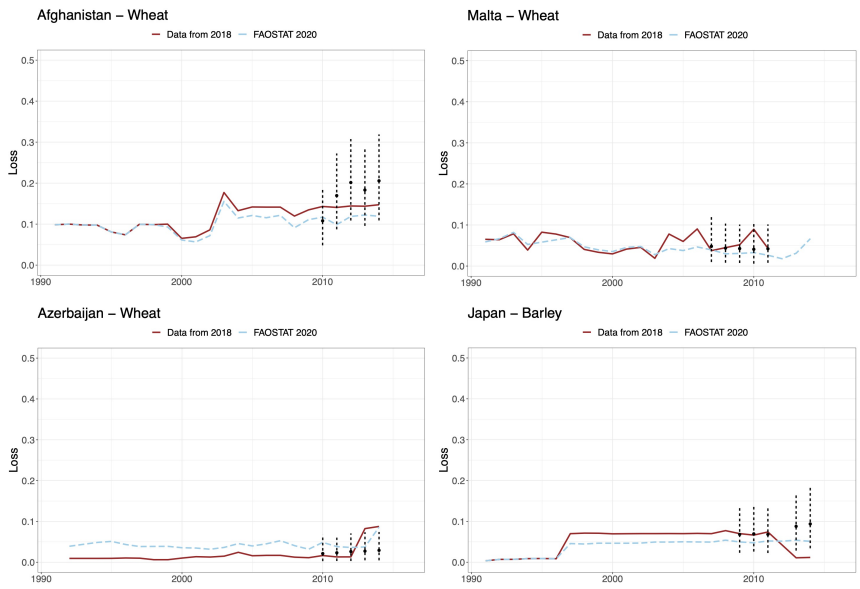


Figure A.5. Comparison of model predictions obtained using old losses data in FAOSTAT and the current stored values.

Appendix B

Bayesian hierarchical modeling and analysis for physical activity trajectories using actigraph data

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., 2017). We did not consider the Response NNGP because it does not recover the latent component. The first one is described in Section B.1 and includes simulated observations for one single individual; the second one includes simulated observations for multiple individuals and is described in Section B.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`.

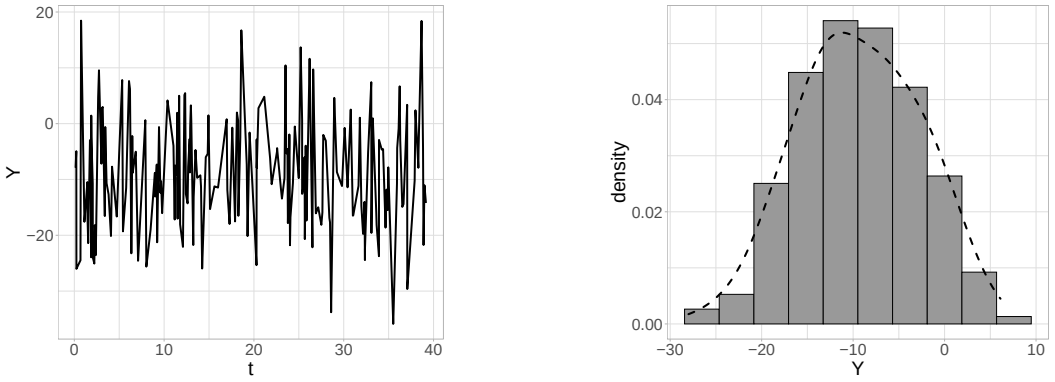
B.1 Experiment 1

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:

$$\mathbf{Cov}_\theta\left[Y(t), Y(t')\right] = c_\theta(t, t') = \sigma^2 e^{(\phi|t-t'|)}, \quad \sigma^2, \phi \in \mathbb{R}^+, \tag{B.1}$$

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 B.1a and B.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. $\{\tilde{\boldsymbol{s}}_j\}_{j=1}^{\top} = \{(t_j, 0)\}_{j=1}^{\top}$). 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 Gamma prior $\mathcal{IG}(2, 2)$, and the decay parameter ϕ was ascribed a $\mathcal{G}(1, 1)$. 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



(a) **(b)**

Figure B.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	0.95	0.99		0.96	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		4.78	4.47	
Run time (h)	1.77			1.86		

Table B.1. Parameter estimates, predictive validation and fitting times (hours) on the simulated dataset for all the considered models.

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 B.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).

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

$T \times 10^3$	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 B.2. Time (in seconds) of one MCMC iteration for the two considered algorithms with increasing sample size (T) and fixed $m = 30$.

time of one sampler iteration, fixing $m = 30$ for all the considered algorithms has been recorded for $\tilde{M} = 100$ times. Figure B.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 B.2.

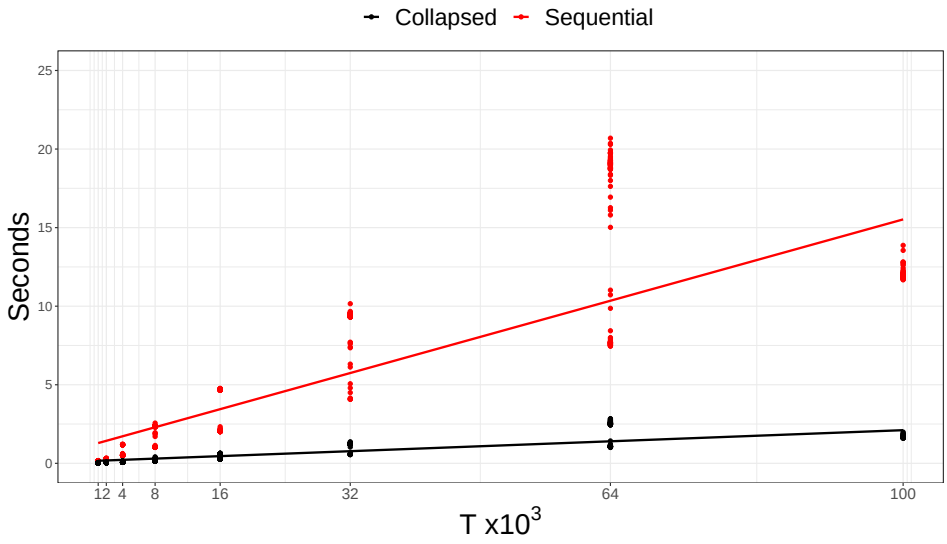


Figure B.2. Time elapsed (in seconds) for 1 MCMC iteration for the two considered algorithms with increasing sample size T and fixed $m = 30$.

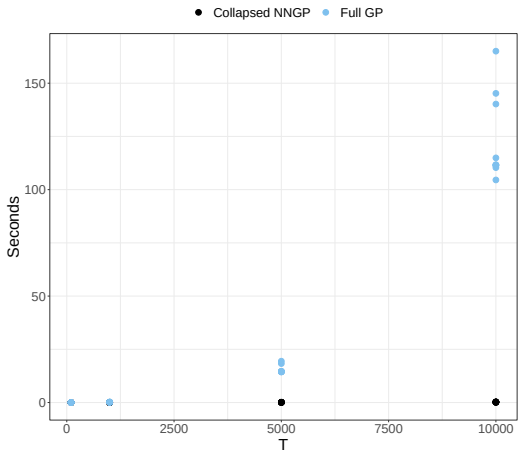


Figure B.3. Time elapsed (in seconds) for 1 MCMC iteration for the Collapsed NNGP and the Full GP with increasing sample size T.

Additionally, we wanted to quantify the computational advantage of the proposed NNGP-based collapsed algorithm over the standard MCMC update of the full GP model (Cressie, 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 B.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.

B.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 B.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 B.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

			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}	-0.82	-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 B.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.

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 $\mathcal{G}(1, 1)$. 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.

Appendix C

Nowcasting COVID-19 incidence indicators during the Italian first outbreak

C.1 Gradients

In order to make the optimization procedure robust, gradients and Hessians used for the estimation (optimization routine on the log-likelihood) have been computed analytically. This section provides insights about their derivation for the log-likelihoods at hand. For the sake of clarity, in the sequel, we will invert the previous notation and denote the functions of interest as functions of the parameters, given the observed time points: e.g. $\tilde{\mu}_{\boldsymbol{\theta}}(t)$ becomes $\tilde{\mu}_t(\boldsymbol{\theta})$. We first provide the computations for the gradient of the log-likelihood for both Poisson and Negative Binomial distributions by considering their mean function $\tilde{\mu}_t(\boldsymbol{\theta})$ as a whole. Afterwards, we show the gradients and introduce the Hessians specific to $\tilde{\lambda}_t(\boldsymbol{\gamma})$, as it is the most cumbersome component of the mean to derive with respect to its parameters.

Poisson Gradient

Let q denote any of the elements of $\boldsymbol{\theta}$, vector of parameters characterizing the mean function $\tilde{\mu}_t(\boldsymbol{\theta})$. The generic derivative with respect to the component q of $\boldsymbol{\theta}$ for the Poisson log-likelihood $\text{Poi}(\tilde{\mu}_t(\boldsymbol{\theta}))$ is:

$$\begin{aligned}\frac{\partial}{\partial q} l_{Poi}(\boldsymbol{\gamma}|\mathbf{y}) &= -\sum_{t=1}^{T} \frac{\partial}{\partial q}\tilde{\mu}_t(\boldsymbol{\theta}) + \sum_{t=1}^{T} y_t \frac{\partial}{\partial q}\log(\tilde{\mu}_t(\boldsymbol{\theta})) = \\ &= -\sum_{t=1}^{T} \frac{\partial}{\partial q}\tilde{\mu}_t(\boldsymbol{\theta}) + \sum_{t=1}^{T} y_t \frac{1}{\tilde{\mu}_t(\boldsymbol{\theta})}\frac{\partial}{\partial q}\tilde{\mu}_t(\boldsymbol{\theta}).\end{aligned} \tag{C.1}$$

Negative Binomial Gradient

The Negative Binomial $\text{NB}\left(\nu, \mu_t(\boldsymbol{\theta})\right)$ presents the additional parameter ν, which does not affect the mean function but controls for the dispersion. In the following, we provide the first derivative with respect to ν and with respect to the generic element q of $\boldsymbol{\theta}$, respectively.

The first derivative with respect to ν of the log-likelihood is:

$$\begin{aligned}\frac{\partial}{\partial \nu} l_{NB}(\nu, \boldsymbol{\theta}|\mathbf{y}) &= T(\log(\nu) - \psi(\nu)) + \\ &+ \sum_{t=1}^{T}\left(\psi(\nu + y_t) - log\left(\mu_t(\boldsymbol{\theta}) + \nu\right) + \frac{\mu_t(\boldsymbol{\theta}) - y_t}{\mu_t(\boldsymbol{\theta}) + \nu}\right)\end{aligned} \tag{C.2}$$

where $\psi(\cdot)$ denotes the *digamma* function. The generic derivative with respect to q of the log-likelihood is:

$$\frac{\partial}{\partial q} l_{NB}(\nu, \boldsymbol{\gamma}|\mathbf{y}) = -\sum_{t=1}^{T} \frac{y_t + \nu}{\tilde{\mu}_t(\boldsymbol{\theta}) + \nu} \frac{\partial}{\partial q} \tilde{\mu}_t(\boldsymbol{\theta}) + \sum_{t=1}^{T} \frac{y_t}{\tilde{\mu}_t(\boldsymbol{\theta})} \frac{\partial}{\partial q} \tilde{\mu}_t(\boldsymbol{\theta}). \tag{C.3}$$

Richards' Gradient

Derivation of the gradient $\tilde{\mu}_t(\boldsymbol{\theta})$ can be obtained by deriving separately (but appropriately) each of the pieces composing it. Computations are straightforward for all components, but for the Richards' first differences parameters, which can in turn be divided as:

$$\frac{\partial}{\partial \gamma_i} \tilde{\lambda}_t(\boldsymbol{\gamma}) = \frac{\partial}{\partial \gamma_i} \lambda_t(\boldsymbol{\gamma}) - \frac{\partial}{\partial \gamma_i} \lambda_{t-1}(\boldsymbol{\gamma})$$

