UNIVERSITÀ DEGLI STUDI DI BERGAMO

Scuola di Alta Formazione Dottorale

Dipartimento di Ingegneria e Scienze Applicate

Corso di Dottorato in Mathematics and Applied Statistics

Ciclo XXXVIII

A STATE-SPACE MODELLING FRAMEWORK IN GEOSTATISTICS WITH APPLICATION TO ENVIRONMENTAL DATA

Supervisore:

Chiar.mo Prof. Alessandro Fassò

Tesi di dottorato

Jacopo RODESCHINI

Matricola n. 1046083

ANNO ACCADEMICO 2024/ 2025

Acknowledgments

First, I would like to thank my advisor, Alessandro Fassò, for introducing me to the field of spatial statistics a long time ago and for all the help and many interesting discussions. His constant guidance, encouragement, and ability to open new perspectives have been invaluable throughout my academic journey. He shaped the way I think about research.

I am also grateful to Francesco Finazzi for all the help and many interesting discussions in making this thesis. Our many discussions have not only helped me solve technical problems but have also given me a broader understanding of the subject, making this work a much richer experience.

During my PhD journey, I visited the University of Glasgow, Department of Mathematics and Statistics, and I would like to thank Philipp Otto for all his great ideas, enthusiasm, and help along the way, especially in connection with our joint work on Paper Otto, Fusta Moro, et al. (2024). My thanks also to the fellow PhD student Lorenzo Tedesco for numerous comments on all parts of this thesis and, in particular, for our joint work Rodeschini, Tedesco, et al. (2025);

I would also like to thank all my colleagues at the Engineering and Applied Sciences Department at the University of Bergamo who, in one way or another, have supported me during the past three years. I am especially grateful to Alessandro and Andrea for their friendship and for making this PhD journey truly meaningful.

As far as I am concerned, I'm very happy to complete my PhD in such a beautiful year. The year of the Jubilee. This has been a very important experience, one that has given me the key to reading the world in a new light. To bring hope to the world is one of life's most beautiful callings, and I wish to keep this moment alive in my heart for a long time.

Finally, I am sincerely grateful to my friends and family for always being there.

Nothing I've done is important in absolute terms; that's obvious, but it belongs to me totally and identifies me.

Walter Bonatti - *Montagne di una Vita, 1995*

Table of Contents

Abstract

This thesis proposes a framework based on geostatistical models for the analysis of environmental datasets. Geostatistical models rely on stochastic fields (also called random fields or stochastic processes), which are usually described by a small number of parameters that can be estimated and modelled; they aim at building a process that only mimics some patterns of the observed space-time variability, without necessarily following the underlying governing "*true*" equations. The model randomness may not originate from the actual physical process; it is most often a (model-dependent) measure of uncertainty regarding the unique physical process, whose parameters are not completely known. Stochastic models are then considered as alternative approaches until more elaborate physically based models become available.

Geostatistical space–time models, also referred to as spatio-temporal models, are rapidly evolving to face this surge of novel data structures in various application fields: geology, meteorology and climate science, ecology, epidemiology, economics, politics, and more. Research in space and space-time analysis is of broad and current interest in the statistical research literature and applications. Stochastic models that extend spatial statistics to include the additional time dimension are presented throughout the thesis with a common notation to facilitate comparison. Two main conceptual viewpoints are distinguished: *(i)* approaches involving a single spatio-temporal random function model, and *(ii)* dynamical-statistical approach, which involves vectors of space random functions or vectors of time series. Moreover, links between these two viewpoints are then revealed, with advantages and shortcomings highlighted.

While there are many problems for which the statistical approach *(i)* provides satisfactory solutions, it does not explicitly model the aetiology of the phenomenon under study. In contrast, the dynamical-statistical approach models how the current state depends on previous states through dynamical relationships that are either mechanistic or probabilistic. The rationale for adopting the dynamical specification is discussed in detail in this thesis. These models are typically specified through a set of parameters that are unknown in practice and must be estimated from observed data. Estimation is naturally formulated within a state-space framework, where the unobserved state process can be in-

ferred using the Kalman filter or Kalman smoother when parameters are known. In the more common case where parameters are unknown, the standard approach is to use the expectation–maximisation (EM) algorithm to get the maximum-likelihood estimates (MLE).

We will focus on the Gaussian state-space formulation; thus, the space random functions are assumed Gaussian. Furthermore, we consider the joint modelling of multiple spatial variables, known as multivariate analysis, which allows for improved predictive performance and enables a deeper understanding of the cross-dependencies among processes observed in the same spatial domain. We use a variant of the state-space model to explain the particulate matter, $PM_{2.5}$ and PM_{10} respectively. The results of leave-one-out cross-validation show that the state-space model outperforms benchmark approaches. Building upon these results, we propose a heteroskedastic space-time model used to estimate the impact of ammonia (NH_3) emissions on the concentration of $PM_{2.5}$ in the Lombardy region, Italy.

With the modern set of larger and larger spatial datasets, however, exact inference of Gaussian processes becomes computationally infeasible. Datasets from remote-sensing platforms and sensor networks are often very large. Processing massive amounts of data to provide current estimates of the state from current and past data is challenging, even for the Kalman filter. A large number of spatial locations observed through time can quickly lead to an overwhelmingly high-dimensional statistical model. As a result, efficient approximations and scalable methods are needed for Gaussian processes inference to modern spatio-temporal datasets.

To deal with the large spatial-temporal regression problems, we develop a variant of the state-space model by exploiting the low-rank approximation of a Gaussian process. The method is based on that the Gaussian process, under general conditions, can be viewed as a solution to a certain stochastic partial differential equation (SPDE). This reduces the computational cost of the Kalman filter, while preserving the physical interpretation of the original model. We prove theoretical results, including asymptotic properties and error bounds for the proposed low-rank model. The MLE of the model parameters is carried out via EM, which results in closed-form updates for most parameters. Extensive simulations validate both the accuracy and efficiency of the method. Although our discussion focuses on domains $\Omega \subset \mathbb{R}^2$, the approach extends naturally to higher-dimensional manifolds, and future studies may further adapt it to networks and graphs, highlighting the model's flexibility and broad applicability.

Chapter 1

Introduction

This thesis presents a modelling framework for spatio-temporal processes. Spatio-temporal statistics is a discipline rooted in geostatistics and focuses on the analysis of data indexed simultaneously by location (space) and time. In recent years, spatio-temporal analysis has gained increasing relevance in the geosciences, driven by large-scale sampling campaigns, the rapid growth of data availability, and diverse application domains. These include environmental monitoring, climate science, geology, forestry, and astronomy, as well as any discipline that relies on spatially distributed data and seeks to model dependencies across both space and time.

Environmental time series often exhibit distinctive features that set them apart from other types of temporal data. Classical time series models are typically built on the assumption of identically distributed errors that are dependent across equally spaced time points. The equal spacing and unidirectional flow of time underlie the construction of these models. In contrast, an important characteristic of environmental time series is the spatial correlation among observations. Environmental phenomena are inherently interconnected not only over time but also across space, requiring models that account for both dimensions simultaneously. Such spatio-temporal models must be more flexible than purely temporal ones. For instance, it is generally unrealistic to assume that spatial observation sites are arranged on a regular grid, or that concepts such as “past”, “present”, and “future” apply in the same way to the spatial dimension as they do to the temporal one.

As discussed above, environmental time series often exhibit dependencies that extend beyond the temporal dimension. This motivates the use of spatio-temporal statistics, a discipline that analyses data indexed by both space and time (Cressie and Wikle, 2015). Temporal dependence arises because phenomena evolve in such a way that past observations causally influence future ones. Spatial dependence, by contrast, is typically symmetric and instantaneous: nearby locations exert mutual influence, with the strength of this dependence generally decaying with distance. Introducing spatial correlation

amounts to assuming that a single underlying process is observed across multiple spatial locations.

The literature on spatial methods is relatively recent but already extensive, reflecting the fact that each application often requires specific modelling strategies tailored to the data type and spatial structure at hand. Different dependence structures inevitably lead to distinct probabilistic assumptions and modelling tools. A comprehensive review of spatial methods from various perspectives can be found in Cressie and Wikle (2015), Stein (1999), and Wikle, Zammit-Mangion, and Cressie (2019), while further approach-specific accounts are available in Cressie and Johannesson (2008), Lawson (2018), and Banerjee et al. (2008).

For the reasons discussed so far, an important extension of geostatistics is to jointly accommodate both spatial and temporal correlation, a task generally referred to as spatio-temporal modelling. While spatial statistics is often framed in terms of subsets of $\mathbb{R}^2$ equipped with the Euclidean norm, from a purely mathematical perspective, it is natural also to consider time as an additional dimension. Indeed, the temporal domain $\mathcal{T} \subset \mathbb{R}^+$ can be viewed as a one-dimensional space endowed with an ordered structure, where the notions of "before" and "after" play a central role. Historically, however, the high costs of collecting data across both space and time limited the development of spatio-temporal statistics, confining it for decades to small, application-specific studies in the environmental sciences.

In recent years, large-scale monitoring campaigns have produced multiple spatial variables across environmental, agricultural, and climate studies. The observed values of these variables constitute multivariate spatial data and often exhibit an additional form of spatial dependence, namely spatial cross-correlation, describing the correlation between different variables measured at the same or at different locations. The most informative configuration is the isotopic case, in which all variables are observed at every site. At the opposite extreme, the completely heterotopic configuration occurs when each variable is measured only at distinct, non-overlapping locations. Between these two extremes lies the partially heterotopic setting, where some variables share some sites. Modelling both the spatial correlation and the spatial cross-correlation can lead to improvements in prediction and interpretation (H. Zhang, 2007; Wackernagel, 2003). To this end, multivariate spatio-temporal models for heterotopic data have been proposed, designed to capture such complex interdependencies (Finazzi, E. M. Scott, and Fassò, 2013; Cressie and Wikle, 2002; Kyriakidis and Journel, 1999). These models not only improve predictive accuracy but also enhance our ability to understand the underlying processes and to support evidence-based management and decision-making. To highlight the practical implications of spatio-temporal modelling, two motivating examples of spatial prediction are presented in Section 1.1, where the goal is to predict pollutant concentrations at unobserved locations and quantify the associated uncertainty.

A particularly prominent class of models for both univariate and multivariate spatio-temporal data are based on Gaussian random fields, also known as Gaussian processes (GPs; Rasmussen (2003)). Among the various modelling approaches, GPs have become a widely used and powerful tool for regression in spatial and spatio-temporal settings. Their popularity stems from their flexibility and their natural ability to capture spatial and temporal dependencies (Stein, 1999). In this framework, the underlying regression function is modelled as a realisation of a GP, fully characterised by its mean and covariance functions. Exploiting the properties of the multivariate normal distribution, GPs yield closed-form expressions for both predictions and their associated uncertainty.

While GPs provide a flexible and theoretically appealing framework, their practical use is often limited by the high computational cost associated with large covariance matrices. This challenge is particularly acute in modern applications, where the growing availability of massive climate datasets generated by remote sensing satellites has made computational efficiency a crucial requirement. A significant part of this thesis is therefore devoted to the development of models capable of handling multivariate spatio-temporal phenomena while remaining computationally feasible. Many real-world applications demand a careful balance between model flexibility and computational efficiency. To highlight this trade-off, a motivating example of spatial prediction is presented in Section 1.1.2, where the goal is to perform inference and prediction on a spatio-temporal process at linear computational complexity, while handling a large spatio-temporal dataset. The examples in Section 1.1 illustrate the practical challenges that motivate the methodological contributions of this thesis.

1.1 Motivating example

1.1.1 Air pollution and livestock

Air pollution represents a major public health and environmental challenge worldwide, with particularly acute problems in industrialised and densely populated regions. According to the World Health Organisation (WHO) Henschel and Chan (2013), the most harmful air pollutants to human health and ecosystems are ground-level ozone (O_3), nitrogen dioxide (NO_2), and particulate matter (PM). Among these, fine particulate matter ($PM_{2.5}$), defined as a mixture of airborne particles with a diameter smaller than 2.5 μm (see also Jerrett et al., 2005, for a review), has been identified as the most hazardous air pollutant (European Environmental Agency, 2022). Exposure to high concentrations of $PM_{2.5}$ has direct and well-documented effects on human health, especially over long periods, and is strongly associated with respiratory and cardiovascular diseases (Dominici et al., 2006; Yang et al.,

2019; Boldo et al., 2014). In Europe, the European Environment Agency (EEA) estimates that about 400.000 premature deaths per year can be attributed to long-term exposure to $PM_{2.5}$ concentrations (European Environment Agency (EEA), 2022).

On a global scale, WHO benchmark limits are set at 20 $\mu g/m^3$ for PM_{10} and 10 $\mu g/m^3$ for $PM_{2.5}$, yet these thresholds are frequently exceeded (European Environmental Agency (EEA) - Air pollution section, 2021). It was estimated that in 2018, around 90% of the world's population was exposed to polluted air (European Environmental Agency, 2018). In Europe specifically, between 6–8% of the population was exposed to $PM_{2.5}$ concentrations above the WHO limit, and 13–19% to levels of PM_{10} exceeding the threshold (European Environmental Agency (EEA) - Air pollution section, 2021). The European Union has adopted less restrictive limits of 50 $\mu g/m^3$ for PM_{10} and 25 $\mu g/m^3$ for $PM_{2.5}$ as annual averages. However, even under these standards, compliance remains a challenge, with several countries unable to meet them consistently; Italy is a notable example (Kiesewetter et al., 2015).

The Lombardy region, located in the heart of the Po Valley in Northern Italy, is among the most polluted areas in Europe, largely due to the natural barrier created by the Alps, which hinders the dispersion of air pollutants (see, e.g., Pernigotti et al., 2012). The geographical stability of the valley, combined with limited air circulation, often leads to the accumulation of fine particulate matter ($PM_{2.5}$) at high concentrations, with particularly severe effects on air quality. This situation is exacerbated during the colder months, when the low altitude of the atmospheric mixing layer coincides with elevated emissions from road traffic and domestic heating systems. Figure 1 shows the weekly averages of $PM_{2.5}$ concentrations across Lombardy. As expected, a clear seasonal pattern emerges, with higher concentrations during winter and lower levels in summer. Median concentrations range between 10 $\mu g/m^3$ and 40 $\mu g/m^3$ throughout the year, consistently exceeding the WHO guideline of 10 $\mu g/m^3$ for long-term exposure (WHO, 2021). Figure 2 shows a scatter plot of the average concentrations of PM_{10} and $PM_{2.5}$ for each monitoring station, over the augmented Lombardy region obtained by applying a 0.3° buffer around the regional borders. The figure highlights that annual concentration levels in the Po Valley, located in the southern part of the region, are higher than in the northern mountainous areas. Furthermore, spatial homogeneity is supported by the map, which reveals clusters of neighbouring stations with similar concentration levels, suggesting pronounced spatial dependence.

PM emission can have both primary and secondary origins. Primary PM is emitted directly into the atmosphere from various sources, while secondary PM forms in the atmosphere through chemical transformations of precursor pollutants such as ammonia (NH_3), nitrogen oxides (NO_x), and sulphur dioxide (SO_2). In the Po Valley, the composition of atmospheric particulate matter during winter contains a substantial secondary component, often exceeding 50%. This proportion can be even

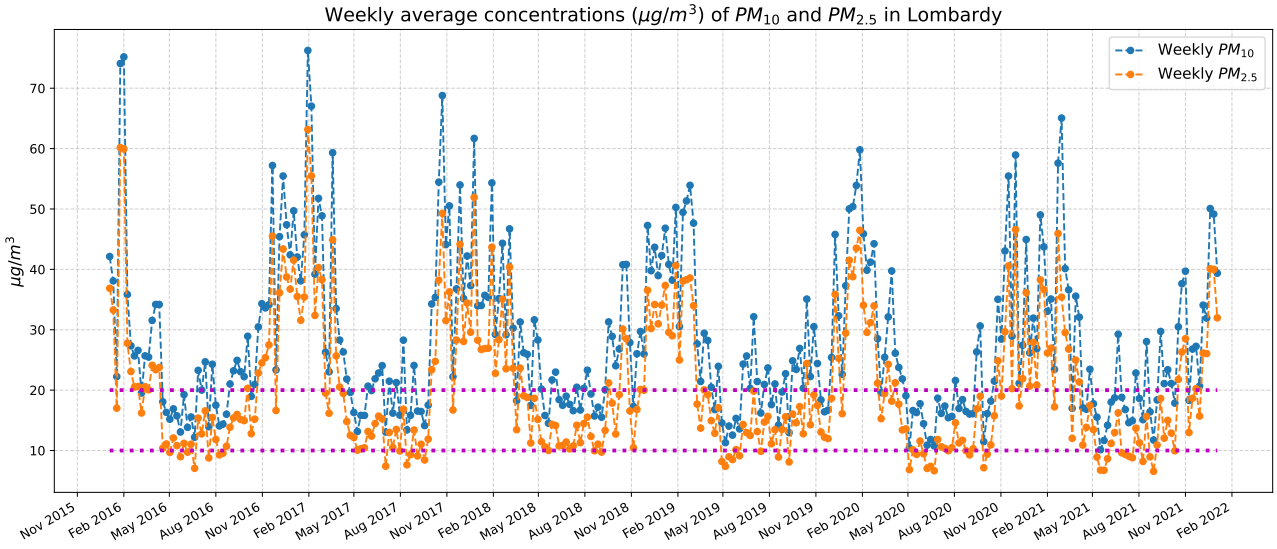


Figure 1: Time series of weekly averaged PM_{10} and $PM_{2.5}$ concentrations in the Lombardy region from January 2016 to December 2021. The weekly averages are computed from 105 stations for PM_{10} and 49 stations for $PM_{2.5}$. Pink horizontal lines indicate the WHO annual thresholds of 20 $\mu g/m^3$ for PM_{10} and 10 $\mu g/m^3$ for $PM_{2.5}$.

higher during accumulation episodes under meteorological conditions that are unfavourable for dispersion. Furthermore, the region is surrounded on three sides by the Alpine and Apennine mountain chains, which limit the dispersion of pollutants; it is also characterised by weak ventilation, favouring the stagnation and accumulation of fine particulate matter. In addition, Lombardy is Italy's leading industrial and agricultural hub (Lovarelli, Conti, et al., 2020; Fassò, Rodeschini, et al., 2023a), and the associated emissions enhance the secondary PM emissions, further worsening the already critical air quality. Consequently, Lombardy is considered one of the most polluted regions in Europe and exhibits among the highest $PM_{2.5}$-related mortality rates on the continent, with an estimated 100–150 premature deaths per 100.000 inhabitants (European Environment Agency (EEA), 2022).

Ammonia (NH_3) is a key precursor of secondary PM. The association between ammonia and PM has been extensively studied in the literature (Clappier et al., 2021; Lonati and Cernuschi, 2020; Lovarelli, Fugazza, et al., 2021; Veratti et al., 2023; Fassò, 2023). There is a broad scientific consensus that livestock and fertilisers are the primary sources of ammonia emissions (Nenes et al., 2020; Thunis et al., 2021). In the Lombardy region, these activities account for up to 97% of all ammonia emissions (ARPA Lombardia Settore Monitoraggi Ambientali, INEMAR, 2022), which then undergo chemical reactions in the atmosphere to form $PM_{2.5}$. Figure 3 illustrates this relationship: areas with higher livestock density, including both pigs and cattle, correspond to higher ammonia emissions, and vice versa. According to the Lombardy Regional Environment Protection Agency (ARPA Lombardia Settore Monitoraggi Ambientali, 2019), the reaction of ammonia with nitric acid can contribute up to 60% of the $PM_{2.5}$ total mass concentration. Consequently, incorporating livestock data into air quality

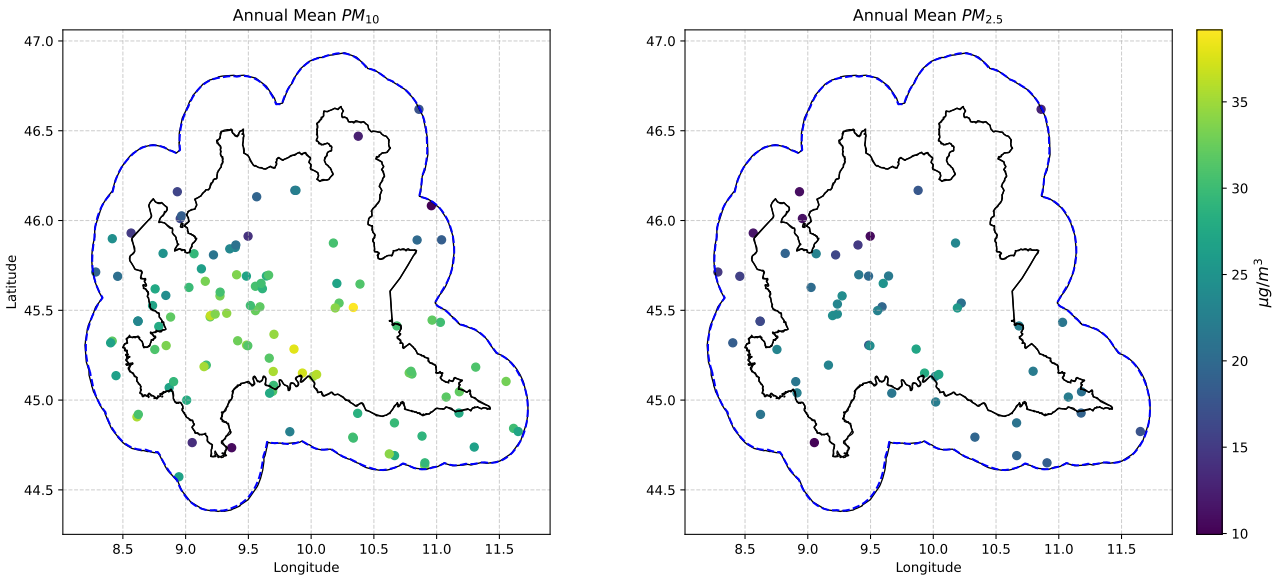


Figure 2: Average concentration of PM_{10} and $PM_{2.5}$ ($\mu g/m^3$) by 105 and 49 stations in the Lombardy region. The black line shows the Lombardy regional borders. While the blue solid line depicts the buffer of 0.3° that defines the neighbouring area useful to avoid boundary effects.

models can improve understanding of, and inform strategies to mitigate, the impact of livestock on $PM_{2.5}$ concentrations.

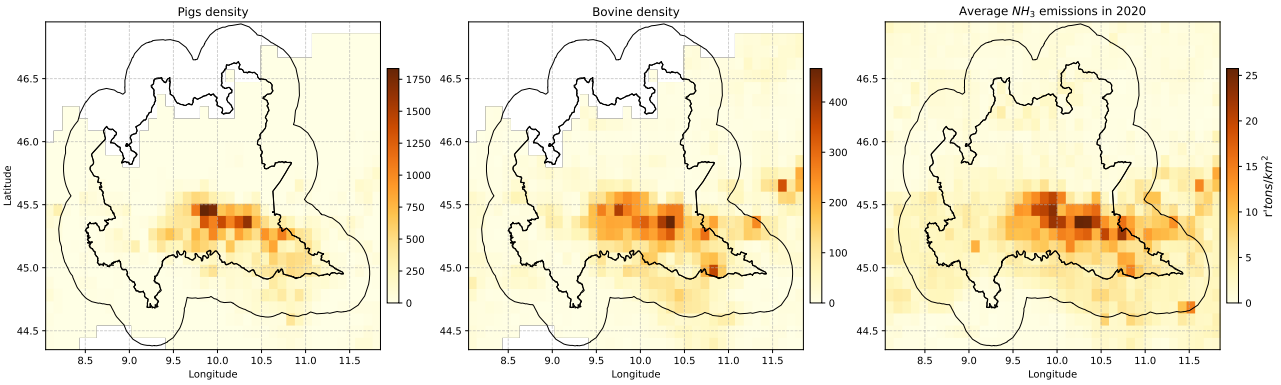


Figure 3: Density of pigs and bovines ($number/km^2$) in the left and centre panels, and ammonia (NH_3) emissions (t/km^2) over the Lombardy region in 2020 on the right. The black line indicates the augmented Lombardy region, obtained by applying a 0.3° buffer around the regional borders.

In light of the above discussion, the main scientific questions of interest are the following:

- How can datasets from satellites, numerical models, and in situ measurements, which have different spatial and temporal resolutions and originate from national and international agencies, be combined? In particular, how can these heterogeneous data be harmonised in time (to a daily resolution) and space (at the air quality station level) to enable integrated analysis?