The Richards' function gradient is composed of the following four terms:

$$\nabla \lambda_t(\boldsymbol{\gamma}) = \left[\frac{\partial}{\partial r} \lambda_t(\boldsymbol{\gamma}),\ \frac{\partial}{\partial h} \lambda_t(\boldsymbol{\gamma}),\ \frac{\partial}{\partial p} \lambda_t(\boldsymbol{\gamma}),\ \frac{\partial}{\partial s} \lambda_t(\boldsymbol{\gamma}) \right]^\top$$

which can be computed as follows.

$$\frac{\partial}{\partial r} \lambda_t(r) = \frac{\partial}{\partial r} \left(b + \frac{r}{(1 + 10^{h(p-t)})^s} \right) = \frac{1}{(1 + 10^{h(p-t)})^s},$$

$$\begin{aligned} \frac{\partial}{\partial h} \lambda_t(h) &= \frac{\partial}{\partial h} \left(b + \frac{r}{(1 + 10^{h(p-t)})^s} \right) = \\ &= -r \cdot s \cdot (1 + 10^{h(p-t)})^{-s-1} 10^{h(p-t)} (p - t) \log(10), \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial p} \lambda_t(p) &= \frac{\partial}{\partial p} \left(b + \frac{r}{(1 + 10^{h(p-t)})^s} \right) = \\ &= -r \cdot s \cdot (1 + 10^{h(p-t)})^{-s-1} 10^{h(p-t)} h \log(10), \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial s} \lambda_t(s) &= \frac{\partial}{\partial s} \left(b + \frac{r}{(1 + 10^{h(p-t)})^s} \right) = \\ &= -r \cdot \left(1 + 10^{h(p-t)}\right)^{-s} \log\left(1 + 10^{h(p-t)}\right) \end{aligned}$$

Log-scale

In the `R` implementation, the log-likelihood has been parametrized on the log-scale for all the parameters defined on $\mathbb{R}^+$ in order to ease the optimization process under the positivity constraint. This means that given $q \in \{b, r, p, s\}$, the log-likelihood uses $\log(q) = v$, where $q = e^v$. This implies that, when we differentiate, we have to take into account the Jacobian as a result of the transformation:

$$\frac{\partial}{\partial v} \lambda_t(\boldsymbol{\gamma}) = \frac{\partial}{\partial e^v} \lambda_t(\boldsymbol{\gamma}) \frac{\partial e^v}{\partial v} = \frac{\partial}{\partial e^v} \lambda_t(\boldsymbol{\gamma}) \cdot e^v = \frac{\partial}{\partial q} \lambda_t(\boldsymbol{\gamma}) \cdot q. \tag{C.4}$$

Therefore, each derivative must be multiplied by $e^v = q$.

C.2 Hessians

Hessians used for the estimation procedure of the model (optimization routine on the log-likelihood) have been computed analytically. In the sequel, we first provide the Hessian for the log-likelihod of Poisson and Negative Binomial by considering $\lambda_t(\gamma)$ as a whole.

Poisson Hessian

Let q and f denote any pair of the parameters characterizing the mean function $\tilde{\mu}_t(\boldsymbol{\theta})$.

The mixed second derivative with respect to the components q and f of $\boldsymbol{\theta}$ for the Poisson log-likelihood is:

$$\frac{\partial^2}{\partial x f} l_{Poi}(\boldsymbol{\theta}|\mathbf{y}) = \sum_{t=1}^{T} \frac{y_t - \tilde{\mu}_t(\boldsymbol{\theta})}{\tilde{\mu}_t(\boldsymbol{\theta})} \frac{\partial^2}{\partial q f} \tilde{\mu}_t(\boldsymbol{\theta}) - \sum_{t=1}^{T} \frac{y_t}{\tilde{\mu}_t(\boldsymbol{\theta})^2} \frac{\partial}{\partial q} \tilde{\mu}_t(\boldsymbol{\theta}) \frac{\partial}{\partial f} \tilde{\mu}_t(\boldsymbol{\theta}).$$

The second derivative with respect to the components q for the Poisson log-likelihood is:

$$\frac{\partial^2}{\partial q^2} l_{Poi}(\boldsymbol{\theta}|\mathbf{y}) = \sum_{t=1}^{T} \frac{y_t - \tilde{\mu}_t(\boldsymbol{\theta})}{\tilde{\mu}_t(\boldsymbol{\theta})} \frac{\partial^2}{\partial q^2} \tilde{\mu}_t(\boldsymbol{\theta}) - \sum_{t=1}^{T} \frac{y_t}{\tilde{\mu}_t(\boldsymbol{\theta})^2} \left(\frac{\partial}{\partial q} \tilde{\mu}_t(\boldsymbol{\theta}) \right)^2.$$

Negative Binomial Hessian

Let q and f denote any pair of the parameters characterizing the mean function $\mu_t(\boldsymbol{\theta})$.

The mixed second derivative with respect to q and f of the Negative Binomial log-likelihood is:

$$\begin{aligned} \frac{\partial^2}{\partial x f} l_{NB}(\boldsymbol{\theta}|\mathbf{y}) = \sum_{t=1}^{T} \left(\frac{y_t + \nu}{(\mu_t(\boldsymbol{\theta}) + \nu)^2} - \frac{y_t}{\mu_t(\boldsymbol{\theta})^2} \right) \frac{\partial}{\partial q} \tilde{\mu}_t(\boldsymbol{\theta}) \frac{\partial}{\partial f} \tilde{\mu}_t(\boldsymbol{\theta}) \quad + \\ + \sum_{t=1}^{T} \left(\frac{y_t}{\tilde{\mu}_t(\boldsymbol{\theta})} - \frac{y_t + \nu}{\tilde{\mu}_t(\boldsymbol{\theta}) + \nu} \right) \frac{\partial^2}{\partial q f} \tilde{\mu}_t(\boldsymbol{\theta}). \end{aligned}$$

The second derivative with respect to q of the Negative Binomial log-likelihood is:

$$\begin{aligned} \frac{\partial^2}{\partial q^2} l_{NB}(\boldsymbol{\theta}|\mathbf{y}) = \sum_{t=1}^{T} \left(\frac{y_t + \nu}{(\tilde{\mu}_t(\boldsymbol{\theta}) + \nu)^2} - \frac{y_t}{\tilde{\mu}_t(\boldsymbol{\theta})^2} \right) \left(\frac{\partial}{\partial q} \tilde{\mu}_t(\boldsymbol{\theta}) \right)^2 \quad + \\ + \sum_{t=1}^{T} \left(\frac{y_t}{\tilde{\mu}_t(\boldsymbol{\theta})} - \frac{y_t + \nu}{\tilde{\mu}_t(\boldsymbol{\theta}) + \nu} \right) \frac{\partial^2}{\partial q^2} \tilde{\mu}_t(\boldsymbol{\theta}). \end{aligned}$$

In the Negative Binomial case, we must recall the presence of the additional parameter ν. The second derivative with respect to ν of the Negative Binomial log-likelihood is:

$$\frac{\partial^2}{\partial \nu^2} l_{NB}(\boldsymbol{\theta}|\mathbf{y}) = T \left(\frac{1}{\nu} - \psi'(\nu) \right) + + \sum_{t=1}^{T} \left(\psi'(\nu + y_t) - \frac{1}{\tilde{\mu}_t(\boldsymbol{\theta}) + \nu} - \frac{\tilde{\mu}_t(\boldsymbol{\theta}) - y_t}{(\tilde{\mu}_t(\boldsymbol{\theta}) + \nu)^2} \right)$$

where $\psi(\cdot)$ and $\psi'(\cdot)$ denote the *digamma* and the *trigamma* function, respectively.

The mixed derivative with respect to ν and the generic element q of $\boldsymbol{\gamma}$ is:

$$\frac{\partial^2}{\partial \nu q} l_{NB}(\boldsymbol{\gamma}|\mathbf{y}) = \sum_{t=1}^{T} \frac{y_t - \tilde{\mu}_t(\boldsymbol{\theta})}{(\tilde{\mu}_t(\boldsymbol{\theta}) + \nu)^2} \frac{\partial}{\partial q} \tilde{\mu}_t(\boldsymbol{\theta})$$

Richards' Hessian

As for the gradient, the same holds for the Hessian of the first differences of the Richards function, which would be only one interesting computation to show. Also here:

$$\frac{\partial^2}{\partial \gamma_i \gamma_j} \tilde{\lambda}_t(\boldsymbol{\gamma}) = \frac{\partial^2}{\partial \gamma_i \gamma_j} \lambda_t(\boldsymbol{\gamma}) - \frac{\partial^2}{\partial \gamma_i \gamma_j} \lambda_{t-1}(\boldsymbol{\gamma})$$

In particular, the resulting Hessian is a 4×4 matrix such that:

$$\left[\mathbf{H}\left(\lambda_t(\boldsymbol{\gamma})\right)\right]_{ij} = \frac{\partial^2}{\partial \boldsymbol{\gamma}_i \boldsymbol{\gamma}_j} \lambda_t(\boldsymbol{\gamma}), \quad i, j \in \{1, \ldots, 4\}.$$

Computations are straightforward for most of the terms, but the final result counts 10 terms (the Hessian matrix is symmetric) and some of those terms are cumbersome to report. Therefore, we won't include these here. The reader, if interested, is invited to contact the authors for further details on their detailed computation.

Log-scale

In the `R` implementation, the log-likelihood has been parametrized on the log-scale for all the parameters defined on $\mathbb{R}^+$ in order to ease the optimization process under the positivity constraint. This means that given two generic elements, say q and f, of the parameters' vector $\boldsymbol{\gamma}$, the log-likelihood uses $\log(q) = v$ and $\log(f) = u$, where $q = e^v$ and $f = e^u$. The Jacobian inclusion has two implications on the Hessian. When computing the mixed derivative, we need to account for the transformation of both terms (if both are on the log scale):

$$\begin{aligned}
\frac{\partial^2}{\partial v \partial u} \lambda_t(\boldsymbol{\gamma}) &= \frac{\partial}{\partial v} \left(\frac{\partial}{\partial u} \lambda_t(\boldsymbol{\gamma}) \right) = \frac{\partial}{\partial v} \left(\frac{\partial}{\partial e^u} \lambda_t(\boldsymbol{\gamma}) \frac{\partial e^u}{\partial u} \right) = \\
&= \frac{\partial}{\partial v} \left(\frac{\partial}{\partial e^u} \lambda_t(\boldsymbol{\gamma}) \cdot e^u \right) = \frac{\partial}{\partial e^v} \left(\frac{\partial}{\partial e^u} \lambda_t(\boldsymbol{\gamma}) \cdot e^u \right) \frac{\partial e^v}{\partial v} = \\
&= \frac{\partial}{\partial e^v} \left(\frac{\partial}{\partial e^u} \lambda_t(\boldsymbol{\gamma}) \cdot e^u \right) \cdot e^v = \frac{\partial}{\partial q} \frac{\partial}{\partial s} \lambda_t(\boldsymbol{\gamma}) \cdot q \cdot f.
\end{aligned} \tag{C.5}$$

Therefore, each mixed derivative $\frac{\partial^2}{\partial x s} \lambda_t(\boldsymbol{\gamma})$ must be multiplied by both $e^v = q$ and $e^u = f$. When computing the second derivative for v, we need to recall that the first derivative contains the Jacobian, so:

$$\begin{aligned}
\frac{\partial^2}{\partial v} \lambda_t(\boldsymbol{\gamma}) &= \frac{\partial}{\partial v} \left(\frac{\partial}{\partial v} \lambda_t(\boldsymbol{\gamma}) \right) = \frac{\partial}{\partial e^v} \left(\frac{\partial}{\partial e^v} \lambda_t(\boldsymbol{\gamma}) \cdot e^v \right) \cdot e^v = \\
&= \left(\frac{\partial}{\partial e^v} \frac{\partial}{\partial e^v} \lambda_t(\boldsymbol{\gamma}) \cdot e^v + \frac{\partial}{\partial e^v} \lambda_t(\boldsymbol{\gamma}) \cdot \frac{\partial}{\partial e^v} e^v \right) \cdot e^v = \\
&= \left(\frac{\partial}{\partial q} \frac{\partial}{\partial q} \lambda_t(\boldsymbol{\gamma}) \cdot q + \frac{\partial}{\partial q} \lambda_t(\boldsymbol{\gamma}) \cdot \frac{\partial}{\partial q} q \right) \cdot q = \\
&= \frac{\partial^2}{\partial q^2} \lambda_t(\boldsymbol{\gamma}) \cdot q^2 + \frac{\partial}{\partial q} \lambda_t(\boldsymbol{\gamma}) \cdot q.
\end{aligned} \tag{C.6}$$

C.3 Model on *daily deceased*

The same analysis carried out for the *daily positives* has been conducted on the *daily deceased*. Comparisons in terms of goodness of fit measures are reported for both models in Table C.1. The best model in terms of all the goodness of fit scores (AIC, AICc and BIC) is the model with baseline. The resulting estimated parameters $\hat{\boldsymbol{\theta}}$ and the respective intervals are shown in Table C.2, where the baseline α is estimated to be $\hat{\alpha} = 3.3$ (sensibly larger than 0). Hence, in the considered time horizon, we expect the endemic fatality rate to be of ≈ 3 deaths per day. The final outbreak size r is estimated to be $\approx 35,810$, an amount that would have been reached in $\approx 10,851$ days at the endemic fatality rate. As for the daily positives, the parameters h, p, s and ν do not have an easily quantifiable and absolute interpretation, but are useful for comparisons.