- Can statistical approaches quantify the relationship between ammonia (livestock) and $PM_{2.5}$ concentrations? How do meteorological factors, such as temperature, relative humidity, and rainfall, influence this relationship?

- Is it possible to assess the contribution of ammonia emissions to $PM_{2.5}$ concentrations using a spatio-temporal geostatistical model? Can we produce a map of this impact across the Lombardy region, along with associated prediction uncertainties?

- Sampling irregularities, often due to limited access to monitoring sites or technical constraints, result in missing observations. Can these gaps be efficiently addressed using spatio-temporal modelling approaches that exploit both spatial and temporal dependencies?

1.1.2 Big dataset and multivariate analysis

With the widespread availability of remote sensing platforms, particularly satellite-based systems, both the volume and resolution of spatial data have increased dramatically (Datta et al., 2016; Heaton et al., 2019), thereby significantly amplifying the computational demands associated with Gaussian process (GP) modelling. This challenge is commonly referred to as the "Big n problem" (Jona Lasinio, Mastrantonio, and Pollice, 2013): the computational cost of inverting a covariance matrix scales poorly with the number of spatio-temporal observations $n = mT$, where m is the total number of locations and T is the number of time steps, since covariance matrices grow with complexity $O(m^2T^2)$ in storage and $O(m^3T^3)$ in inversion. These demands quickly exceed available resources under increasing data volume, rendering exact inference infeasible even with a moderately large n. As a result, efficient approximations and scalable methods are indispensable for applying Gaussian models to modern spatio-temporal datasets.

In addition to large data volumes, many environmental applications involve multivariate spatial processes, where multiple variables are observed simultaneously. Multivariate spatio-temporal models are designed to capture these complex interdependencies, enabling the joint modelling of several processes, such as $PM_{2.5}$ and PM_{10}, which often exhibit similar behaviour, see for example Figure 2. By exploiting the cross-dependence among variables, these models can improve predictive accuracy and provide a more comprehensive understanding of underlying environmental processes. However, incorporating multiple datasets simultaneously further increases computational demands, highlighting the need for efficient algorithms capable of handling both high-dimensional and multivariate spatio-temporal data.

An important application of multivariate analysis is "data fusion". The main assumption underlying

this problem is the existence of a 'true' unobserved process, with observations linked to it through measurement error models. In this framework, different data sources are interpreted as different realisations of the same process, which itself is endowed with a spatial correlation structure. Typically, a set of sources is widespread but noisy (e.g., model outputs), while another is more accurate but sparse (e.g., ground-based observations): "how to fuse all the sources to recover the unique true process?". A space-time fusion model has been proposed by McMillan et al. (2010). Following Fuentes and Raftery (2005), McMillan et al. (2010) postulated the existence of a true spatial process related to both observational data and numerical model output. However, rather than specifying the true process at the point level, McMillan et al. (2010) specified the underlying process at the block level.

However, with the increasing use of remote sensing satellites and the consequent availability of massive climate datasets, computational efficiency has become a crucial requirement. Thus, many applications demand a careful balance between the flexibility of the statistical model and the constraints imposed by computational feasibility. This need has given rise to an entire line of research within random field methods where computational considerations are placed at the centre of methodological development, and the so-called 'Big n problem' has emerged as one of the main challenges, since the increasing dimensionality of spatio-temporal data makes exact Gaussian field inference computationally prohibitive.

In light of the above discussion, the main scientific questions of interest are the following:

- Is it possible to develop a multivariate spatio-temporal model suitable for a big dataset (where big refers to both space and temporal dimension) and let the user decide the trade-off between the accuracy and computational bounds? If the model relies on an approximation, is it possible to define the approximation error upper bounds or in other words, control the approximation error?

- How can we quantify the uncertainty associated with predictions at different spatial supports (point vs. block), particularly when combining heterogeneous sources such as satellite products, numerical models, and in-situ measurements?

- Is it possible to cast such a framework to address the "data fusion" challenge by leveraging the multivariate analysis and offer a solution to the problem of upscaling and downscaling?

- Can the proposed multivariate spatio-temporal framework be tailored to provide interpretable insights into environmental processes, for example, by improving the mapping accuracy of $PM_{2.5}$ concentrations, while ensuring that results remain robust under model misspecification and data sparsity?

1.2 Outline of the thesis

The structure of this thesis summary is as follows. Chapter 2 provides an introduction to stochastic processes and their use in modelling spatial variation over continuous-space domains, with both theoretical and practical examples. In particular, Section 2.1 and Section 2.2 present the definitions of stochastic processes for univariate and multivariate outcomes, respectively, while Section 2.3 introduces spatio-temporal processes, with a focus on continuous space and discrete time formulations. Finally, section 2.4 describes statistical models for data assumed to be generated by these processes.

Chapter 3 turns to introduce spatial variation over discrete-space domains modelling. Specifically, Section 3.1 introduces a range of modelling approaches that are particularly useful for the analysis of discrete space processes, highlighting their strengths and typical applications. Section 3.2 provides a detailed discussion of Gaussian Markov Random Fields (GMRFs), emphasising their theoretical foundations, computational advantages, and their central role in modern spatial statistics.

Chapter 4 describes the creation of the *Agrimonia* dataset. This dataset is designed to support quantitative analyses of the contribution of agricultural and livestock sectors to air quality. To this end, Sections 4.1 and 4.2 provide an overview of the dataset together with a description of the data sources. Section 4.3 outlines the methodology used to harmonise the source data. Section 4.4 describes the construction of the gridded representation of the Agrimonia dataset. Most of the material used throughout this chapter is drawn from Fassò, Rodeschini, et al. (2023a).

Chapter 5 uses the Agrimonia dataset to study the daily $PM_{2.5}$ concentrations and assess the contribution of ammonia (NH_3) emissions using a spatio-temporal modelling framework. In particular, Section 5.1 provides a preliminary analysis of the dataset. To identify which model best explains the relationship between predictor variables and daily $PM_{2.5}$, in Section 5.2, three classes of models, including machine learning and statistical approaches, are compared. The comparison includes a cross-validation study and an interpretation of the estimates. The comparison study presented here is based on Otto, Fusta Moro, et al. (2024). Section 5.3 develops a spatio-temporal model to assess the sensitivity of $PM_{2.5}$ to NH_3, followed by a scenario analysis discussed in Section 5.4. Section 5.5 further illustrates these methods through a multivariate example where PM_{10} and $PM_{2.5}$ are jointly modelled. Most of the material used throughout this section originates from Rodeschini, Fassò, et al. (2024).

The Agrimonia dataset is a medium-sized dataset covering a limited study area. However, the increasing dimensionality of spatio-temporal data renders exact Gaussian field inference computationally prohibitive. To address this issue, Chapter 6 presents the development of a new multivariate

spatio-temporal model, specifically designed for large spatial datasets and to tackle the "Big n problem" which has emerged as a central challenge in modern data analysis. Section 6.1 details the big-n problem for the state-space model and highlights the computational bottleneck. Section 6.2 details a range of strategies that have been developed to address the high computational cost associated with large datasets. Section 6.3 formalises the framework underlying our approach and presents the equations defining the proposed Low-Rank State-Space Model (LR-SSM), tailored for large-scale multivariate spatio-temporal datasets. We also provide theoretical results and examine how the low-rank approximation affects inference. In Section 6.5, we detail the ML estimation procedure, with a focus on the low-rank representation derived via the GMRF approach. Section 6.6 presents numerical studies illustrating the performance of the method and confirming the theoretical findings. In Section 6.7, we apply the methodology to bivariate air quality data. Most of the material used throughout this chapter is drawn from Rodeschini, Tedesco, et al. (2025).

Chapter 7, the conclusion chapter, highlights the main contributions of this thesis and outlines potential directions for future work.

Chapter 2

Fundamentals of continuous-space and space-time processes

As discussed in Chapter 1 and illustrated in the motivating example in Section 1.1, environmental time series often exhibit dependencies that extend beyond the temporal dimension. In particular, spatial dependence arises when nearby locations exert mutual influence, with the strength of this dependence generally decaying with distance. Introducing spatial correlation amounts to assuming that a single underlying process is observed across multiple spatial locations. A key features in spatial statistics is the correlation of data (or equivalently about the covariance); according to the first law of geography, decades later made famous by W.Tobler: "*everything is related to everything else, but near things are more related than distant things*", meaning that, observations at locations in close spatial proximity often tend to be more similar (equivalently high correlated or with high covariance) than observations at locations far apart.

Spatial statistics is the scientific discipline of statistical modelling and analysis of spatially structured phenomena. The goal is to obtain accurate predictions and a reasonable assessment of its associated uncertainty at an unobserved location. Important works by Krige (1951), Matérn (1960) and Matheron (1962) laid the ground for the field of *geostatistics*, proposing some of the earliest methods for modelling spatial dependence that remain fundamental in modern spatial data analysis. While spatial statistics is often introduced in the context of two-dimensional Euclidean space $\mathbb{R}^2$, most theory and methods are valid for any subspace $\mathcal{D} \subset \mathbb{R}^d$ equipped with some notion of norm with $d = 1, 2, 3, \ldots$, leaving computational considerations aside.

Considering the type of the domain $\mathcal{D}$, some possible distinctions are made. When $\mathcal{D}$ is a fixed continuous subset of $\mathbb{R}^d$, we talk about a continuous-space process or a *geostatistical* process, i.e. the index $\mathbf{s}$ varies continuously through the region $\mathcal{D}$. In case $\mathcal{D}$ is a fixed collection of countable

points of $\mathbb{R}^d$ it originates a discrete-space process or *lattice* process. In this chapter, we focus on the continuous case, while the discrete case is presented in Chapter 3.

An important class of covariance functions suitable to model spatial correlation was introduced by Matérn (1960). Thanks to its flexibility and asymptotic properties, the Matérn covariance function is widely used to model spatial dependence (Stein, 1999). Accounting for the spatial covariance is essential to obtain accurate inference on spatial prediction. The use of the word "kriging" in spatial statistics has come to be synonymous with "optimal prediction" in space, using observations taken at known nearby locations. Kriging was originally a linear predictor, and it is in this context that it will be used here, although methods of optimal nonlinear spatial prediction have become part of the "kriging family".

This chapter introduces notation and basic concepts to deal with spatial and temporal dependence. We present statistical models for fixed-domain settings, where the set of locations (e.g., spatial coordinates and/or time instants) $\mathcal{S}$ within the domain $\mathcal{D}$ is known and fixed. The rest of the chapter is organised as follows: we start by presenting the univariate spatial stochastic processes (or simply spatial processes) along with some notions of stationarity in Section 2.1. In Section 2.2, we will extend the univariate case to the multivariate case by considering multivariate processes which allow us to model multiple processes jointly. Temporal correlations are discussed in Section 2.3. Finally, in Section 2.4 we introduce such a model for spatial data, which has a very simple structure that is flexible enough to handle a large class of problems, including multivariate and spatio-temporal correlations.

2.1 Univariate continuous-space processes

Definition 1. *(Stochastic process) A stochastic process (or random field or stochastic field) is defined as a collection of random variables* $\{z(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$. *The distribution of the stochastic process* $z(\mathbf{s})$ *is defined by its finite-dimensional joint (multivariate) distributions*

$$\mathrm{F}_{z(\cdot)}(y_1, \dots, y_n; \mathbf{s}_1, \dots, \mathbf{s}_n) = \mathrm{P}(z(\mathbf{s}_1) \leq y_1, \dots, z(\mathbf{s}_n) \leq y_n)$$

for every finite n and for every finite collection $\{\boldsymbol{s}_1, \dots, \boldsymbol{s}_n\}$.

To emphasise the source of randomness, the stochastic process is sometimes written as $\{z(\boldsymbol{s}; \omega) : \boldsymbol{s} \in \mathcal{D}; \omega \in \Omega\}$ where $(\Omega, \mathcal{F}, \mathrm{P})$ is the a probability space; at every location $\mathbf{s} \in \mathcal{D}$, the mapping function $z(\mathbf{s}, \omega) : \mathcal{D} \to \mathbb{R}$ is a random variable indexed by $\boldsymbol{s}$ where the event ω lies in the sample space Ω. In other words, the ω is the outcome of the process, or the response variable.

An important example of a stochastic process is the Gaussian process (GP; Williams and Rasmussen (2006)), see the Definition 2. Due to its flexibility and since the joint distribution is a multivariate Normal and thus characterised entirely and independently by its mean and covariance terms, the GP is extensively used for solving regression problems in many fields, such as spatial and spatio-temporal statistics.

Definition 2. *(Gaussian process) A univariate stochastic field* $\{z(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ *is said Gaussian process if its finite-dimensional joint distribution is defined as a multivariate Gaussian distribution, that is* $\mathrm{F}_{z(\cdot)}(y_1, \ldots, y_n; \mathbf{s}_1, \ldots, \mathbf{s}_n) \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *for every finite collection* $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ *and where* $\boldsymbol{\mu}$ *is the mean and* $\boldsymbol{\Sigma}$ *is a valid covariance matrix.*

Under suitable conditions that we will present hereafter, the multivariate distribution of $\{z(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ reflects the spatial structure of the process at any other set of locations. The ultimate objective is therefore to draw inferences on this general structure by exploiting the information contained in the realisation of the process that generated the observed data. For this purpose, we characterise the process through the notion of stationarity, namely strict, weak (or second-order), and intrinsic stationarity. The stationarity property is quite important since it allows us to generalise information gathered from the finite set $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\} \subset \mathcal{D}$ to any other set $\{\boldsymbol{s}_1^*, \ldots, s_m^*\} \subseteq \mathcal{D}$.

Definition 3. *(Strictly stationary) A stochastic field* $\{z(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ *is called strictly stationary if for any vector* $\boldsymbol{h} \in \mathcal{D}$ *and for every collection* $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset \mathcal{D}$ *the finite-dimensional joint distribution is such that*

$$\mathrm{F}_{z(\cdot)}(y_1, \ldots, y_n; \boldsymbol{s}_1 + \boldsymbol{h}, \ldots, \mathbf{s}_n + \boldsymbol{h}) = \mathrm{F}_{z(\cdot)}(y_1, \ldots, y_n; \mathbf{s}_1, \ldots, \mathbf{s}_n).$$

In practice, for empirical applications, it is usually sufficient for the process to be stationary in some weaker sense. Strict stationarity implies that the finite-dimensional joint distribution of the process is invariant under shifts in location, which, when the mean and covariance functions exist, leads to shift-invariant mean and covariance functions. Suppose that $\mu(\boldsymbol{s}) = \mathrm{E}[z(\boldsymbol{s})]$ exist for all $\boldsymbol{s} \in \mathcal{D}$ and the $\mathrm{V}[z(\boldsymbol{s})]$ is finite $\forall s \in D$; under these conditions, it is possible define the weak stationarity and intrinsic stationarity process.

Definition 4. *(Weakly stationary) A stochastic field* $\{z(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ *is called weakly stationary if for any vector* $\boldsymbol{h} \in \mathcal{D}$ *and any locations* $\mathbf{s}, \mathbf{s}' \in \mathcal{D}$ *holds:*

$$\mathrm{E}[z(\mathbf{s} + \boldsymbol{h})] = \mathrm{E}[z(\mathbf{s})] = \mu, \quad \textit{and} \quad \mathrm{Cov}(z(\mathbf{s} + \boldsymbol{h}), z(\mathbf{s}' + \boldsymbol{h})) = c(\mathbf{s} - \mathbf{s}'),$$

where $c(\cdot) : \mathcal{D} \to \mathbb{R}$ *is a positive-definite function called the covariance function.*

Considering the weakly stationary, in Definition 4, the second property implies that the covariance function depends only on the difference between $\boldsymbol{s}$ and $\boldsymbol{s}'$. It follows that a strictly stationary process with finite variance is also weakly stationary. It is not true the vice versa, since a weakly stationary field is in general not strictly stationary.

Definition 5. *(Positive definite function) A function $c(\cdot) : \mathcal{D} \to \mathbb{R}$ is positive definite if and only if for any finite set of points $\{\mathbf{s}_1, \cdots, \mathbf{s}_n\} \in \mathcal{D}$ the covariance matrix, said $\boldsymbol{\Sigma} = [c(\boldsymbol{s}_i - \boldsymbol{s}_j)]_{i,j=1}^n$, shall verify Kolmogorov consistency, thus be a non-negative definite matrix.*

When the covariance between two points depends only on their distance, i.e., $\mathrm{Cov}[z(\mathbf{s}), z(\mathbf{s}')] = c(\|\mathbf{s} - \mathbf{s}'\|)$, the process is said to be isotropic, since the covariance does not depend on direction. Otherwise, the process is anisotropic. While isotropy or even stationarity may not always be realistic, particularly over large spatial domains, they often serve as reasonable working assumptions and provide a foundation for more sophisticated anisotropic and non-stationary models. Covariance functions are often parametrised by the set $\vartheta \in \Theta$, denoted as $c(\cdot; \vartheta)$, which defines a family of covariance functions. The parameter set ϑ is generally unknown in practice, and it is typically estimated from the observed data.

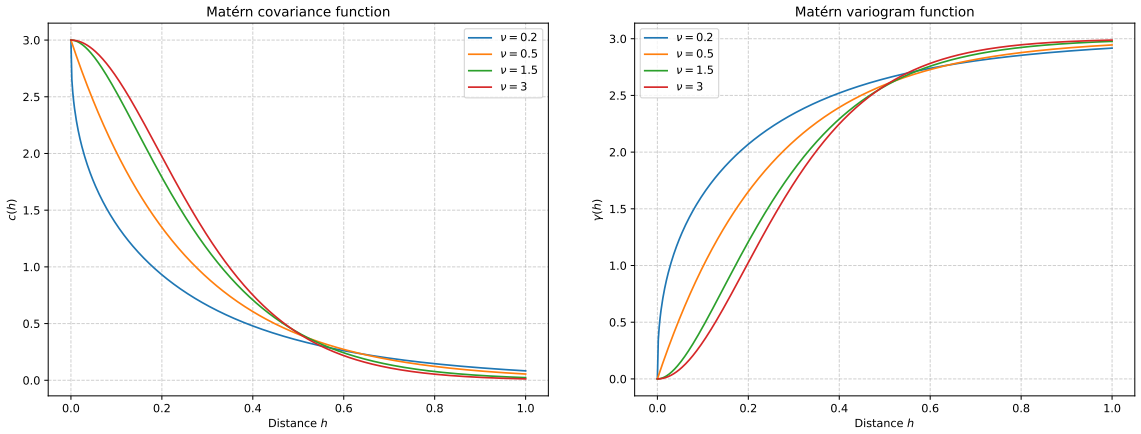
Building on the concept of isotropic covariance functions, one of the most widely used families is the Matérn class of positive definite functions (Matérn, 1960), which has become the standard choice for univariate isotropic fields (Guttorp and Gneiting, 2006; Stein, 1999). A key feature of the Matérn covariance is the inclusion of a smoothness parameter, which directly controls the correlation at short distances and determines the differentiability of the process at the origin.

Definition 6. *(Matérn covariance function) Let $K_\nu(\cdot)$ be the modified Bessel function of the second kind of order ν. The Matérn class of covariance function, parametrised by $\vartheta = \{\nu, \sigma^2, k\}$ is then given by*

$$c(\boldsymbol{h}; \vartheta) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (k\|\boldsymbol{h}\|)^\nu K_\nu(k\|\boldsymbol{h}\|) \tag{2.1}$$

where $\nu > 0$ is the smoothness parameter, $k > 0$ is a length scale parameter that controls the rate of decay of correlation at larger distances, and $\sigma^2 > 0$ is the marginal variance of the field.

The popularity in large part is due to the work of Stein (1999), who showed that the behaviour of the covariance function near the origin has fundamental implications on predictive distributions, particularly predictive uncertainty. A process having the Matérn covariance function, given in Eq. (2.1), is $\lceil \nu \rceil - 1$ times mean square differentiable, where $\lceil \nu \rceil$ is the largest integer less than or equal to ν. The smoothness of the field increases with ν, as shown in Figure 4. Other classes of covariance



(a) (b)

Figure 4: (left) Matérn covariance and (right) variogram functions for different values of the smoothness parameter $\nu \in \{0.2, 0.5, 1.5, 3\}$, over the domain $\mathcal{D} = [0, 1]$. $\nu = 0.5$ corresponds to the exponential covariance, while larger values of ν yield smoother sample realisation.

functions do not have such a parameter to yield a preferred mean square differentiability. A more natural interpretation of the scaling parameter k is as a range parameter ρ; the distance where $z(\boldsymbol{s})$ and $z(\boldsymbol{s}')$ are almost independent. Lacking a simple closed-form relationship, an empirically derived definition is $\rho = \sqrt{8\nu}/k$, which corresponds to a correlation of approximately 0.1 at the distance ρ, for all ν. Moreover, the exponential and Gaussian covariance functions are particular cases of this for $\nu = 1/2$ and $\nu \longrightarrow +\infty$, respectively.

Another way to characterise the stochastic process is through its increments, that is, the difference between two random variables of the same stochastic process. For a random field $z(\boldsymbol{s})$, spatial variation can be studied by considering the increments between values at pairs of points, $z(\mathbf{s} + \boldsymbol{h}) - z(\mathbf{s})$ for $\mathbf{s} + \boldsymbol{h}, \mathbf{s} \in \mathcal{D}$. First proposed by Georges Matheron (1963), the variogram $2\gamma(\mathbf{h})$ is defined as

$$2\gamma(\boldsymbol{h}) = \mathrm{V}[z(\mathbf{s} + \boldsymbol{h}) - z(\mathbf{s})]. \tag{2.2}$$

If a process has a constant mean value function (if it exists) and a variogram only depending on $\boldsymbol{h}$, it is called *intrinsically* stationary. Hence, all weakly stationary processes are intrinsically stationary. It is possible to show that for an intrinsically stationary process, the variogram can be deduced from the covariance function by $\gamma(\boldsymbol{h}) = c(\mathbf{0}) - c(\boldsymbol{h})$, as shown in Figure 4b. In general, the reverse is not true since the variogram is not necessarily bounded. The correlation function, $-1 \leq \rho(\boldsymbol{h}) \leq 1$ is given by $\rho(\boldsymbol{h}) = c(\boldsymbol{h})/c(\mathbf{0})$.

If the process is intrinsically stationary with a constant mean, $2\gamma(\boldsymbol{h}) = \mathrm{E}[(z(\mathbf{s}+\boldsymbol{h}) - z(\mathbf{s}))^2]$, thus a natural estimator of the variogram $2\hat{\gamma}(\boldsymbol{h})$ based on the method-of-moments is:

$$2\hat{\gamma}(\boldsymbol{h}) = \frac{1}{|\mathcal{N}(\boldsymbol{h})|} \sum_{(\boldsymbol{s}_i, \boldsymbol{s}_j) \in \mathcal{N}(\boldsymbol{h})} (z(\boldsymbol{s}_i) - z(\boldsymbol{s}_j))^2, \tag{2.3}$$

where $\mathcal{N}(\boldsymbol{h}) = \{(\boldsymbol{s}, \boldsymbol{s}') : \boldsymbol{s} - \boldsymbol{s}' = \boldsymbol{h}\}$ is the set of pair location at distance $\boldsymbol{h}$ and $|\mathcal{N}(\boldsymbol{h})|$ is the cardinality of the set. Although the estimator $\hat{\gamma}(\boldsymbol{h})$ is appealing, it cannot be used for prediction, as it may lead to an invalid variogram and consequently to a negative mean squared prediction error. The idea then is to search in the class of a valid variogram for one that is close to the spatial dependence in the data. Formally, let $\mathcal{P} = \{2\gamma(\cdot, \vartheta); \vartheta \in \Theta\}$ denote a parametric family of valid variograms. Several criteria have been proposed to identify the best element in $\mathcal{P}$ (see (Cressie and Kornak, 2003) for a complete review), among which the weighted least squares approach has become one of the most widely used in spatial statistics (Cressie and Kornak, 2003):

$$\hat{\vartheta} = \arg\min_{\Theta} \Big\{ \sum_{j=1}^{K} |\mathcal{N}(\boldsymbol{h}(j))| \Big(\frac{\hat{\gamma}(\boldsymbol{h}(j))}{\gamma(\boldsymbol{h}(j); \vartheta)} - 1 \Big)^2 \Big\}, \tag{2.4}$$

at some lags $\boldsymbol{h}(1), \ldots, \boldsymbol{h}(K)$. $\hat{\gamma}(\boldsymbol{h}(j))$ is the method-of-moments estimator of $\gamma(\boldsymbol{h}(j))$. Minimising (2.4) is a good approximation to the weighted least squares. The fitting criterion (2.4) is sensible from the viewpoint that the more pairs of observations $|\mathcal{N}(\boldsymbol{h}(j))|$ there are, the more weight the residual at lag $\boldsymbol{h}(j)$ receives in the overall fit. Also, the smaller the value of the theoretical variogram, the more weight the residual receives at that lag, which is an attractive property because it is important to obtain a good fit of the variogram near the origin. Viewed the minimisation Eq. (2.4) as a fitting criterion for weighted nonlinear regression, the estimates $\hat{\vartheta}$ can be obtained from an iterative procedure like the Gauss–Newton algorithm.