Table C.1. *Log-likelihood, AIC, BIC and AICc for the model without baseline and the model with baseline on daily deceased.*

Index	Model without baseline	Model with baseline
log-likelihood	-735.8	-732.5
AIC	1461.6	1452.9
AICc	1471.2	1464.4
BIC	1446.7	1435.1

Table C.2. *Parameters' points estimates and* 95% *confidence intervals for the model with baseline on daily deceased.*

Parameter	Point estimate	95% Interval
α	3.3	$(1.97, 5.54)$
r	35.73×10^3	$(35.34 \times 10^3, 36.11 \times 10^3)$
h	0.0247	$(0.0244, 0.0249)$
p	-50.50	$(-52.07, -48.93)$
s	171.58	$(130.33, 201.83)$
ν	12.36	$(11.73, 13.02)$

We can then obtain point predictions $\{\hat{y}_t\}_{t=1}^T$ and prediction intervals $\left\{(\widehat{y}_t^l; \widehat{y}_t^u)\right\}_{t=1}^T$ through the parametric bootstrap procedure described in the Main Text. Figure C.1 shows the fit on the whole available time series of counts: the former on the daily series, the latter on the cumulative one. Also in the case of the deceased the estimated curve does catch the observed general behavior. The same metrics are used to evaluate the fitting performances, which correspond to an $R^2 = 0.90$ and a coverage $\overline{\text{Cov}}_{95\%} = 0.95$. Also here, we performed a diagnostic check on both the Pearson and the Deviance residuals. The plots in Figure C.2 show the Deviance residuals behavior: histogram (a), including the p-value from the Shapiro test; Normal qq-plot (b); auto-correlation plot (c); plot of the residuals vs. fitted values (d).

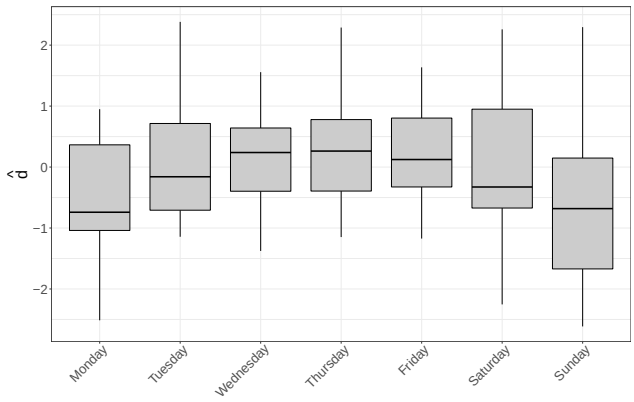


Figure C.3. Deviance residuals distribution aggregated by day of the week for *daily deceased*.

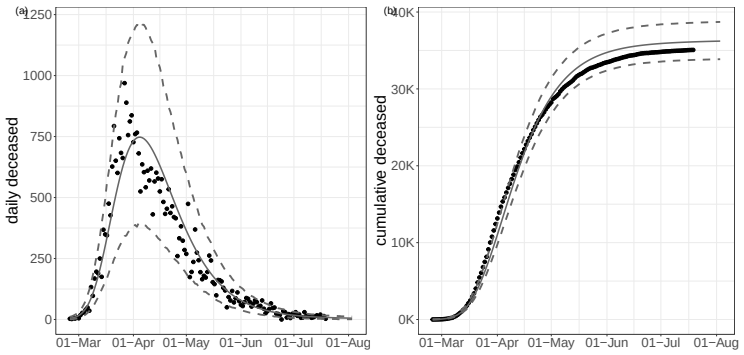


Figure C.1. Observed (black dots) and fitted values (grey solid lines) with 95% confidence intervals (grey dashed lines) for model with baseline on *daily deceased*.

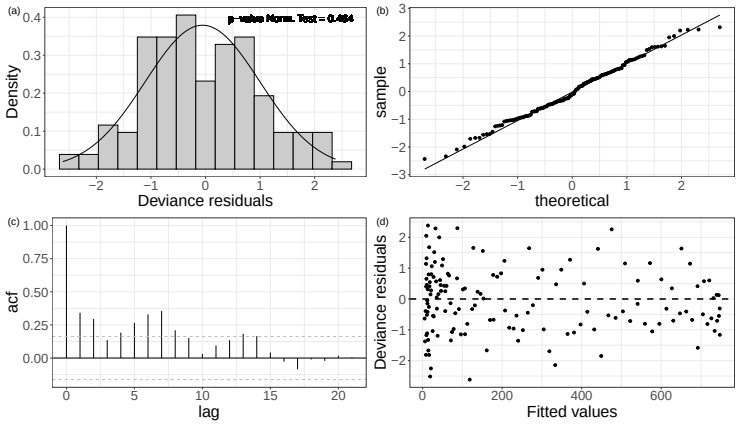


Figure C.2. Deviance residuals for the model with baseline on *daily deceased*.

Weekly seasonality

As in the case of *daily positives*, the diagnostic check on the residuals for the *daily deceased* highlights a slight week seasonality pattern for the auto-correlations. In addition, also the residual Normality hypothesis is rejected. Potentially, the inclusion of a week-day effect may solve both problems. In order to decide what set of week-days to group together, we visualize the residuals' distribution aggregated by week (see Figure C.3). The pattern is not as evident as in the case of *daily positives*, but we can still detect some undesirable overestimation on Mondays and Sundays.

Table C.3. *Log-likelihood, AIC, BIC and AICc for the models with baseline including additive or multiplicative week-day effect on daily deceased.*

Index	Additive effect	Multiplicative effect
log-likelihood	-725.77	-725.30
AIC	1437.78	1436.61
AICc	1450.97	1449.81
BIC	1416.90	1415.73

Table C.4. *Intercept β_0 and week-day effect β_{wd} point estimates and 95% confidence intervals for the additive model with baseline on daily deceased.*

Parameter	Point estimate	95% Interval
β_0	1.85	$(1.73, 1.98)$
β_{wd}	-510.36	$(-573.84, -446.88)$
r	35.81×10^3	$(33.58 \times 10^3, 38.19 \times 10^3)$
h	0.0251	$(0.0246, 0.0255)$
p	-58.88	$(-63.52, -54.22)$
s	297.24	$(199.12, 390.36)$
ν	13.58	$(10.01, 18.41)$

Therefore, on the line of the previous application, we decide to include a dichotomous week-day fixed effect on the pair Sunday-Monday. As before, this effect may be included either in an additive or a multiplicative fashion and, again, we may pick the version that achieves the best AIC, AICc and BIC scores. However, as shown in Table C.3, differences in these scores are almost negligible and choice based on such a small improvement would not be robust. Therefore, we checked the Pearson residuals for both alternatives and we selected the additive model because of the improved residuals behavior (Normality is accepted, autocorrelation at lag 7 is reduced). The resulting fit is shown in Figure C.4 where: on the left, we can observe the fitted curve and the 95% confidence intervals; on the right, we can observe the cumulative fit. Estimated parameters are shown in Table C.4, where the Sunday-Monday effect is estimated to have a strong reducing effect on the daily baseline rate of ≈ -510 on the log-scale, i.e. $\exp\{-510\} \approx 0$, which shrinks to 0 the *baseline* on Mondays and Sundays. The estimates of the outbreak size $\hat{r}$ and of the infection rate $\hat{h}$ of the two models are in agreement, while the point estimates of the asymmetry parameter $\hat{s}$ are different but both large and mutually included in the corresponding 95% intervals. This is reasonable since we would not expect the outbreak size, rate and symmetry to vary wildly after accounting for week-day heterogeneity. On the other hand, the new estimate $\hat{p}$ of p detects a longer lag-phase and hence a slightly slower approach to the descending phase. Finally, the estimate

of the dispersion parameter $\hat{\nu}$ is slightly larger than in the model without covariates, denoting less over-dispersion with respect to the equi-variance hypothesis. This is completely reasonable since the week-day effect is able to explain some of the previously unaccounted heterogeneity. The inclusion of the Sunday-Monday effect allows for an increase of the R^2 to 0.91, whilst keeping the coverage $\overline{\text{Cov}}_{95\%}$ steady at 0.95. The diagnostic check shown in Figure C.5 shows how Residual Normality is now accepted and the previously evident correlation pattern is slightly reduced.

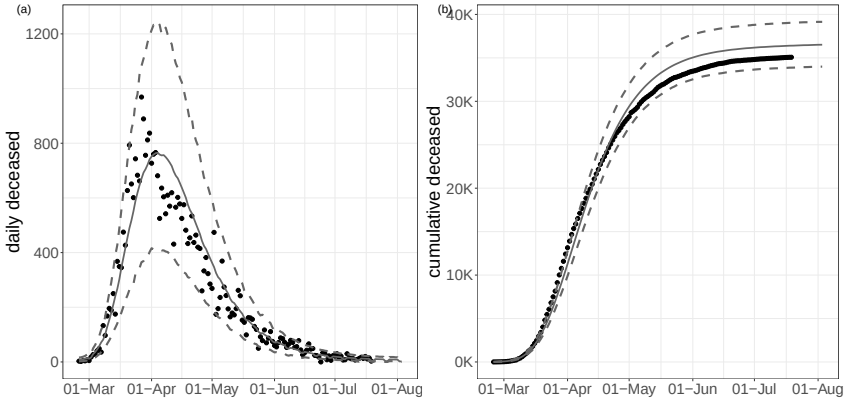


Figure C.4. Observed (black dots) and fitted values (grey solid lines) with 95% confidence intervals (grey dashed lines) for model with baseline and week-day additive effect, estimated on the *daily deceased*.

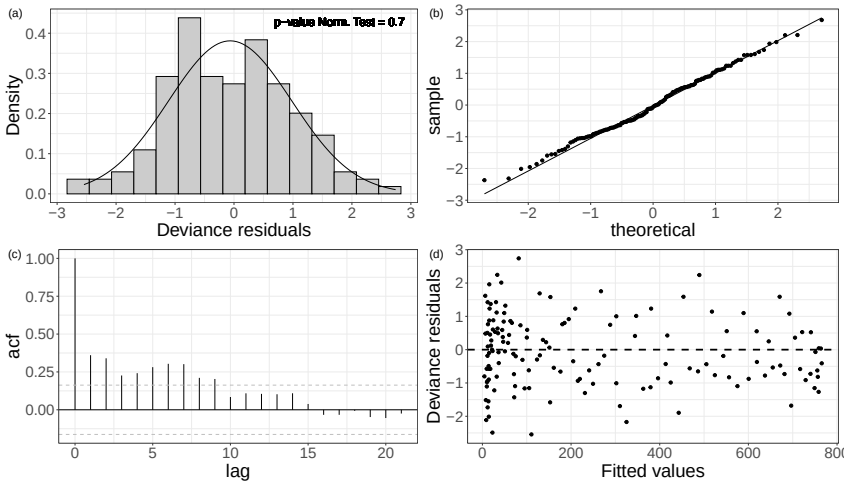


Figure C.5. Deviance residuals for the model with baseline and week-day additive effect on *daily deceased*.