2.2 Multivariate continuous-space processes

In recent years, multivariate data indexed by spatial coordinates have become ubiquitous in geophysical, ecological, environmental, and climate sciences. The importance of modelling such datasets has already been discussed in Section 1.1.2. In this section, we therefore extend the definition of a stochastic process to the multivariate setting.

Definition 7. *(Multivariate stochastic process) A multivariate stochastic process is defined as a p-variate random vector* $\boldsymbol{z}(\mathbf{s}) = (z_1(\mathbf{s}), \ldots, z_p(\mathbf{s}))', \mathbf{s} \in \mathcal{D}$ *where* $\{z_i(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ *is the i-th process at location* $\boldsymbol{s}$.

The definition of weak stationarity in the case of a multivariate process is straightforwardly obtained by extending the Definition 4. Indeed, the process $\boldsymbol{z}(\mathbf{s})$ is said to be weak stationary if for

all $\boldsymbol{s}, \boldsymbol{s}' \in \mathcal{D}$ holds $\mathrm{E}[z_i(\mathbf{s})] = \boldsymbol{\mu}_i$ and $\mathrm{Cov}(z_i(\mathbf{s}), z_j(\mathbf{s}')) = c_{ij}(\mathbf{s} - \mathbf{s}')$ where the function $c_{ij}(\cdot)$ is the covariance function (or direct covariance function) if $i = j$ and the cross-covariance for $i \neq j$. Thus, the multivariate covariance function is $\boldsymbol{C}(\mathbf{s}, \mathbf{s}') = [c_{ij}(\mathbf{s} - \mathbf{s}')]_{i,j=1}^{p}$, where $c_{ij}(\cdot) : \mathcal{D} \to \mathbb{R}$ is a valid covariance function and needs to be described to fully specify the multivariate random field. It is isotropic if $\mathrm{Cov}(z_i(\mathbf{s}), z_j(\mathbf{s}')) = c_{ij}(\|\mathbf{s} - \mathbf{s}'\|)$. Otherwise, it is said to be anisotropic. A summary of the properties of cross-covariance functions can be found in Genton and Kleiber (2015) and Wackernagel (2003).

As well as for the covariance, the variogram can be extended to multivariate random fields in two different ways: a covariance-based cross-variogram (Myers, 1992) defined as

$$2\gamma_{ij}(\boldsymbol{h}) = \mathrm{Cov}[z_i(\boldsymbol{s} + \boldsymbol{h}) - z_i(\boldsymbol{s}), z_j(\boldsymbol{s} + \boldsymbol{h}) - z_j(\boldsymbol{s})], \quad \boldsymbol{s} + \boldsymbol{h}, \boldsymbol{s} \in \mathcal{D}, \tag{2.5}$$

and a variance-based cross-variogram (or pseudo cross-variogram) (Myers, 1991),

$$2\gamma_{ij}(\boldsymbol{h}) = \mathrm{Var}[z_i(\boldsymbol{s} + \boldsymbol{h}) - z_j(\boldsymbol{s})], \quad \boldsymbol{s} + \boldsymbol{h}, \boldsymbol{s}' \in \mathcal{D}. \tag{2.6}$$

That is, the quantity in (2.5) corresponds to the covariance of direct increments, while the quantity in (2.6) represents the variance of cross-increments. Cressie and Wikle (1998) reviewed the differences between the two formulations and argued that (2.6) is preferable, as it yields the same optimal predictor as the one obtained with the cross-covariance function c_{ij}. In contrast, the pseudo cross-variogram (2.5) suffers from the drawback that it cannot be estimated when the variables are not observed at the same spatial locations. A different viewpoint is provided in Wackernagel (2003). In fact, because the pseudo cross-variogram (2.5) only takes positive values, it cannot adequately model negative correlations.

Although the variogram is usually used in the univariate spatial process, in the multivariate context, it is more natural to work with the direct and cross-covariance c_{ij}. The key difficulty arises in specifying the cross-covariance function, that is, the function responsible for the relationship between distinct processes. Indeed, these cross-covariance functions must be chosen to be consistent with direct covariance functions in such a way that yields a non-negative definite covariance matrix. Because of this constraint, it is a difficult problem to specify a valid multivariate covariogram that is not too complex to be estimated and yet capable of modelling a wide range of spatial correlations.

An interesting property of the multivariate cross-covariance matrix is that of separability, given in Definition 8.

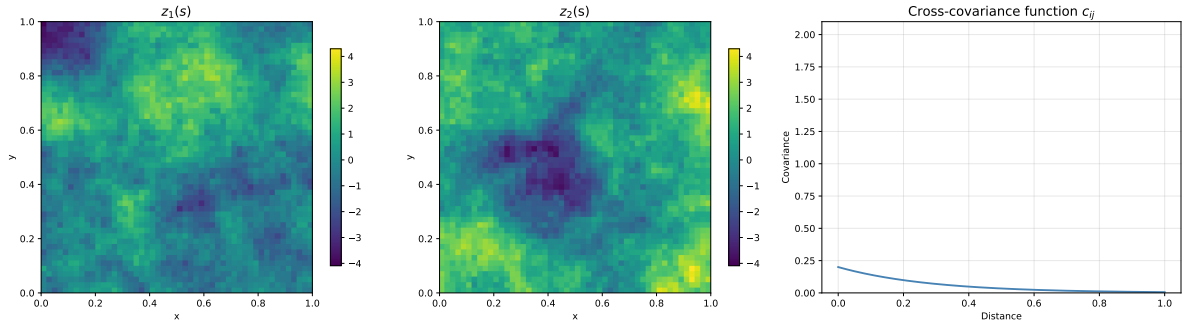


Figure 5: An example of the bivariate spatial process $\boldsymbol{z}(\boldsymbol{s}) = (z_1(\boldsymbol{s}), z_2(\boldsymbol{s}))$ defined on $\mathcal{D} = [0,1]^2$, with an intrinsic multivariate covariance model $C(\boldsymbol{h}) = \boldsymbol{V}\rho(\boldsymbol{h};\vartheta)$, where $v_{ii} = 1$ and $v_{ij} = 0.2$ for $i, j = 1, 2$. The simulated fields z_1 and z_2 are shown in the left and centre panels, respectively. The right panel displays the stationary Matérn correlation function $\rho(\boldsymbol{h};\vartheta)$, parametrized by $\vartheta = \sigma^2 = 2, \nu = 0.5, k = 0.2$.

Definition 8. *(Separable cross-covariance matrix) A cross-covariance matrix* $\boldsymbol{C}(\boldsymbol{s},\boldsymbol{s}') = [c_{ij}(\boldsymbol{s},\boldsymbol{s}')]_{i,j=1}^{p}$ *function is separable for* $\boldsymbol{s},\boldsymbol{s}' \in \mathcal{D}$ *if holds*

$$c_{ij}(\boldsymbol{s},\boldsymbol{s}') = v_{ij}\rho(\boldsymbol{s},\boldsymbol{s}'), \tag{2.7}$$

for all $i, j = 1, \dots, p$*, where* $\rho(\cdot) : \mathcal{D} \to [-1,1]$ *is a valid correlation function and* $\boldsymbol{V} = (v_{ij})_{i,j=1}^{p}$ *is the (non-spatial) covariance between the process* i *and* j*.*

An important application of the separability properties for a weakly stationary process is provided by the *intrinsic* correlation function. In this context, the multivariate correlation structure of a set of variables is independent of the spatial correlation. Under this assumption, the simplest multivariate covariance model that can be adopted for a $\boldsymbol{C}(\boldsymbol{h})$ consists in describing the relations between variables by the variance-covariance matrix $\boldsymbol{V}$ (sometimes called coregionalisation matrix) and the relations between points in space by a spatial correlation function $\rho(\boldsymbol{h}) : \mathcal{D} \to [-1,1]$, such that $\boldsymbol{C}(\boldsymbol{h}) = \boldsymbol{V}\rho(\boldsymbol{h})$. The intrinsic correlation model is sometimes also referred to as the coregionalisation model (Wackernagel, 2003) or the proportional model (Chiles and Delfiner, 2012). It is termed intrinsic because the correlation ρ_{ij} between two variables does not depend upon spatial scale; they are intrinsically correlated by

$$\frac{\sigma_{ij}\rho(\boldsymbol{h})}{\sqrt{\sigma_{ii}\rho(\boldsymbol{h})\sigma_{jj}\rho(\boldsymbol{h})}} = \frac{\sigma_{ij}}{\sigma_{ii}\sigma_{jj}} = \rho_{ij}.$$

Examples of bivariate processes simulated from the intrinsic model on $\mathcal{D} = [0,1]^2$ are shown in Figures 5 and 6. Both use a coregionalisation matrix $\boldsymbol{V}$ and a Matérn correlation function $\rho(\boldsymbol{h};\vartheta)$. In Figure 5, $\boldsymbol{V}$ has $v_{ii} = 1$ and $v_{ij} = 0.2$, while in Figure 6, $\boldsymbol{V}$ has $v_{ii} = 1$ and $v_{ij} = 1.9$. These two scenarios illustrate how different levels of cross-correlation between variables affect the joint spatial behaviour of the process.

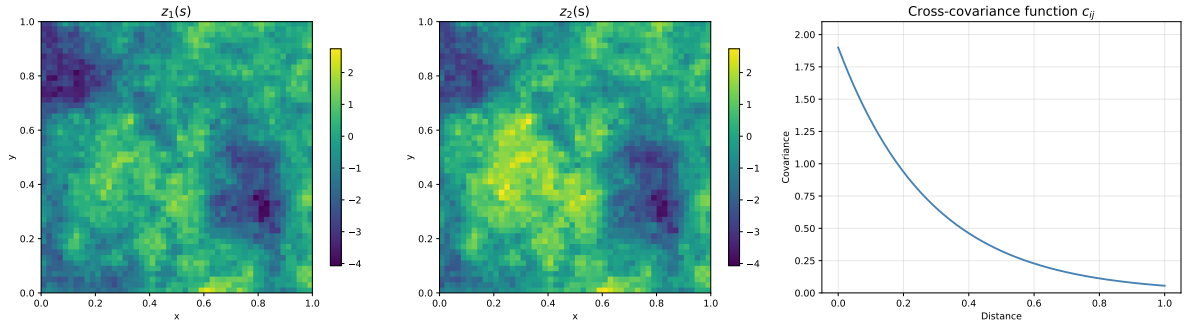


Figure 6: An example of the bivariate spatial process $\boldsymbol{z}(\boldsymbol{s}) = (z_1(\boldsymbol{s}), z_2(\boldsymbol{s}))$ defined on $\mathcal{D} = [0,1]^2$, with an intrinsic multivariate covariance model $C(\boldsymbol{h}) = \boldsymbol{V}, \rho(\boldsymbol{h};\vartheta)$, where $v_{ii} = 1$ and $v_{ij} = 1.9$ for $i, j = 1, 2$. The simulated fields z_1 and z_2 are shown in the left and centre panels, respectively. The right panel displays the stationary Matérn correlation function $\rho(\boldsymbol{h};\vartheta)$, parametrized by $\vartheta = \sigma^2 = 2, \nu = 0.5, k = 0.2$.