C.3.1 Prediction of future cases and of the peak date

Validation performances on the *daily deceased* are analogous to the ones on *daily positives*. As in the main text, also here we highlight how the peak is accurately predicted with a shorter delay and generally smaller uncertainty for the *daily deceased* than for the *daily positives*. This is probably related to the more regular behavior of the series, due to a likely more homogeneous collection process of the records.

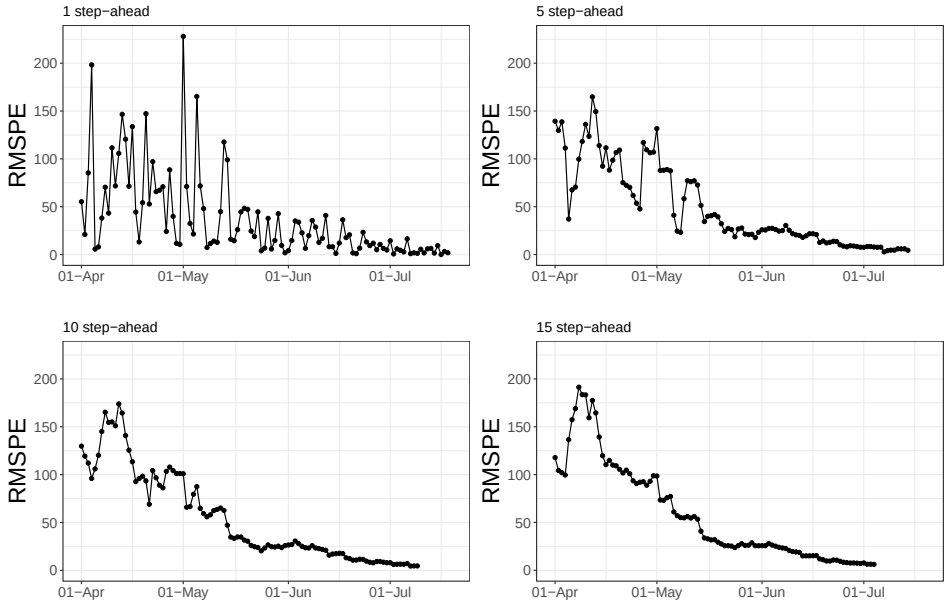


Figure C.6. RMSPE for *daily deceased* at different steps-ahead.

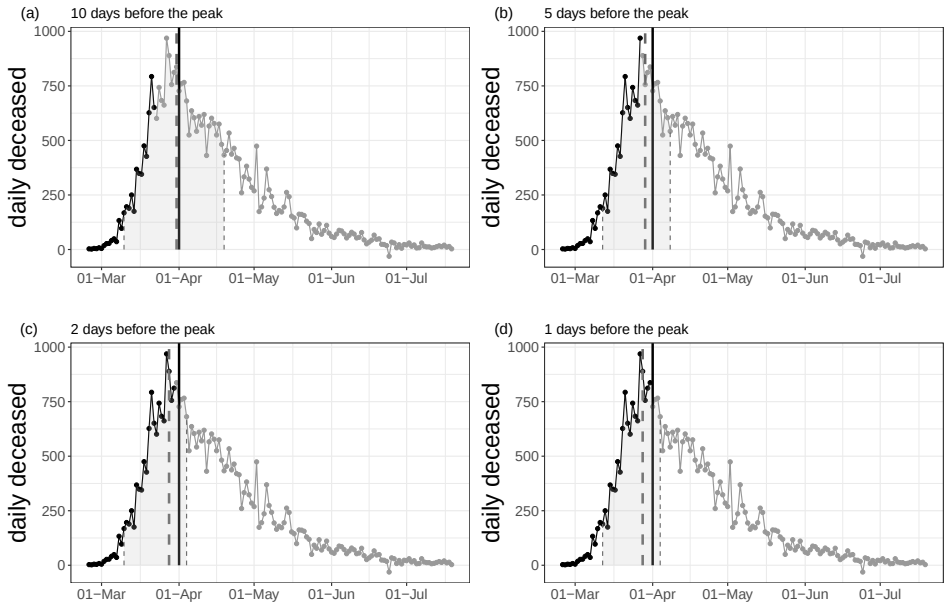


Figure C.7. Estimation of the date of the peak for *daily deceased* at different steps-before.

C.4 Regional *daily positives*

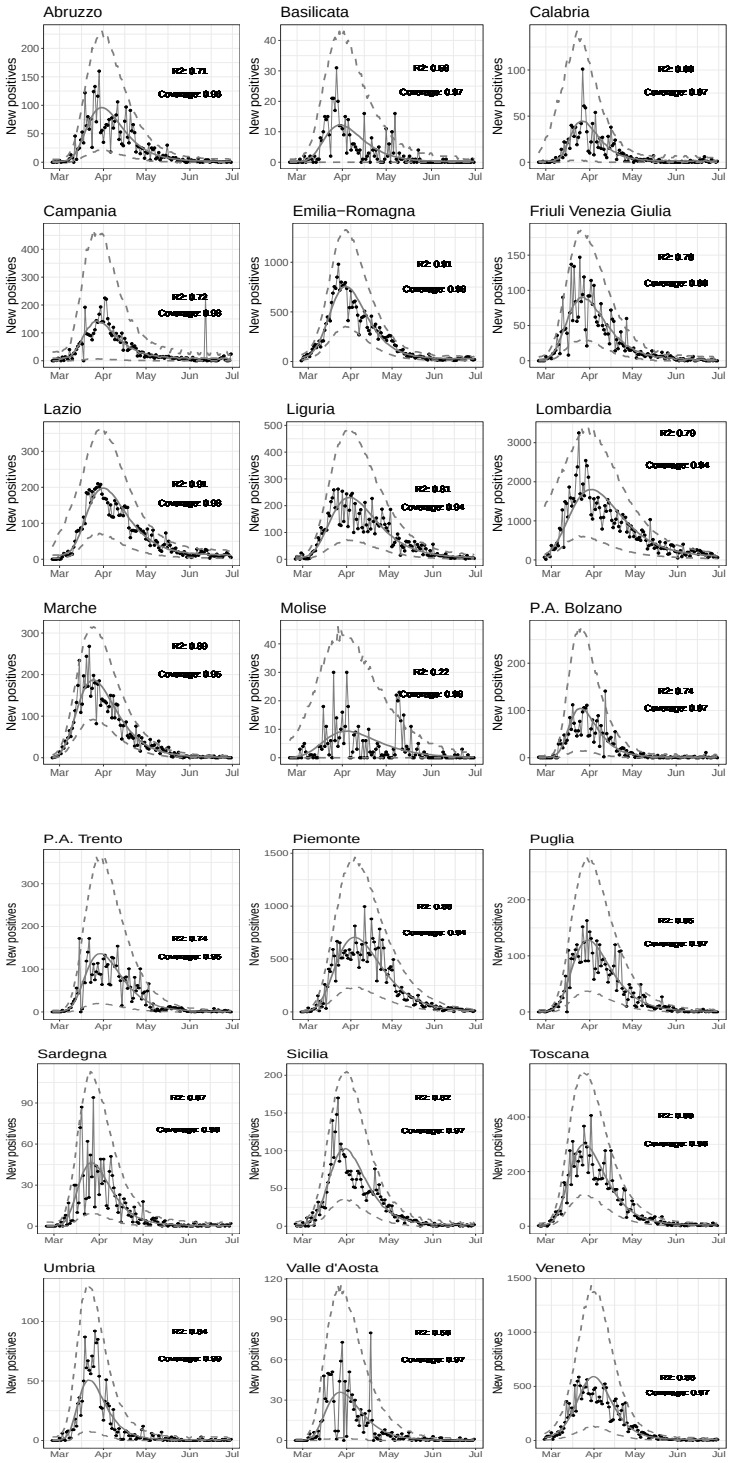


Figure C.8. Observed (black dots) and fitted values (grey solid lines) with 95% confidence intervals (grey dashed lines) for model with baseline and week-day additive effect, estimated on the *daily positives* at the regional level.

C.5 Regional *daily deceased*

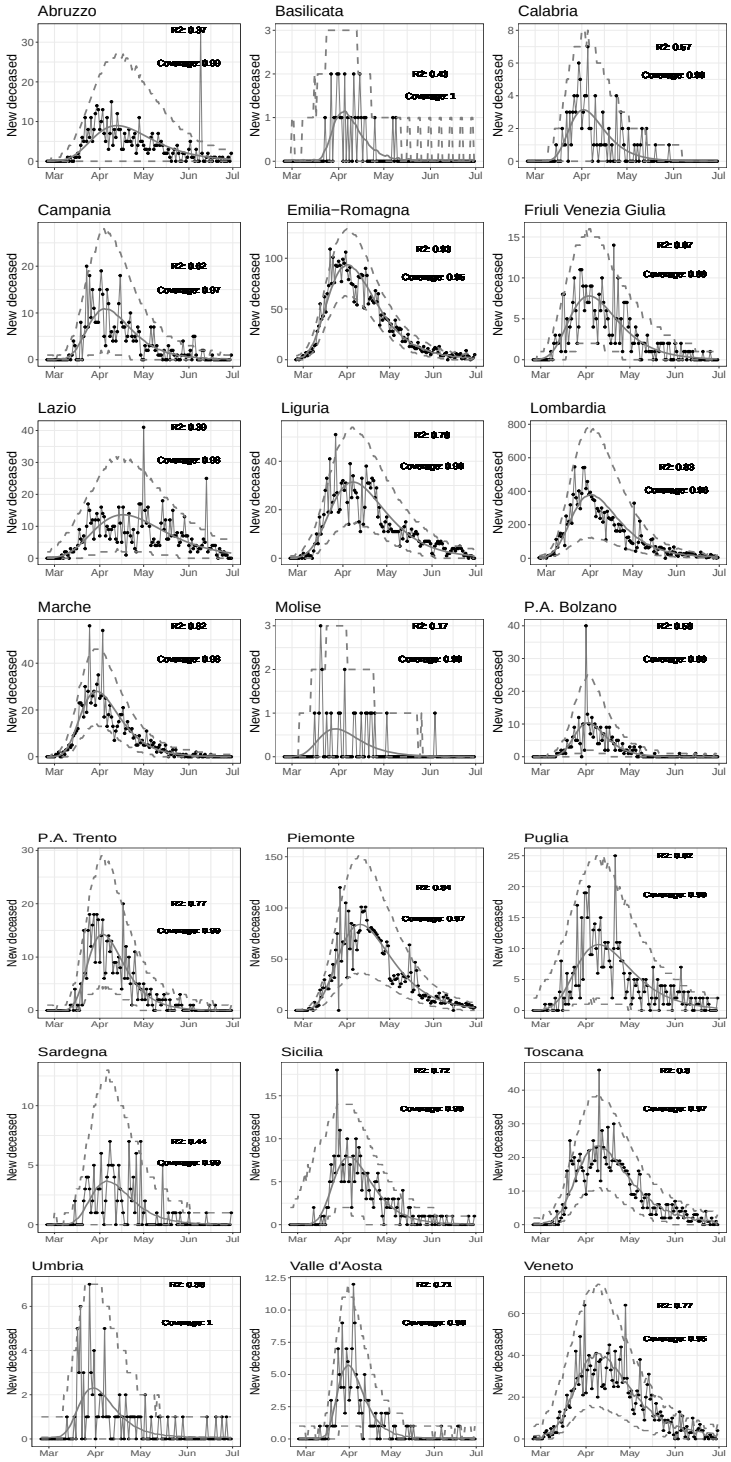


Figure C.9. Observed (black dots) and fitted values (grey solid lines) with 95% confidence intervals (grey dashed lines) for model with baseline and week-day additive effect, estimated on the *daily deceased* at the regional level.

Appendix D

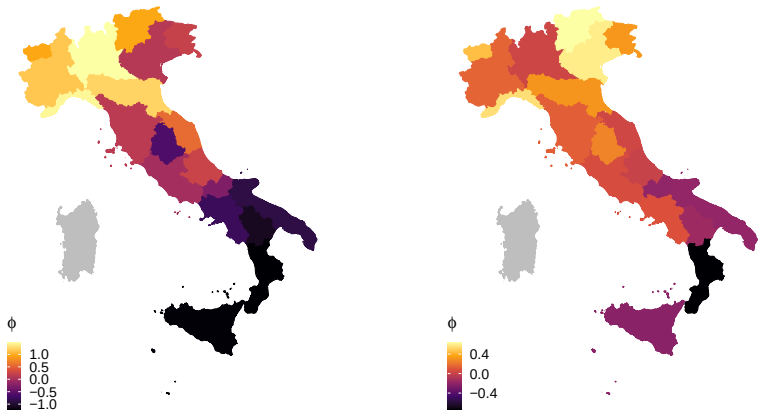
Spatio-temporal modelling of COVID-19 incident cases using Richards' curve

D.1 Exact-sparse CAR algorithm

```
functions {
    // Sparse computation of the log-posterior contribution of the single
        spatial vector phi_t (phi) given the previous value phi_{t-1} (
        phi_old)
    real sparse_carar_lpdf(vector phi, vector phiOld, real rho, real tau,
        real alpha,
    int[,] W_sparse, vector W_weight, vector D_sparse, vector lambda, int n,
        int W_n) {
        row_vector[n] phit_D;
        row_vector[n] phit_W;
        vector[n] ldet_terms;
        vector[n] phiNew = phi-rho*phiOld;
        phit_D = (phiNew .* D_sparse);
        phit_W = rep_row_vector(0, n);
        for (i in 1:W_n) {
            phit_W[W_sparse[i, 1]] = phit_W[W_sparse[i, 1]] + W_weight[i]*
                phiNew[W_sparse[i, 2]];
            phit_W[W_sparse[i, 2]] = phit_W[W_sparse[i, 2]] + W_weight[i]*
                phiNew[W_sparse[i, 1]];
        }
      for (i in 1:n) ldet_terms[i] = log1m(alpha * lambda[i]);
      return 0.5*(n*log(tau) + sum(ldet_terms) - tau*(phit_D*phiNew - alpha*(
          phit_W*phiNew)));
  }
}
data {
  int<lower=0> nTimes;     // Number of times
  int<lower=0> nReg;       // Number of regions
  int W_n;                 // Number of adjacent region pairs
  int W_sparse[W_n, 2];    // adjacency pairs
  vector[W_n] W_weight;    // Connection weights
  vector[Nreg] D_sparse;   // diagonal of D
  vector[Nreg] lambda;     // eigenvalues of invsqrtD * W * invsqrtD
}
parameters {
    vector[Nreg] phi[Ntimes];
    real<lower=0, upper=1> alpha;
    real<lower=-1, upper=1> rho;
}
model {
    alpha ~ beta(0.5, 0.5);
    phi[1] ~ sparse_carar(zeros, rho, tau, alpha,
                          W_sparse, W_weight, D_sparse, lambda, Nreg, W_n);
    for (i in 2:Ntimes){
            phi[i] ~ sparse_carar(phi[i-1], rho, tau, alpha,
                          W_sparse, W_weight, D_sparse, lambda, Nreg, W_n);
        }
}
```

Algorithm D.1. Core of the `Stan` code for the sparse implementation of the spatio-temporal CAR-AR.

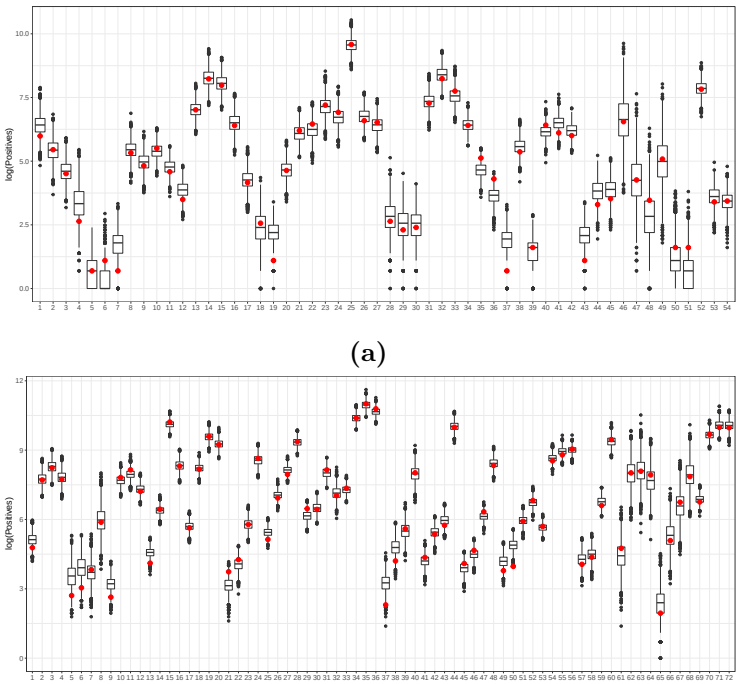
D.2 Estimated average spatial random effects



(a) (b)

Figure D.1. Average estimated spatial random effect $\bar{\phi}_g$ by M_2 for the first (a) and the second (b) wave.

D.3 Posterior distribution of out-of-sample predictions



(a)

(b)

Figure D.2. Observed points (red) and simulated predictions (boxplots) for the first (a) and the second wave (b) test sets.

D.4 Forecasting performances

As pointed out by an anonymous referee, during the COVID-19 pandemic the objective of authorities and researchers was to predict the future incidence. Hence, it would be interesting to see which of the proposed models performs “better" when that is the objective in mind. We would like to remark that this model has been originally conceived as an explanatory tool, not as a forecasting tool. That is why the prediction performances have been assessed on values occurring within the waves in the main text. This conceptual limitation depends on a number of factors, listed here below.

1. The expression of the mean depends on the number of weekly swabs, which is not known in advance at future times. In order to provide valid and accurate forecasts (together with the corresponding uncertainty) the number of weekly swabs and positives shall be modeled jointly. This is something worth of exploration in the future.

2. The model includes a complex and highly parametrized spatio-temporal structure. Hence, it likely need a relative large number of time-points to estimate its components without introducing bias.

3. The main focus on the paper is on identifying the most informative spatial structure in the observed data. In principle, spatial dependence is much more helpful when predicting the counts of some regions at some time t, when other regions' counts at the same time are available. On the contrary, the impact of spatial dependence can rapidly fade when considering future outcomes, where temporal dependence shall dominate the process behavior (unless suitable space-time neighboring structure are considered). In any case, when there is no information from other regions at the same time, the spatial dependence shall propagate through time and its effect is inevitably diminished week after week.

All things considered, we still believe the proposed model may provide reasonable predictions up to 15 days (2 weeks) ahead, as for all our other works on this topic (please see Farcomeni et al. (2021) and Alaimo Di Loro et al. (2021a)). Hence, in this section, we show predictions at one- and two-weeks ahead, comparing the different specifications of $\boldsymbol{W}$. We only did this for the *common* Richards': the best model, as discussed in Subsection 5.2.3. Issue 1 is overcome by assuming constant tracing effort: the weekly number of swabs used to predict the positive cases for the weeks ahead is assumed to be the same as the previous (in sample) week. This is a very strong (and possibly wrong) working assumption, which is particularly ill-suited for the first wave when the tracing was improving its capabilities week after week. Therefore, forecasts shall be interpreted only in relative terms as a comparison between the three dependence structures and not in absolute terms. The prediction experiment was developed as follows.

First wave. We estimate four examples building training sets up to April 12-19-26 and May 03, 2020. We predict the following first and second week count in each case.

Second wave. We estimate four examples building training sets up to November 22-29 and December 06-13, 2020. We predict the following first and second week count in each case.

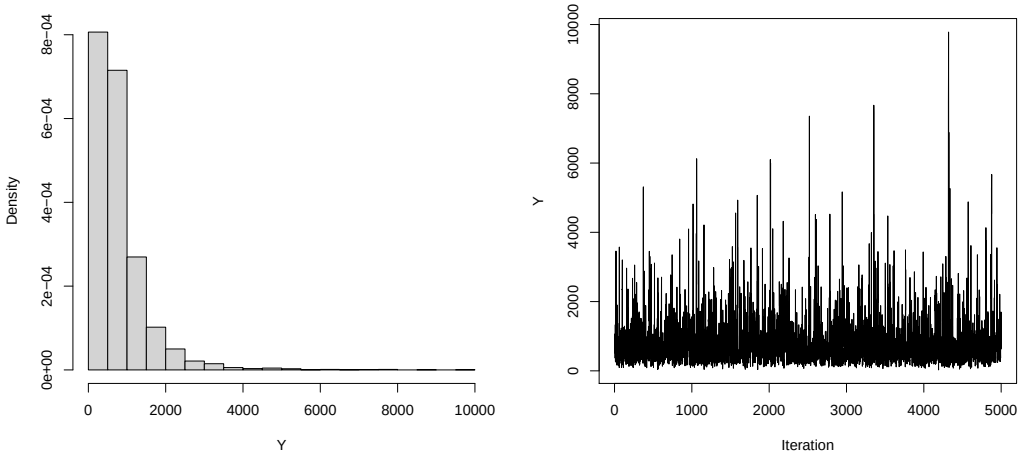


Figure D.3. Density and traceplot of a predicted value.

		Week(s) ahead	
Wave	**Model**	**1**	**2**
I	M_0	244 (47)	259 (54)
	M_1	242 (46)	256 (53)
	M_2	267 (49)	282 (45)
II	M_0	1670 (728)	2544 (1522)
	M_1	1388 (496)	2189 (1005)
	M_2	1749 (825)	2578 (1543)

Table D.1. Average (Median) RMSPE for each model specification for the first and the second wave at different steps ahead.

Figure D.3 shows an example of posterior predictive distribution. Notice how this is strongly skewed, hence we also considered posterior medians as point estimates for the forecasts.

Performances are evaluated in terms of root mean square predictive error (RMSPE), whose distribution by wave, window, and dependence structure can be observed in Figure D.4. The errors are ultimately averaged over the various regions and fitting windows, keeping 1-week and 2-weeks ahead as separate objects, and the resulting estimates are reported in Table D.1.

Eventually, what is found in terms of fit quality in the original application is also confirmed for these forecasts. Along the first wave, there is no clear dominance of one dependence structure over the others: apparently, the most of the information is carried by the temporal process. Instead, along the second wave, the transport flows dependence structure provides clearly better forecasts than all its competitors. Note that the two-weeks-ahead prediction error is uniformly larger than the one-week-ahead as expected, for both waves and for each specification of $\boldsymbol{W}$.

Some examples of the point forecasts, together with the corresponding uncertainty, are shown in Figure D.5.

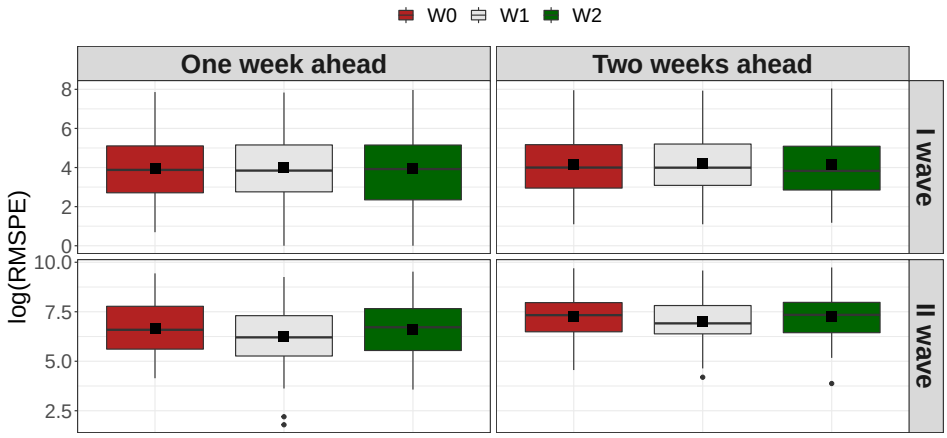


Figure D.4. Prediction error (on the log scale) at different steps ahead, for each specification of $\boldsymbol{W}$ and for each wave.