The intrinsic correlation model implies an even cross-covariance function because $\rho(\boldsymbol{h})$ is an even function and it is a normalised direct covariance function. The empirical estimator of the cross-covariance matrix function, based on the method of moments, extends naturally to the multivariate process covariogram $c_{ij}(\boldsymbol{h})$.

$$\hat{c}_{ij}(\boldsymbol{h}) = \frac{1}{|\mathcal{N}(\boldsymbol{h})|} \sum_{(\boldsymbol{s},\boldsymbol{s}')\in\mathcal{N}(\boldsymbol{h})} (z_i(\boldsymbol{s}) - z_i(\boldsymbol{s}'))(z_j(\boldsymbol{s}) - z_j(\boldsymbol{s}')). \tag{2.8}$$

where $\mathcal{N}(\boldsymbol{h}) = \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : \boldsymbol{s}_i - \boldsymbol{s}_j = \boldsymbol{h}\}$. Then a valid parametric model is typically fit by least squares methods to the empirical estimates in (2.8). An example to estimate the $\boldsymbol{V}$ of the intrinsic correlation model is proposed by Goulard and Voltz (1992) while the parameters of the correlation function $\rho(\cdot;\vartheta)$ are supposed known. Goulard and Voltz (1992) proposes an iterative algorithm: first calls for the empirical non-parametric estimation of the covariance $\hat{c}_{ij}(\cdot)$ at some lags $\{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N\}$. Then, for each (i, j) and lag $\mathbf{h}_k$, minimises through as iterative procedure the quantity

$$\hat{\vartheta}_{\text{OLS}} = \arg\min_{\theta} \sum_{i,j,k} \|\hat{c}_{ij}(\mathbf{h}_k) - c_{ij}(\mathbf{h}_k;\vartheta)\|^2,$$

where $c_{ij}(\mathbf{h}_k;\vartheta)$ is a valid parametric model for $\boldsymbol{c}_{ij}(\boldsymbol{h}_k)$ and $\|\cdot\| : \mathbb{R}^{p\times p} \to \mathbb{R}^+$ is any valid norm in the vector space of matrices $p \times p$. A useful norm is the $\|\cdot\|_2^2 = \text{tr}(\cdot)$ known as the Frobenius norm, which is an extension of the Euclidean norm and comes from the Frobenius inner product on the space of all matrices. The minimisation is subject to the constraint that the $\boldsymbol{V}$ is positive semi-definite. Note that this algorithm does not apply to the completely heterotopic case since the empirical cross-covariograms defined in Eq. (2.8) cannot be computed in this case. Other empirical estimators and the associated properties of the cross-variogram (2.5) and (2.6) are discussed in Papritz, Künsch, and Webster (1993), while robust estimators and asymptotic properties of the estimator are discussed in Lark (2003) and Furrer (2005), respectively.

The intrinsic correlation model can be used as a building block for the nested multivariate covariogram in the so-called linear coregionalisation model (LCM; Wackernagel (2003)). The LCM consists of representing the multivariate random field as a linear combination of $K > 0$ independent p-variate stationary processes, said $\boldsymbol{w}_k$ for $k = 0, \dots, K$, with mean 0 and multivariate (separable) covariogram $\boldsymbol{V}_k \rho_k(\boldsymbol{h}, \vartheta_k)$ where $\rho_k(\boldsymbol{s}; \vartheta_k)$ is a valid correlation function and $\boldsymbol{V}_k$ is a $p \times p$ positive definite matrix which models the (non-spatial) covariance between the process i and j. In addition, the $\boldsymbol{w}_k$ processes are spatially uncorrelated in the sense that for any $k \neq j$, $\text{Cov}[\boldsymbol{w}_k(\boldsymbol{s}), \boldsymbol{w}_j(\boldsymbol{s}')] = 0$ for all $\boldsymbol{s}, \boldsymbol{s}' \in \mathcal{D}$. The LCM is defined as a linear mixture, that is:

$$\boldsymbol{z}(\boldsymbol{s}) = \boldsymbol{\mu}(\boldsymbol{s}) + \sum_{k=0}^{K} \boldsymbol{w}_k(\boldsymbol{s}). \tag{2.9}$$

Given the assumption of Eq. (2.9), the resulting multivariate covariance functions take the form:

$$\boldsymbol{C}(\boldsymbol{h}) = \sum_{k=0}^{K} \boldsymbol{V}_k \rho_k(\boldsymbol{h}; \vartheta_k). \tag{2.10}$$

The idea behind the LCM is that several coregionalisation matrices $\boldsymbol{V}_k$ can describe the correlation structure among variables at different spatial scales of the phenomena. Usually the $\rho_i(\boldsymbol{s}; \vartheta_i) = \rho(\boldsymbol{s}; \vartheta_i)$.

With a large number of processes, the number of parameters can quickly become unwieldy and the resulting estimation difficult. H. Zhang (2007) described MLE of the spatial LCM based on an EM algorithm. Importantly, H. Zhang (2007) show that the estimates in each iteration automatically satisfy the necessary constraints for a valid multivariate covariogram. A second drawback of the LCM is that the smoothness of any component of the multivariate random field is restricted to that of the roughest underlying univariate process.

Along with the LCM, other possible approaches can be used to build valid multivariate covariance, such as convolution methods, the multivariate Matérn and the non-stationary approach. See Genton and Kleiber (2015) for a complete overview.

2.3 Spatio-temporal processes

Until now, we have considered univariate spatial stochastic processes, i.e., $z(\cdot) : \mathcal{D} \to \mathbb{R}$, as well as multivariate processes, $\mathbf{z}(\cdot) : \mathcal{D} \to \mathbb{R}^d$. In real applications, however, data are not only indexed in space but also contain temporal information, since attributes are often collected at specific time points. Spatio-temporal problems arise in the analysis of phenomena that evolve in both space and time. The aim of this Section is therefore to discuss stochastic processes defined over spatio-temporal domains. We will focus mainly on the univariate spatio-temporal process, as the multivariate case follows naturally.

Definition 9. *(Spatio-temporal stochastic process) A univariate spatio-temporal stochastic process is defined as a collection of random variables* $\{z(\boldsymbol{h}, t) : \boldsymbol{s} \in \mathcal{D}, t \in \mathcal{T} \subset \mathbb{R}^+\}$ *indexed in space and in time. The distribution of the stochastic process* $z(\mathbf{s}, t)$ *is defined by its finite-dimensional joint (multivariate) distribution*

$$\mathrm{F}_{z(\cdot,\cdot)}(y_1, \ldots, y_n; (\mathbf{s}_1, t_1), \ldots, (\mathbf{s}_n, t_n)) = \mathrm{P}(z(\mathbf{s_1}, t_1) \leq y_1, \ldots, z(\mathbf{s}_n, t_n) \leq y_n),$$

for every finite collection $\{(\boldsymbol{s}_1, t_1), \ldots, (\boldsymbol{s}_m, t_T)\} \subset \mathcal{D} \times \mathcal{T}$.

Definition 9 extends naturally to the multivariate spatio-temporal processes. The concepts of strict and weak stationarity extend to the spatio-temporal context as well. In particular, strict stationarity requires the finite-dimensional distributions to be shift-invariant (see Definition 10), whereas weak stationarity requires only the first and second moments to be shift-invariant (see Definition 11).

Definition 10. *(Strictly stationary) A univariate spatio-temporal stochastic process* $\{z(\mathbf{s}, t) : (\mathbf{s}, t) \in \mathcal{D} \times \mathcal{T}\}$ *is called strictly stationary if for any pair* $(\boldsymbol{h}, \delta) \in \mathcal{D} \times \mathcal{T}$ *and for every collection* $\{(\boldsymbol{s}_1, \boldsymbol{t}_1), \ldots, (\boldsymbol{s}_n, \boldsymbol{t}_T)\} \subset \mathcal{D} \times \mathcal{T}$ *the finite-dimensional joint distribution is such that*

$$\mathrm{F}_{z(\cdot,\cdot)}(y_1, \ldots, y_n; (\boldsymbol{s}_1 + \boldsymbol{h}, t_1 + \delta), \ldots, (\mathbf{s}_n + \boldsymbol{h}, t_T + \delta)) = \mathrm{F}_{z(\cdot,\cdot)}(y_1, \ldots, y_n; (\mathbf{s}_1, t_1), \ldots, (\mathbf{s}_n, t_T)).$$

Definition 11. *(Weak stationary) A spatio-temporal stochastic process* $\{z(\mathbf{s}, t) : (\mathbf{s}, t) \in \mathcal{D} \times \mathcal{T}\}$ *is called weak stationary if for any pair* $(\boldsymbol{h}, \delta) \in \mathcal{D} \times \mathcal{T}$

$$\begin{aligned}
&\mathrm{E}[z(\mathbf{s} + \boldsymbol{h}, t + \delta)] = \mathrm{E}[z(\mathbf{s}, t)], \\
&\mathrm{Cov}[z(\mathbf{s} + \boldsymbol{h}, t + \delta), z(\mathbf{s}, t)] = \mathrm{Cov}(z(\mathbf{0}, 0), z(\boldsymbol{h}, \boldsymbol{\delta})) = c(\boldsymbol{h}, \boldsymbol{\delta}),
\end{aligned}$$

where $c(\boldsymbol{h}, \boldsymbol{\delta}) : \mathcal{D} \times \mathcal{T} \to \mathbb{R}$ *is a valid space-time covariance function.*

To guarantee continuous realisation and second-order stationarity, the function $c(\boldsymbol{h}, \boldsymbol{\delta})$ must belong to the class of differentiable functions. Its margins $c(\mathbf{0}, \cdot)$ and $c(\cdot, \mathbf{0})$ are purely temporal, purely spatial and covariance functions, respectively.

Although observations are placed into the space–time $\mathcal{D} \times \mathcal{T}$ domain, there are fundamental differences between the coordinate axes. First of all, time moves only forward, while there is no preferred direction in space (usually). Thus, the specification of the combined space–time domain $\mathcal{D} \times \mathcal{T}$ must acknowledge these differences. A simple way to build valid space-time covariance functions is to use the fact that the class of positive definite functions is closed with respect to the operators of sum and product. Therefore, we can construct valid spatio-temporal covariance functions

simply by taking the sum or the product of a purely spatial and a purely temporal covariance function. These two decompositions are:

$$c(\boldsymbol{s}, \boldsymbol{t}; \vartheta) = c_s(\boldsymbol{s}; \vartheta_s) + c_t(t; \vartheta_t), \tag{2.11}$$

and:

$$c(\boldsymbol{s}, \boldsymbol{t}; \vartheta) = c_s(\boldsymbol{s}; \vartheta_s) c_t(t; \vartheta_t) \tag{2.12}$$

where $\vartheta = \{\vartheta_s, \vartheta_t\}$ and $c_s(\boldsymbol{s}; \vartheta_s)$, $c_t(t; \vartheta_t)$ are the purely spatial and purely temporal covariance functions. Such models arise from two separate processes (one temporal and one spatial) that act independently of each other. Nevertheless, while convenient in terms of computation and interpretability, separable models do not allow for arbitrary interaction of the spatial and temporal dimensions (e.g. different spatial dependence at different temporal lags and vice versa) and are necessarily fully symmetric. These yield simplistic dependence patterns that frequently fail to represent the complex spatio-temporal variations of real phenomena. More sophisticated methods for building valid non-separable space-time covariance are available in the literature (Stein, 2005; Cressie and H.-C. Huang, 1999; Gneiting, 2002). For example, Stein (2005) extended the Matérn family of spatial covariance functions to account for temporal covariance. Gneiting (2002) proved that given a completely monotone function $g(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ and a positive function with completely monotone derivative $\psi(\cdot) : \mathbb{R}^+ \to \mathbb{R}$, then any function of the following form:

$$c(\boldsymbol{s}, t; \vartheta) = \frac{1}{\psi(t^2)^{d/2}} g\left(\frac{\|\boldsymbol{h}\|^2}{\psi(t^2)}\right) \tag{2.13}$$

is a valid space-time covariance function over the product space $(\boldsymbol{s}, t) \in \mathbb{R}^d \times \mathbb{R}$.

However, spatio-temporal processes are often not observed continuously along time, whereas observed into ordered integer-spaced intervals such that $\mathcal{T} = \{0, 1, \ldots, T\}$ where T is the temporal length. In this case, the problem reduces to the analysis of a collection of temporally varying spatial processes $\{z_t(\boldsymbol{s}), s \in \mathcal{D}\}$ for $t \in \mathcal{T}$: meaning different continuous-space process on the same spatial domain at different time steps t (previously referred to as stochastic imaging, since a new map is considered for every t). With an abuse of notation, for simplicity, we will continue to use the notation $z(\boldsymbol{s}, t) = z_t(\boldsymbol{s})$. If time is assumed to not play a direct role in the explanation of the observed phenomenon, each spatial process may be assumed to evolve independently, that is $\mathrm{Cov}(z(\boldsymbol{s}, t_i), z(\boldsymbol{s}, t_j)) = 0$ for all $t_i \neq t_j$, perturbed with independent random errors. Therefore, all the independent realisations can be used to make an inference about the same underlying spatial process $z(\boldsymbol{s})$ characterised only by the spatial covariance function $c_s(\boldsymbol{s}; \vartheta_s)$. Otherwise, the potential temporal dependence between the points belonging to different surfaces at subsequent times must be taken into

account.

Statistics for spatio-temporal data have developed along two main paradigms. The first treats time as an additional dimension alongside the spatial ones, leading to covariance models and analyses that extend to higher-dimensional space, as discussed so far. This approach often provides satisfactory solutions but does not explicitly model the underlying dynamics of the phenomenon. In contrast, the dynamical statistical approach models how the current state depends on previous states through either mechanistic or probabilistic relationships. These models describe the temporal evolution of spatial processes by specifying suitable transition equations in the space of latent features, thereby inducing temporal dependence between residuals at different times. (H.-C. Huang and Cressie, 1996; Calculli et al., 2015; Shumway, Stoffer, and Stoffer, 2000).

One of the best-known examples of dynamical relationships in the purely temporal context is the discrete-time Markov process (or Markov chain since the time is discrete). It describes the sequence of possible events in which the probability of each event depends only on the state attained in the previous event:

Definition 12. *(Discrete-time Markov process) A discrete-time Markov process (or Markov chain) is a ordered sequence of random variables* $(z_1, z_2, \ldots, z_T)$ *with the Markov property, namely that the probability of moving to the next state depends only on the present state and not on the previous states:* $\mathrm{P}(z_t \mid z_{t-1}, z_{t-2}, \ldots, z_0) = \mathrm{P}(z_t \mid z_{t-1})$ *if both conditional probabilities are well defined, that is, if the join probability* $\mathrm{P}_{z_1,\ldots,z_T}(\cdot) > 0$. *The countable set* $\mathcal{S} = \{z_1, z_2, \ldots, z_T\}$ *is called the state space of the chain.*

The Markov property can be formulated more generally by introducing the concepts of graphs, adjacency, and neighbourhood. We will return to this point in Chapter 3. Importantly, the Markov property leads to a factorisation of the likelihood function, which simplifies the optimisation.

2.4 Statistical modelling of multivariate spatio-temporal processes

A statistical model can be seen as a mathematical abstraction of a data-generating mechanism; it may or may not explain why a particular phenomenon occurs. In this work, we focus on the inference problem instead of the causality one. Any statistical inference requires some assumptions, and a statistical model is a set of assumptions concerning the generation of the observed data. For spatio-temporal phenomena, the model is usually a stochastic process (or random field) $\{z(\boldsymbol{s}, t) : \mathcal{D} \times \mathcal{T}\}$ indexed

both in space and time, see Definition 9. For any given ω in the sample Ω, the mapping function $z(\boldsymbol{s}, t) : \mathcal{D} \times \mathcal{T} \to \mathbb{R}$ is called a realisation of the stochastic process. This means that for a fixed ω there exists a function that maps the index set $\mathcal{D} \times \mathcal{T}$ to $\mathbb{R}$, the actual observed data. In the sequel, we assume that the domain $\mathcal{D} \subset \mathbb{R}^d$ with $d \geq 1$ and $\mathcal{T} = \{0, \dots, T\}$ are known and fixed.

Hence, our final objective is to draw inferences on the general structure of the stochastic process, exploiting the information contained in a realisation of the process that produced the observed data. Throughout this work, estimation will refer exclusively to inference on fixed but unknown parameters of the model (e.g., those governing the mean, covariance, or dependence structure), while prediction will denote inference on unobserved random variables. The observed data are therefore used to perform inference on the underlying process, such as predicting a known functional $g(z(\cdot))$. In many geostatistical applications, the primary goal is spatial prediction. A class of linear prediction methods, known as kriging, has been developed for this purpose. Stein (1999) provides a rigorous account of the mathematical theory underlying linear kriging. In this framework, spatial prediction corresponds to the case $g(z(\cdot)) = z(\mathbf{s}_0)$, where $\mathbf{s}_0$ is an unobserved location. The kriging predictor of $z(\mathbf{s}_0)$ is derived as the minimum mean squared error linear predictor (BLUP), i.e. the linear predictor that minimises mean squared prediction error under unbiasedness constraints. Formal definition of BLUP is provided in Definition 13.

Definition 13. *(BLUP) A predictor $\hat{z}(\boldsymbol{s}_0) \in \Theta$ where Θ is the class of the linear predictor is called a Best Linear Unbiased Predictor (BLUP) if it is unbiased and has minimum mean square error among all linear unbiased predictors, i.e.* $\mathrm{E}[\hat{z}(\boldsymbol{s}_0)] = \mathrm{E}[z(\boldsymbol{s}_0)] = z(\boldsymbol{s}_0)$ *and* $\hat{z}(\boldsymbol{s}_0) = \arg\min_{\hat{z}(\boldsymbol{s}_0) \in \Theta} \mathrm{E}[(\hat{z}(\boldsymbol{s}_0) - z(\boldsymbol{s}_0))^2]$.

We have already introduced the Gaussian process (GP), which is a common assumption in the analysis of spatio-temporal processes. Moreover, in the Gaussian case, weak stationarity implies strict stationarity. Therefore, as there is no distinction between the two concepts of stationarity in the Gaussian case, one simply writes that the process is stationary. Furthermore, if $z(\cdot)$ is a GP, then the kriging predictor is the $\mathrm{E}(z(\boldsymbol{s}_0)|\boldsymbol{z})$ where $\boldsymbol{z}$ is the vector of the observed data and under the square error loss, the optimal predictor and the kriging are the same. The cokriging procedure is a natural extension of kriging when a multivariate covariance model and multivariate data are considered.

In what follows, we present a very simple model structure that is flexible enough to handle a large class of problems, including univariate and multivariate spatial processes as well as discrete-time spatio-temporal processes. In particular, we will delve into mixed-effect models, which are extensively used in geostatistics. For extensive details on kriging, including asymptotic theory and

estimation details, see Stein (1999), Cressie and Kornak (2003), and Cressie and Wikle (2015) for kriging and Wackernagel (2003) for the co-kriging.

2.4.1 Univariate spatial modelling

Let $\{z(\boldsymbol{s}) : s \in \mathcal{D}\}$ be a real valued random process (see the Definition 1). We are interested in making inferences on the $z(\boldsymbol{s})$ given a finite number of observations $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))'$. The set of observed locations $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, with $\boldsymbol{s}_i \in \mathcal{D}$ for $i = 1, \ldots, n$, is assumed to be know and fixed as well as the domain $\mathcal{D}$. The measurements $\boldsymbol{y}$ are often sampled under measurement noise, and a statistical model for the data has to take this into account. We can consider the following Spatio Random Effect (SRE) model (Cressie and Johannesson, 2008),

$$y(\boldsymbol{s}) = \mu + z(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \quad \text{with } \boldsymbol{s} \in \mathcal{D}, \tag{2.14}$$

where $\mu \in \mathbb{R}$ is unknown and represents the mean structure, sometimes referred to as trend or large-scale variation. $z(\cdot)$ is sometimes called small-scale variation, and it is assumed to be a zero-mean intrinsically stationary GP with covariance function $c(\boldsymbol{h}) = c(\boldsymbol{h}; \vartheta)$ parametrised by ϑ. $\epsilon(\cdot)$ is an uncorrelated random noise, sometimes referred to as measurement error, with zero mean and marginal variance σ^2_ϵ. $\epsilon(\cdot)$ and $z(\cdot)$ are uncorrelated, that is $\text{Cov}(z(\cdot), \epsilon(\cdot)) = 0$. Under the SRE model (2.14), the process $y(\boldsymbol{s})$ is stationary with a constant mean. One possible generalisation of (2.14) is to assume that the $\text{E}[z(\boldsymbol{s})] = \mu$ is no longer constant but is an unknown linear combination of a set of determinist covariates, i.e $\mu(s) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta}$ where $\boldsymbol{x}(\boldsymbol{s}) : \mathcal{D} \to \mathbb{R}^b$ is a vector of known covariates (which may include an intercept) and $\boldsymbol{\beta}$ is an unknown parameter vector. Although the observed process often exhibits a non-constant mean across space, subtracting the trend yields residuals that can be modelled as a stationary process with a constant mean. In particular, when the covariates include an intercept term, the residual process $y(\boldsymbol{s}) - \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta}$ has zero mean by construction. Building upon these intuitions, the model (2.14) can be expressed as

$$y(\boldsymbol{s}) - \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} = z(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \quad \text{with } \boldsymbol{s} \in \mathcal{D}. \tag{2.15}$$

The parameter set of model (2.15) is $\Pi = \{\boldsymbol{\beta}, \vartheta, \sigma^2_\epsilon\}$. An important quantity is the signal-to-noise ratio (SNR), often defined as the ratio between the marginal variance of the latent GP (which can be considered the variance of the signal) and the noise variance, i.e. $\text{SNR} = \sigma^2/\sigma^2_\epsilon$. A high SNR indicates that the latent spatial signal dominates over observational noise, leading to more confident predictions, whereas a low SNR suggests that noise obscures the spatial structure, making parameter estimation and kriging less reliable.

The SRE model (2.15) defines the process at the continuous domain level $\boldsymbol{s}$. Once observations are available, the model (2.15) must be restricted to the sampled locations. The corresponding SRE data model is then obtained by restricting the process to the observation set $\mathcal{S}$. Let $\boldsymbol{y}$ be the $n \times 1$ vector of the observation, $\boldsymbol{z} = (z(\boldsymbol{s}_1), \dots, z(\boldsymbol{s}_n))'$ be the $n \times 1$ random vector. Since $\boldsymbol{z}(\boldsymbol{s})$ is a GP, then $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = (c(\boldsymbol{s}_i - \boldsymbol{s}_j; \vartheta))_{i,j=1}^n$. Finally, let $\boldsymbol{\epsilon} = (\epsilon(\boldsymbol{s}_1), \dots, \epsilon(\boldsymbol{s}_n))'$ be the corresponding noise vector. The SRE data model associated with the SRE model (2.15) becomes:

$$\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\beta} = \boldsymbol{z} + \boldsymbol{\epsilon} \tag{2.16}$$

The SRE data model (2.16) is sometimes also called the measurement equation since it specifies how the observed data are generated as a function of the latent process, and $\boldsymbol{\epsilon}$ as the measurement error. Under Eq. (2.16), the SRE model is cast into a Gaussian process regression. Figure 7 shows an example of the SRE data model where $\boldsymbol{y} = (y(\boldsymbol{s}_1), \dots, y(\boldsymbol{s}_n))'$ with $n = 100$ and $\mathcal{D} = [0, 100]^2$ and $c(\boldsymbol{h}; \vartheta)$ is the Matérn covariance function parametrised by $\vartheta = \{\sigma^2 = 1, \nu = 1, k = 40\}$.

In the SRE model, the parameter set Π is fixed but unknown. In a frequentist framework, inference on Π can be performed using the observed data through the maximum likelihood (ML), which selects the parameter values that maximise the observed likelihood. Alternative approaches also exist for inference on Π. In particular, Bayesian hierarchical frameworks have proven to be highly flexible and efficient (Banerjee et al., 2008; Datta et al., 2016), allowing the incorporation of prior information, modelling of spatial structures, and natural quantification of uncertainty in both parameter estimates and predictions. These methods are especially useful when the process exhibits multilevel or nested spatial-temporal dependencies. In this work, however, we focus on inference through ML framework. If the $\vartheta, \sigma_\epsilon^2$ are known, the ML estimation (MLE) of β reduces to the generalised least squares (GLS) estimator as detailed in Example 1. The MLE for all the parameters in Π, based on the data model (2.16) is illustrated in Example 2, which serves as a fundamental building block in geostatistics.