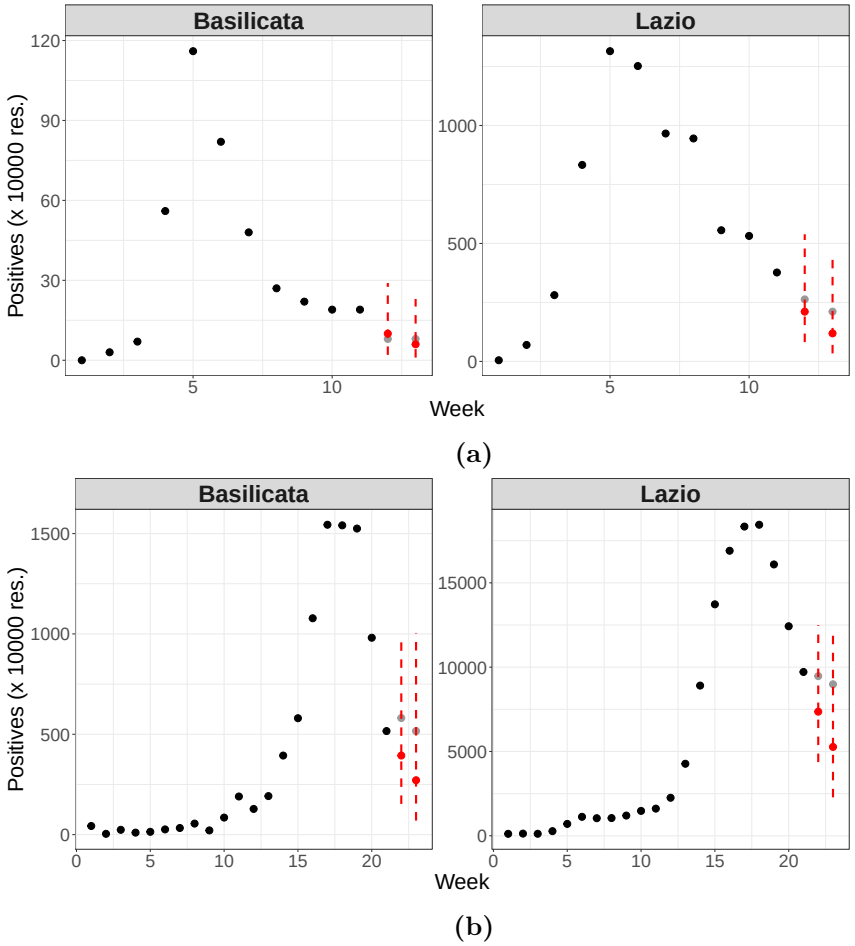


Figure D.5. Observed time series: black dots are in-sample, grey dots are out-of-sample. Predicted values (red dots) with 95% prediction intervals (red dashed) for 2 randomly chosen regions in the first (top panels) and the second wave (bottom panels).

Bibliography

Boxall, R.A (1986). A critical review of the methodology for assessing farm-level grain losses after harvest. Accessed: 2020-01-18.

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.

Abramson, B., Brown, J., Edwards, W., Murphy, A., and Winkler, R. L. (1996). Hailfinder: A bayesian system for forecasting severe weather. *International Journal of Forecasting*, 12(1):57–71.

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., Divino, F., Farcomeni, A., Jona Lasinio, G., Lovison, G., Maruotti, A., and Mingione, M. (2021a). Nowcasting covid-19 incidence indicators during the italian first outbreak. *Statistics in Medicine*, 40(16):3843–3864.

Alaimo Di Loro, P., Mingione, M., Lipsitt, J., Batteate, C. M., Jerrett, M., and Banerjee, S. (2021b). Bayesian hierarchical modeling and analysis for physical activity trajectories using actigraph data. *arXiv preprint arXiv:2101.01624*.

Alder, B. J. and Wainwright, T. E. (1959). Studies in molecular dynamics. i. general method. *The Journal of Chemical Physics*, 31(2):459–466.

Allaire, J. (2012). RStudio: integrated development environment for R. *Boston, MA*, 770:394.

Allen, M. (2017). *The SAGE encyclopedia of communication research methods*. Sage Publications.

Alvarez, M. A. and Lawrence, N. D. (2011). Computationally efficient convolved multiple output gaussian processes. *The Journal of Machine Learning Research*, 12:1459–1500.

Baek, J., Farias, V. F., Georgescu, A., Levi, R., Peng, T., Sinha, D., Wilde, J., and Zheng, A. (2020). The limits to learning an SIR process: granular forecasting for COVID-19. *arXiv.2006.06373*.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

Banks, R. B. (1993). *Growth and diffusion phenomena: Mathematical frameworks and applications*, volume 14. Springer Science & Business Media.

Barbarossa, M. V., Fuhrmann, J., Meinke, J. H., Krieg, S., Varma, H. V., Castelletti, N., and Lippert, T. (2020). Modeling the spread of covid-19 in germany: Early assessment and possible scenarios. *Plos one*, 15(9):e0238559.

Barbieri, M. M., Berger, J. O., et al. (2004). Optimal predictive model selection. *The annals of statistics*, 32(3):870–897.

Bartolucci, F. and Farcomeni, A. (2021). A spatio-temporal model based on discrete latent variables for the analysis of COVID-19 incidence. *Spatial Statistics*, page 100504.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365.

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.

Betancourt, M. (2016a). Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo. *arXiv preprint arXiv:1604.00695*.

Betancourt, M. (2016b). Identifying the optimal integration time in hamiltonian monte carlo. *arXiv preprint arXiv:1601.00225*.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Betancourt, M. and Girolami, M. (2015). Hamiltonian monte carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79(30):2–4.

Bhadra, A., Datta, J., Polson, N. G., Willard, B., et al. (2019a). Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. T. (2019b). The horseshoe-like regularization for feature subset selection. *Sankhya B*.

Birch, C. P. (1999). A new generalized logistic sigmoid growth equation compared with the richards growth equation. *Annals of botany*, 83(6):713–723.

Bock, R. D. (2014). *Multilevel analysis of educational data*. Elsevier.

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.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brooks, S. (1998). Markov chain monte carlo method and its application. *Journal of the royal statistical society: series D (the Statistician)*, 47(1):69–100.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC 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.

Cabras, S. (2020). A Bayesian deep learning model for estimating COVID-19 evolution in Spain. *arXiv.2005.10335*.

Cao, L., Shi, P.-J., Li, L., and Chen, G. (2019). A new flexible sigmoidal growth model. *Symmetry*, 11(2):204.

Car, Z., Baressi Šegota, S., Anđelić, N., Lorencin, I., and Mrzljak, V. (2020). Modeling the spread of covid-19 infection using a multilayer perceptron. *Computational and mathematical methods in medicine*, 2020.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: a probabilistic programming language. *Grantee Submission*, 76(1):1–32.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80. Accessed: 2020-01-18.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Causton, D. (1969). A computer program for fitting the richards function. *Biometrics*, pages 401–409.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., et al. (2006). Deviance information criteria for missing data models. *Bayesian analysis*, 1(4):651–673.

Cf, O. (2015). Transforming our world: the 2030 agenda for sustainable development.

Chen, Y. C., Lu, P. E., and Chang, C. S. (2020). A time-dependent SIR model for COVID-19. *arXiv:2003.00122*.

Chowell, G., Hincapie-Palacio, D., Ospina, J., Pell, B., Tariq, A., Dahal, S., Moghadas, S., Smirnova, A., Simonsen, L., and Viboud, C. (2016). Using phenomenological models to characterize transmissibility and forecast patterns and final burden of zika epidemics. *PLoS currents*, 8.

Cihan, P. (2018). A comparison of five methods for missing value imputation in data sets. *International Scientific and Vocational Studies Journal*, 2(2):80–85. Accessed: 2020-01-18.

Clark, J. S. and Gelfand, A. E. (2006). *Hierarchical modelling for the environmental sciences: statistical methods and applications*. OUP Oxford.

Congdon, P. D. (2019). *Bayesian hierarchical models: with applications using R*. CRC Press.

Corpas-Burgos, F. and Martinez-Beneito, M. A. (2020). On the use of adaptive spatial weight matrices from disease mapping multivariate analyses. *Stochastic Environmental Research and Risk Assessment*, 34(3):531–544.

Costa-Santos, C., Neves, A. L., Correia, R., Santos, P., Monteiro-Soares, M., Freitas, A., Ribeiro-Vaz, I., Henriques, T. S., Rodrigues, P. P., Costa-Pereira, A., et al. (2021). Covid-19 surveillance data quality issues: a national consecutive case series. *BMJ open*, 11(12):e047623.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

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.

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.

Datta, J. and Ghosh, J. K. (2015). In search of optimal objective priors for model selection and estimation. *Current Trends in Bayesian Methodology with Applications*, 225.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, 26(2):403–413.

Dehning, J., Zierenberg, J., Spitzner, F. P., Wibral, M., Neto, J. P., Wilczek, M., and Priesemann, V. (2020). Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science*.

Della Rossa, F., Salzano, D., Di Meglio, A., De Lellis, F., Coraggio, M., Calabrese, C., Guarino, A., Cardona-Rivera, R., De Lellis, P., Liuzza, D., et al. (2020). A network model of Italy shows that intermittent regional strategies can alleviate the COVID-19 epidemic. *Nature communications*, 11(1):1–9.

Di Narzo, A. and Cocchi, D. (2010). A bayesian hierarchical approach to ensemble weather forecasting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(3):405–422.

Diaconis, P., Ylvisaker, D., et al. (1979). Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281.

Dicker, R. C., Coronado, F., Koo, D., and Parrish, R. G. (2006). *Principles of epidemiology in public health practice; an introduction to applied epidemiology and biostatistics*. Centers for Disease Control and Prevention (CDC).

Diekmann, O., Heesterbeek, H., and Britton, T. (2013). *Mathematical Tools for Understanding Infectious Disease Dynamics*. Princeton University Press, Princeton.

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.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, 195(2):216–222.

Duncan, E. W., White, N. M., and Mengersen, K. (2017). Spatial smoothing in bayesian models: a comparison of weights matrix specifications and their impact on inference. *International journal of health geographics*, 16(1):1–16.

Eberhardt, J. J. (2015). Bayesian spam detection. *Scholarly Horizons: University of Minnesota, Morris Undergraduate Journal*, 2(1):2.

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.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632.

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *The Annals of Applied Statistics*, 6(4):1971.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102.

Ejigu, B. A. and Wencheko, E. (2020). Introducing covariate dependent weighting matrices in fitting autoregressive models and measuring spatio-environmental autocorrelation. *Spatial Statistics*, 38:100454.

Evans, M. (1991). Chaining via annealing. *The Annals of Statistics*, pages 382–393.

Fabi, C. and English, A. (2019). Measuring food losses at the national and subnational levels: Fao's methodology for monitoring sustainable development goals. In *The Economics of Food Loss in the Produce Industry*, pages 101–115. Routledge.

Fabi, C., English, A., Mingione, M., and Jona Lasinio, G. (2018). Sdg 12.3. 1: Global food loss index. *FAO, Rome*.

FAO (1980). Assessment and collection of data on post-harvest food-grain losses. `http://www.fao.org/3/ca6157en/CA6157EN.pdf`. Accessed: 2020-01-18.

FAO (2014). SAVE FOOD: Global Initiative on Food Loss and Waste Reduction. Definitional framework of food loss. `http://www.fao.org/3/a-at144e.pdf`. Accessed: 2020-01-18.

FAO (2019). The state of food and agriculture 2019. moving forward on food losses and waste reduction. `http://www.fao.org/3/ca6030en/ca6030en.pdf`. Accessed: 2020-01-18.

FAOSTAT (2016). The United Nations Food and Agriculture Organization Database. Database.

Farcomeni, A., Maruotti, A., Divino, F., Jona-Lasinio, G., and Lovison, G. (2021). An ensemble approach to short-term forecast of covid-19 intensive care occupancy in italian regions. *Biometrical Journal*, 63(3):503–513.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Finley, A., Datta, A., and Banerjee, S. (2017). Spnngp: Spatial regression models for large datasets using nearest neighbor gaussian processes. *R package version 0.1*, 1.

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. 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*, 28(2):401–414.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Coupland, H., Mellan, T. A., Zhu, H., Berah, T., Eaton, J. W., Guzman, P. N. P., Schmit, N., and Callizo, L. (2020). Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in European countries: technical description update. *arXiv:2004.11342*.

Franzin, A., Sambo, F., and Di Camillo, B. (2016). bnstruct: an r package for bayesian network structure learning in the presence of missing data. *Bioinformatics*, 33(8):1250–1252.

Freedman, D. A. (2006). On the so-called "huber sandwich estimator" and "robust standard errors". *The American Statistician*, 60(4):299–302.

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.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402.

Gatto, M., Bertuzzo, E., Mari, L., Miccoli, S., Carraro, L., Casagrandi, R., and Rinaldo, A. (2020). Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences*, 117(19):10484–10491.

Gelfand, A. E. (2012). Hierarchical modeling for spatial data problems. *Spatial statistics*, 1:30–39.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of Spatial Statistics*. Chapman and Hall/CRC Press, Boca Raton, FL.

Gelfand, A. E., Fuentes, M., Hoeting, J. A., and Smith, R. L. (2019). *Handbook of environmental and ecological statistics*. CRC Press.

Gelfand, A. E. and Sahu, S. K. (1994). On markov chain monte carlo acceleration. *Journal of Computational and Graphical Statistics*, 3(3):261–276.

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.

Gelman, A. (2006). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 48(3):432–435.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gelman, A., Gilks, W. R., and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016.

Gelman, A., Roberts, G. O., Gilks, W. R., et al. (1996). Efficient metropolis jumping rules. *Bayesian statistics*, 5(599-608):42.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, pages 721–741.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gibbs, J. W. (1902). Elementary principles in statistical mechanics. *Nature*, 66(1708):291–292.

Gilks, W. R., Roberts, G. O., and Sahu, S. K. (1998). Adaptive markov chain monte carlo through regeneration. *Journal of the American statistical association*, 93(443):1045–1054.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348.

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., and Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*, page to appear.

Girardi, P., Greco, L., Mameli, V., Musio, M., Racugno, W., Ruli, E., and Ventura, L. (2020a). Robust inference for non-linear regression models from the tsallis score: Application to coronavirus disease 2019 contagion in italy. *Stat*, 9(1):e309.

Girardi, P., Greco, L., Mameli, V., Musio, M., Racugno, W., Ruli, E., and Ventura, L. (2020b). Robust inference from robust Tsallis score: application to COVID-19 contagion in Italy. *STAT*.

Girardi, P., Greco, L., and Ventura, L. (2021). Misspecified modeling of subsequent waves during covid-19 outbreak: A change-point growth model. *Biometrical Journal*, 63:In press.

Goldstein, H. (2011). *Multilevel statistical models*, volume 922. John Wiley & Sons.

Gompertz, B. (1825). Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. frs &c. *Philosophical transactions of the Royal Society of London*, 115:513–583.

Good, I. J. (1959). Kinds of probability. *Science*, 129(3347):443–447.

Good, I. J. (1980). Some history of the hierarchical bayesian methodology. *Trabajos de estadística y de investigación operativa*, 31(1):489–519.

Good, I. J. and England, C. B. (1965). The estimation of probabilities.

Goodman, T. and Hardin, D. (2006). Refinable multivariate spline functions. In *Studies in Computational Mathematics*, volume 12, pages 55–83. Elsevier.

Grasselli, G., Pesenti, A., and Cecconi, M. (2020). Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: Early experience and forecast during an emergency response. *Journal of the American Medical Association*, 323:1545–1546.

Green, B. F. and Tukey, J. W. (1960). Complex analyses of variance: general problems. *Psychometrika*, 25(2):127–152.

Groll, A., Hambuckers, J., Kneib, T., and Umlauf, N. (2019). Lasso-type penalization in the framework of generalized additive models for location, scale and shape. *Computational Statistics & Data Analysis*, 140:59–73.

Grossman, M. and Bohren, B. (1985). Logistic growth curve of chickens: heritability of parameters. *Journal of Heredity*, 76(6):459–462.

Guennebaud, G., Jacob, B., et al. (2010). Eigen. *URl: http://eigen. tuxfamily. org*, 3.

Gustavsson, J., Cederberg, C., Sonesson, U., van Otterdijk, R., and Meybeck, A. (2011). Global food losses and food waste. Technical report, Food and Agricultural Organization - FAO.

Haario, H., Saksman, E., and Tamminen, J. (1999). Adaptive proposal distribution for random walk metropolis algorithm. *Computational Statistics*, 14(3):375–395.

Haario, H., Saksman, E., Tamminen, J., et al. (2001). An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242.

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 Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):221–238.

Hänggi, J. M., Phillips, L. R., and Rowlands, A. V. (2013). Validation of the gt3x actigraph in children and comparison with the gt1m actigraph. *Journal of science and Medicine in Sport*, 16(1):40–44.

Hardin, J. W. (2003). The sandwich estimate of variance. In *Maximum likelihood estimation of misspecified models: Twenty years later*, pages 45–73. Emerald Group Publishing Limited.

Harris, K. L. and Lindblad, C. J. (1978). Postharvest Grain Loss Assessment Methods.

Hastie, T., Tibshirani, R., et al. (2000). Bayesian backfitting (with comments and a rejoinder by the authors. *Statistical Science*, 15(3):196–223.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*.

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.

Henderson, C. (1984). Applications of linear models in animal breeding (university of guelph, guelph, on, canada). *Applications of linear models in animal breeding. University of Guelph, Guelph, ON, Canada.*

Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

Howe, C. A., Staudenmayer, J. W., and Freedson, P. S. (2009). Accelerometer prediction of energy expenditure: vector magnitude versus vertical axis. *Med Sci Sports Exerc*, 41(12):2199–206.

Hsieh, Y.-H. (2009). Richards model: a simple procedure for real-time prediction of outbreak severity. In *Modeling and dynamics of infectious diseases*, pages 216–236. World Scientific.

Hsieh, Y.-H. (2010). Pandemic influenza a (h1n1) during winter influenza season in the southern hemisphere. *Influenza and Other Respiratory Viruses*, 4(4):187–197.

Hsieh, Y.-H. and Chen, C. (2009). Turning points, reproduction number, and impact of climatological events for multi-wave dengue outbreaks. *Tropical Medicine & International Health*, 14(6):628–638.

Hsieh, Y.-H., Fisman, D. N., and Wu, J. (2010). On epidemic modeling in real time: An application to the 2009 novel a (h1n1) influenza outbreak in canada. *BMC research notes*, 3(1):1–8.

Hsu, F., Nelson, C., and Chow, W. (1984). A mathematical model to utilize the logistic function in germination and seedling growth. *Journal of Experimental Botany*, 35(11):1629–1640.

Hu, B., Guo, H., Zhou, P., and Shi, Z.-L. (2020). Characteristics of sars-cov-2 and covid-19. *Nature Reviews Microbiology*, pages 1–14.

Huerta, G. and West, M. (1999). Priors and component structures in autoregressive time series models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4):881–899.

Idrovo, A. J. and Manrique-Hernández, E. F. (2020). <? covid19?> data quality of chinese surveillance of covid-19: Objective analysis based on who's situation reports. *Asia Pacific Journal of Public Health*, 32(4):165–167.

IEA, I. E. A. (2019). Statistics resources - balance definitions. `https://www.iea.org/statistics/resources/balancedefinitions/`. Accessed: 2019-11-24.

Ijarchelo, S. M., Afereydoon, K., and Zamanzadeh, L. (2016). Bayesian variable selection under collinearity of parameters. *Reserach journal of applied sciences*, 11:428–438.

Ioannidis, J. P., Cripps, S., and Tanner, M. A. (2020). Forecasting for COVID-19 has failed. *International Journal of Forecasting*.

Ishwaran, H. (1999). Applications of hybrid monte carlo to bayesian generalized linear models: Quasicomplete separation and neural networks. *Journal of Computational and Graphical Statistics*, 8(4):779–799.

Jalilian, A. and Mateu, J. (2021). A hierarchical spatio-temporal model to analyze relative risk variations of COVID-19: a focus on Spain, Italy and Germany. *Stochastic Environmental Research and Risk Assessment*, pages 1–16.

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.

Joseph, M. (2016). Exact sparse CAR models in Stan, 2016. *URL http://mc-stan.org/users/documentation/case-studies/mbjoseph-CARStan. html*.

Kahm, M., Hasenbrink, G., Lichtenberg-Fraté, H., Ludwig, J., and Kschischo, M. (2010). Grofit: fitting biological growth curves. *Nature Precedings*, pages 1–1.

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.

Karantonis, D. M., Narayanan, M. R., Mathie, M., Lovell, N. H., and Celler, B. G. (2006). Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. *IEEE transactions on information technology in biomedicine*, 10(1):156–167.

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.

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.

Koester, U. and Galaktionova, E. (2021). Fao food loss index methodology and policy implications. *STUDIES IN AGRICULTURAL ECONOMICS*, 123(1):1–7.

Kreft, I. G., Kreft, I., and de Leeuw, J. (1998). *Introducing multilevel modeling*. Sage.

Kuiper, M. and Cui, H. D. (2020). Using food loss reduction to reach food security and environmental objectives–a search for promising leverage points. *Food Policy*, page 101915.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Laine, M. et al. (2008). *Adaptive MCMC methods with applications in environmental and geophysical models*. Finnish Meteorological Institute.

LaMotte, L. R. (2014). Fixed-, random-, and mixed-effects models. *Wiley StatsRef: Statistics Reference Online*.

Lang, S. and Brezger, A. (2004). Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1):183–212.

Lauritzen, S. L. (1996). *Graphical Models*, volume 17. Clarendon Press.

Lawson, A. B. (2018). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. CRC press.

Lee, I.-M., Hsieh, C.-c., and Paffenbarger, R. S. (1995). Exercise intensity and longevity in men: The harvard alumni health study. *Jama*, 273(15):1179–1184.

Lee, S. Y., Lei, B., and Mallick, B. (2020). Estimation of covid-19 spread curves integrating global data and borrowing information. *PloS one*, 15(7):e0236860.

Leimkuhler, B. and Reich, S. (2004). *Simulating hamiltonian dynamics*. Number 14. Cambridge university press.

Leroux, B. G., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 179–191. Springer.

Liang, K. (2020). Mathematical model of infection kinetics and its analysis for COVID-19, SARS and MERS. *Infection, Genetics and Evolution*, 82:104306.

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.

Lloyd-Sherlock, P., Sempe, L., McKee, M., and Guntupalli, A. (2021). Problems of data availability and quality for covid-19 and older people in low-and middle-income countries. *The Gerontologist*, 61(2):141–144.

Longford, N. T. (1995). Random coefficient models. In *Handbook of statistical modeling for the social and behavioral sciences*, pages 519–570. Springer.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337.

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, J., Dushoff, J., Bolker, B. M., and Earn, D. J. (2014). Estimating initial epidemic growth rates. *Bulletin of mathematical biology*, 76(1):245–260.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

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.

Mingione, M. and Alaimo Di Loro, P. (2021). Statistical communication of covid-19 epidemic using widely accessible interactive tools. *Book of Short Papers - SIS 2021*, pages 1738–1743.

Mingione, M., Alaimo Di Loro, P., Farcomeni, A., Divino, F., Lovison, G., Jona Lasinio, G., and Maruotti, A. (2021a). Spatial modelling of covid-19 incidence cases using richards' curve: an application to the italian regions. *Spatial Statistics*.

Mingione, M., Fabi, C., and Jona Lasinio, G. (2021b). Measuring and modeling food losses. *Journal of Official Statistics*, 37(1):171–211.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

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.

Morris, A. K. and Silk, W. K. (1992). Use of a flexible logistic function to describe axial growth of plants. *Bulletin of Mathematical Biology*, 54(6):1069–1081.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT press.

Nash, J. C., Varadhan, R., Grothendieck, G., Nash, M. J. C., and Yes, L. (2020). Package 'optimx'.

Neal, P. and Roberts, G. (2006). Optimal scaling for partially updating mcmc algorithms. *The Annals of Applied Probability*, 16(2):475–515.

Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117.

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.

Peeri, N. C., Shrestha, N., Rahman, S., Zaki, R., Tan, Z., Bibi, S., Baghbanzadeh, M., Aghamohammadi, N., Zhang, W., and Haque, U. (2020). The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? *International Journal of Epidemiology*.

Pelagatti, M. M. and Maranzano, P. (2021). Assessing the effectiveness of the italian risk-zones policy during the second wave of covid-19. *University of Milan Bicocca Department of Economics, Management and Statistics Working Paper*.

Peruzzi, M., Banerjee, S., and Finley, A. O. (2020). Highly scalable bayesian geostatistical modeling via meshed gaussian processes on partitioned domains. *Journal of the American Statistical Association*, 0(0):1–14.

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.

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. (2003). JAGS: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). Available at* `https://www.r-project.org/conferences/DSC-2003/Proceedings/`. *Accessed: 2020-01-18*.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics*, 9(501-538):105.

Polson, N. G. and Scott, J. G. (2012). Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):287–311.

Ramsay, J. O. and Silverman, B. W. (2007). *Applied Functional Data Analysis: Methods and Case Studies*. Springer.

Rathod, S. B. and Pattewar, T. M. (2015). Content based spam detection in email using bayesian classifier. In *2015 International Conference on Communications and Signal Processing (ICCSP)*, pages 1257–1261. IEEE.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, volume 1. sage.

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.

Richards, F. (1959). A flexible growth function for empirical use. *Journal of experimental Botany*, 10(2):290–301.

Ritz, C., Baty, F., Streibig, J., and Gerhard, D. (2015). Dose-response analysis using R. *PLoS ONE*, 10:e0146021.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Robinson, G. K. et al. (1991). That blup is a good thing: the estimation of random effects. *Statistical science*, 6(1):15–32.

Rothney, M. P., Schaefer, E. V., Neumann, M. M., Choi, L., and Chen, K. Y. (2008). Validity of physical activity intensity predictions by actigraph, actical, and rt3 accelerometers. *Obesity*, 16(8):1946–1952.

Royle, J. A. and Dorazio, R. M. (2008). *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Elsevier.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Rushworth, A., Lee, D., and Mitchell, R. (2014). A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial and spatio-temporal epidemiology*, 10:29–38.

Rushworth, A., Lee, D., and Sarran, C. (2017). An adaptive spatiotemporal smoothing model for estimating trends and step changes in disease risk. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(1):141–157.

Salje, H., Tran Kiem, C., Lefrancq, N., Courtejoie, N., Bosetti, P., Paireau, J., Andronico, A., Hozé, N., Richet, J., Dubost, C.-L., Le Strat, Y., Lessler, J., Levy-Bruhl, D., Fontanet, A., Opatowski, L., Boelle, P.-Y., and Cauchemez, S. (2020). Estimating the burden of SARS-CoV-2 in France. *Science*.

Santos-Lozano, A., Santin-Medeiros, F., Cardon, G., Torres-Luque, G., Bailon, R., Bergmeir, C., Ruiz, J. R., Lucia, A., and Garatachea, N. (2013). Actigraph gt3x: validation and determination of physical activity intensity cut points. *International journal of sports medicine*, 34(11):975–982.

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.

Schmidt, M. N. (2009). Function factorization using warped gaussian processes. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 921–928.

Scrucca, L. (2013). GA: A package for genetic algorithms in R. *Journal of Statistical Software*, 53:1–37.

Searle, S., Casella, G., and McCulloch, C. (1992). Variance components john wiley and sons. *Inc. New York*.

Sebastiani, G., Massa, M., and Riboli, E. (2020). COVID-19 epidemic in Italy: evolution, projections and impact of government measures. *European Journal of Epidemiology*, 35:341–345.

Shaman, J. and Galanti, M. (2020). Will SARS-CoV-2 become endemic? *Science*, 370:527–529.

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.

Sinharay, S., Stern, H. S., and Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological methods*, 6(4):317.

Smith, A. F. and Roberts, G. O. (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1):3–23.

Smith, A. F. M. (1991). Bayesian computational methods. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 337(1647):369–386.

Snijders, T. A. and Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). Bugs 0.5: Bayesian inference using gibbs sampling manual (version ii). *MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK*, pages 1–59.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):485–493.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

Stan Development Team (2021). Stan modeling language users guide and reference manual, 2.26.

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.

Stekhoven, D. J. and Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Sturtz, S., Ligges, U., and Gelman, A. E. (2005). R2winbugs: a package for running winbugs from r.

Su, Y.-S., Yajima, M., Su, M. Y.-S., and SystemRequirements, J. (2015). Package 'r2jags'. *R package version 0.03-08, URL http://CRAN. R-project. org/package= R2jags*.

Svetliza, C. F. and Paula, G. A. (2003). Diagnostics in nonlinear negative binomial models. *Communications in Statistics-Theory and Methods*, 32(6):1227–1250.

Tanne, J. H. (2021). Covid-19: Fda approves pfizer-biontech vaccine in record time.

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.

Terenin, A., Dong, S., and Draper, D. (2019). Gpu-accelerated gibbs sampling: a case study of the horseshoe probit model. *Statistics and Computing*, 29(2):301–310.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tjørve, E. and Tjørve, K. M. (2010). A unified approach to the richards-model family for use in growth analyses: Why we need only two model forms. *Journal of Theoretical Biology*, 267(3):417 – 425.

Tjørve, K. and Tjørve, E. (2017). The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the unified-richards family. *PLoS ONE*, 12(6):e0178691.

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.

Tsoularis, A. and Wallace, J. (2002). Analysis of logistic growth models. *Mathematical biosciences*, 179(1):21–55.

Varotsos, C. A. and Krapivin, V. F. (2020). A new model for the spread of covid-19 and the improvement of safety. *Safety science*, 132:104962.

Vasudevan, V., Gnanasekaran, A., Sankar, V., Vasudevan, S. A., and Zou, J. (2021). Disparity in the quality of covid-19 data reporting across india. *BMC public health*, 21(1):1–12.

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.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432.

Wachenheim, D. E., Patterson, J. A., and Ladisch, M. R. (2003). Analysis of the logistic function model: derivation and applications specific to batch cultured microorganisms. *Bioresource Technology*, 86(2):157–164.

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., and Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, 3(8).

Waller, L. and Carlin, B. (2010). Disease mapping. In Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M., editors, *Handbook Of Spatial Statistics*, Boca Raton, FL. Chapman and Hall/CRC Press.

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

Werker, A. and Jaggard, K. (1997). Modelling asymmetrical growth curves that rise and then fall: Applications to foliage dynamics of sugar beet (beta vulgarisl.). *Annals of Botany*, 79(6):657–665.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399.

Wu, K., Darcet, D., Wang, Q., and Sornette, D. (2020). Generalized logistic growth modeling of the covid-19 outbreak: comparing the dynamics in the 29 provinces in china and in the rest of the world. *Nonlinear dynamics*, 101(3):1561–1581.

Wu, L. (2009). *Mixed Effects Models for Complex Data*. CRC Press.

Acknowledgements

I would not have gone this far, if it wasn't for some very special people I met along my life path, to whom I owe a large part of what I achieved.

First of all, I wish to thank with all my heart my supervisor, Prof. Giovanna Jona Lasinio. She took care of me since I was a young (and very anxious) student and raised me as the researcher I am today, always bringing out the best in me, never letting me down. She has always been supportive, in the bad and the good times, encouraging me to look at the glass as half-full, and I hope at least to have made her proud.

I am lucky to have found Pierfrancesco Alaimo Di Loro, the best research partner in crime everyone would love to have by their side. He is a true friend and the most brilliant colleague I know. Working with him makes easy even the most challenging task, as his passion and dedication are contagious and inspiring. Grazie assai.

A huge thank is mandatory for the remaining members of StatGroup-19: Prof. Fabio Divino, Prof. Alessio Farcomeni, Prof. Gianfranco Lovison, Prof. Antonello Maruotti. They made me see first-hand the importance of the social role of statistics, involving me in one of the most stimulating research adventures I could ever imagine. They definitely contributed to my growth as an academic and a researcher to a large extent.

I would also like to thank Dr. Carola Fabi for giving me the opportunity of working at FAO. Being part of such a prestigious organization under her guidance, and contributing to the methodology of SDG 12.3.1, made me feel useful for the worldwide community for the first time. I could never repay her for such privilege.

I will always be grateful to Dr. Massimo Bernaschi for introducing me to the academic career. His neat mentoring when I was a researcher at the Institute of Applied Computing "M. Picone" (IAC – CNR) steered me in the right direction everytime I was in need. I always admired his pragmatism and openness. He put his trust in me from the very beginning and made me part of his wonderful team: thanks to Alessandro C., Dario P., Elena A., Enrico M., Marco C. and Stefano G. . A special mention goes to Mario Santoro, who lit the fire of my curiosity and taught me how to "think in `R`". His wholehearted and constant support was priceless. His programming skills and comprehensive expertise makes him my first academic role model.

I wish to offer my deepest thanks to Prof. Sudipto Banerjee, Chair of the Dpt. of Biostatistics at UCLA. Chapter 4 would not exist if it wasn't for him. He supervised the project during my visiting research period in California and made me feel at home even if I was in a foreign country. I studied on his books and followed his research since the beginning of my university career, and I will never forget the thrill I felt when I knew I could collaborate with him. The weekly meetings we had, full of enlightening discussions, are something I will always cherish.

A truthful thank goes to Jonah Lipsitt: a wise researcher, a great teammate, a better friend.

Also a very big hug to Pierre, Geshe, Arslan, Tom, Tommaso, Marco, Rose, Nisa, Tsylia, Ciera, Brian, Fred and all the UCHA-Coop family in Los Angeles for mitigating my homesickness with their warmth.

Even if just a word won't be enough, thanks to the love of my life, Sbiri. I would be half the man I am if it wasn't for her. Her infinite kindness and thoughtful caring brighten my days and I could never imagine my future without her. To her patience and love I owe everything I have. Seeing her smile makes me the happiest person alive, and I hope I could make her feel the same for the rest of my days and beyond. Ti respiro e ti trattengo, per averti per sempre, oltre il tempo di questo momento.

Un grazie infinito a tutta la mia famiglia, per il calore e l'affetto incondizionato che mi hanno sempre dimostrato, e senza il quale non sarei riuscito ad affrontare i momenti più difficili.
In particolare, grazie ai miei genitori, Angelo e Roberta, per avermi supportato e sopportato sempre. Non sono molto bravo a dimostrarvelo e spesso vi rompo le scatole, ma vi voglio un bene infinito, e spero di avervi reso fieri e felici almeno la metà di quanto fiero e felice mi sento io ad essere vostro figlio.
Grazie a mio fratello Mattia, la cui spensieratezza ha reso leggeri anche i giorni di studio più intensi.
Grazie a Nonna Maria, alla sua saggezza e al suo sorriso devo la mia forza.

My truthful thanks to all my friends who keep filling my life with joy and care everyday. It is impossible to list all of them, but I would like to say that each one has been extremely important. A special mention goes:
A Francesco e Matteo, amici di una vita, per la vita.
Ai 'ragazzi', Paola, Claudio, Roberto, Gianni, Nicola, Rosa, Marcella e Alfredo, che hanno sempre nutrito la mia curiosità fin da quando ero bambino. Dalle equazioni sulla tovaglia di carta in pizzeria al dottorato il passo è breve: non poteva finire diversamente con degli insegnanti come voi.
To Eugenio, Enrico, and Jacopo, friends and roommates who made me feel home for the first time far from my family.
To Federico, Kevyn and Riccardo, friends and classmates whom I shared joys and sorrows of the university. They taught me that dedication and hard-work always pay back, and I will always be grateful to them.