Example 1. *(Generalise least square SRE estimation) As an example, we illustrate here the GLS method to estimate the SRE model. Assume that the dataset* $\{(y_i, \boldsymbol{s}_i)\}_{i=1}^n$ *satisfies the SRE data model* (2.16)*. Let* $\boldsymbol{X}' = \mathrm{stack}(x(\boldsymbol{s}_1)', \dots, x(\boldsymbol{s}_n)')$ *be the* $n \times b$ *design matrix of the observed covariates. Assume that the random vector* $\boldsymbol{z}$ *is with* $\mathrm{E}[\boldsymbol{z}|\boldsymbol{X}] = 0$ *and* $\mathrm{Cov}[\boldsymbol{z}|\boldsymbol{X}] = \boldsymbol{\Sigma}(\vartheta)$ *where* $\boldsymbol{\Sigma}(\vartheta) = (c(\boldsymbol{s}_i - \boldsymbol{s}_j; \vartheta))_{i,j=1}^n$ *where* ϑ *is know. Assume that the measurement errors* $\boldsymbol{\epsilon}$ *satisfy* $\mathrm{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = 0$ *and* $\mathrm{Cov}[\epsilon|\boldsymbol{X}] = \sigma_\epsilon^2 I_n$*. The observed vector* $\boldsymbol{y} = (y(\boldsymbol{s}_1), \dots, y(\boldsymbol{s}_n))'$ *then satisfies the SRE data model* (2.16)*, with* $\mathrm{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}'\beta$ *where* $\boldsymbol{\beta} \in \mathbb{R}^b$ *is a vector of unknown parameters. The GLS method estimates* $\boldsymbol{\beta}$ *by minimising the squared Mahalanobis distance of the residual vector* $\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\beta}$*.*

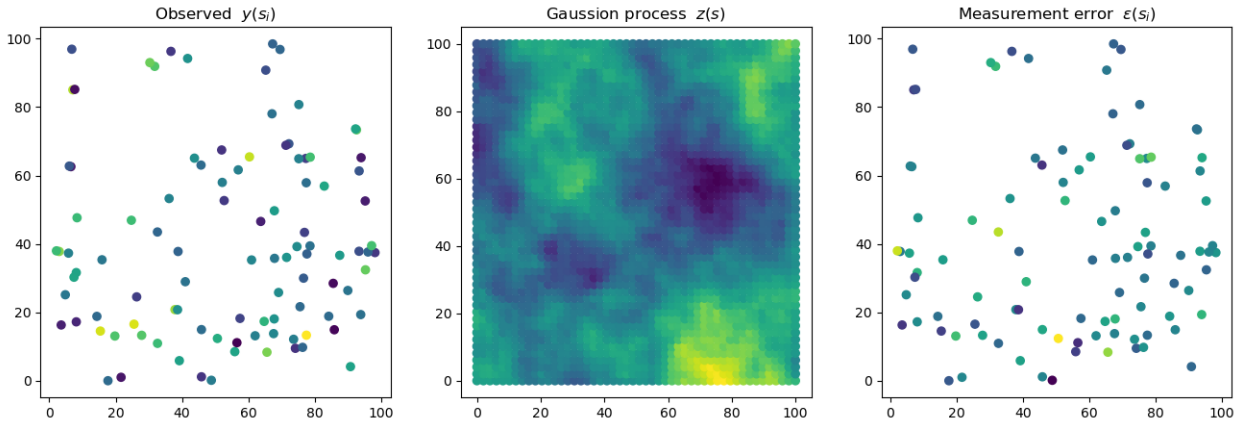


Figure 7: An example of the SRE data model (2.16). Left: the observed data $(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))$ with $n = 100$ and $\mathcal{D} = [0, 100]^2$. Centre: the GP $z(\boldsymbol{s})$ with $\boldsymbol{s} \in \mathcal{D}$ and $c(\boldsymbol{h}; \vartheta)$ is the Matérn covariance function parametrised by $\vartheta = \{\sigma^2 = 1, \nu = 1, k = 40\}$. Right: and the measurement error $(\epsilon(\boldsymbol{s}_1), \ldots, \epsilon(\boldsymbol{s}_n))$.

The resulting estimator is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'(\boldsymbol{\Sigma}(\vartheta) + \sigma_\epsilon^2 I_n)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}(\vartheta) + \sigma_\epsilon^2 I_n)^{-1}\boldsymbol{y}. \tag{2.17}$$

Example 2. *(Maximum-likelihood SRE estimation) Let us assume that the dataset* $\{(y_i, \boldsymbol{s}_i)\}_{i=1}^n$ *satisfy the SRE data model* (2.16). *Levering the definition of GP,* $\boldsymbol{z} \sim \mathrm{N}(0, \boldsymbol{\Sigma}(\vartheta))$ *where* $\boldsymbol{\Sigma}(\vartheta) = [c(\boldsymbol{s}_i - \boldsymbol{s}_j; \vartheta)]_{i,j=1}^n$ *where* $c(\cdot; \vartheta)$ *is a covariance function. The measurement errors are IID and* $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma_\epsilon^2 I_n)$. *Thus, the observed vector* $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}'\boldsymbol{\beta}, \boldsymbol{\Sigma} + \sigma_\epsilon^2 I_n)$. *The parameter set* $\Pi = \{\boldsymbol{\beta}, \vartheta, \sigma_\epsilon^2\}$ *can be estimated by maximising the observed log-likelihood* $\mathrm{logL}(\mathbf{y} \mid \boldsymbol{X}; \Pi)$ *given by*

$$\begin{gathered} 2\,\mathrm{logL}(\mathbf{y} \mid \boldsymbol{X}; \Pi) - n\log(2\pi) - \log|\boldsymbol{\Sigma}(\vartheta) + \sigma_\epsilon^2 I_n| - (\mathbf{y} - \boldsymbol{X}'\boldsymbol{\beta})'(\boldsymbol{\Sigma}(\vartheta) + \sigma_\epsilon^2 I_n)^{-1}(\mathbf{y} - \boldsymbol{X}'\boldsymbol{\beta}), \\ \hat{\Pi}_{MLE} = \arg\max\big(2\,\mathrm{logL}(\mathbf{y} \mid \boldsymbol{X}; \Pi)\big). \end{gathered} \tag{2.18}$$

Several methods are available for this type of optimisation problem, among which gradient descent and the Newton–Raphson method are commonly employed. Initial values for ϑ^0 *can be obtained from the weighted least squares fit of the variogram* (2.4)*, while* β^0 *can be estimated using the GLS. Sometimes the log-likelihood is also written as* $\mathrm{logL}(\mathbf{y} \mid \boldsymbol{X}; \Pi) = \mathrm{logL}_\Pi(\mathbf{y} \mid \boldsymbol{X})$ *to emphasise the dependence on the distributional assumptions specified by* Π.

Theoretical discussions on MLE of data model (2.16) are deeply investigated in H. Zhang (2004), Stein (1999), and Diggle, Tawn, and Moyeed (1998). In spatial statistics, two types of asymptotics are typically considered. Under *increasing-domain* asymptotics, more data are collected by enlarging the spatial domain $\mathcal{D}$, while under *fixed-domain* (or *infill*) asymptotics, more data are obtained by sampling more densely within a bounded region. The latter setting is particularly relevant in geostatistics, where dense measurements are often collected over a fixed study area. More formally, infill

asymptotics means considering the distribution of the ML estimator under an increasing sample size, i.e., let $\mathcal{S}_n$ for $n = 1, 2, \dots$ be an increasing finite subset $\mathcal{S}_n \subset \mathcal{D}$ such that $\cup_{n=1}^{\infty} \mathcal{S}_n$ is bounded and infinite. In this setting, the ML asymptotics describe how the maximum likelihood estimator behaves as $n \to \infty$, typically in terms of consistency and asymptotic normality. Simulation studies are useful to illustrate how fixed-domain asymptotics translate into finite-sample behaviour.

(Mardia and Marshall, 1984) shows that both the variance σ^2 and the scale parameter $\theta = 1/k$ in the exponential covariogram can be estimated consistently under the increasing domain asymptotics. H. Zhang (2004) used properties of equivalence of probability measures to show that not all parameters in a spatial SRE are consistently estimable under infill asymptotics regardless of the estimation methods used. In particular, considering the Matérn covariance function parametrised by $\boldsymbol{\vartheta} = \{\nu = 1/2, \sigma^2, 1/\theta\}$, they show that the distribution of the estimator for the ratio σ^2/θ becomes more symmetric and exhibits smaller variance as the sample size increases, especially when spatial correlation is strong. This improvement contrasts with the estimators of θ and σ^2 individually, whose variances are less sensitive to sample size increases. Hence, while dense sampling in a fixed domain does not substantially improve estimates of θ and σ^2, it markedly improves the estimation of their ratio σ^2/θ. Simulation results show that MLE of σ^2/θ not only display reduced variance and more symmetric distributions but also retain negligible bias even for moderately large sample sizes. By contrast, the asymptotic inconsistency of θ and σ^2 means that their variances may not vanish as the sample size grows, despite the small bias.

Importantly, Kriging predictions are largely unaffected by these inconsistencies: different covariance models can yield very similar predicted values, but their prediction variances differ more substantially depending on the ratio σ^2/θ. This highlights why, in practice, fixed-domain asymptotics reinforce the central role of σ^2/θ in spatial prediction.

In geostatistics, interpolation is achieved through kriging, a method that provides the BLUP of a spatial process at unobserved locations. It combines observed data values with a model for the spatial covariance structure, ensuring predictions are unbiased and have minimal variance among all linear estimators.

Kriging for GP is illustrated in Example 3. However, obtaining the kriging predictor requires inverting the $n \times n$ covariance matrix, where n is the total number of observations, an operation with computational cost $O(n^3)$ and memory demand $O(n^2)$, which quickly becomes prohibitive as n increases. The implications of this computational bottleneck have already been highlighted in Section 1.1.2, and strategies to overcome it in the context of large datasets will be discussed in Chapter 6.

Example 3. *(Kriging for the Gaussian model) As an example, assume that the dataset* $\{(y_i, \boldsymbol{s}_i)\}_{i=1}^n$ *satisfy the SRE data model* (2.16) *with known parameter set* $\Pi = \{\boldsymbol{\beta}, \sigma_\epsilon^2, \vartheta\}$ *and* $\mathcal{S}$ *the set of spatial indices. Let* $\boldsymbol{z} \sim \mathrm{N}(0, \boldsymbol{\Sigma}(\vartheta))$ *where* $\boldsymbol{\Sigma}(\vartheta) = [c(\boldsymbol{s}_i - \boldsymbol{s}_j; \vartheta)]_{i,j=1}^n$ *and* $c(\cdot; \vartheta)$ *is a covariance function. The measurement errors* $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma_\epsilon^2 I_n)$*. Consider now a location set* $\mathcal{S}^* = \{\boldsymbol{s}_0^*, \ldots, \boldsymbol{s}_N^*\}$*. Since* $y(\cdot)$ *is a GP, then* $\boldsymbol{y}^* = \boldsymbol{y}(\mathcal{S}^*)$ *and* $\boldsymbol{y} = \boldsymbol{y}(\mathcal{S})$ *are jointly normal, where* $\boldsymbol{y}(\mathcal{S}) = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))'$*. That is*

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{y}^* \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_{y^*} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{(y,y)} + \sigma_\epsilon^2 I_n & \boldsymbol{\Sigma}_{(y,y^*)} \\ \Sigma_{(y^*,y)} & \Sigma_{(y^*,y^*)} \end{bmatrix} \right), \tag{2.19}$$

where the generics covariance matrix $\Sigma_{(\boldsymbol{y},\boldsymbol{y}^*)} = [\mathrm{Cov}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j^*))]_{\boldsymbol{s}_i \in \mathcal{S}, \boldsymbol{s}_j^* \in \mathcal{S}^*} = [c(\boldsymbol{s}_i - \boldsymbol{s}_j^*; \vartheta)]_{\boldsymbol{s}_i \in \mathcal{S}, \boldsymbol{s}_j^* \in \mathcal{S}^*}$*. For GP, the kriging predictor is the conditional expectation,* $\mathrm{E}(z(\boldsymbol{s}_0) \mid \boldsymbol{z})$ *i.e.*

$$\begin{aligned} &\boldsymbol{y}^* | \boldsymbol{y} \sim \mathrm{N}(\mu_{\boldsymbol{y}^*|\boldsymbol{y}}, \Sigma_{\boldsymbol{y}^*|\boldsymbol{y}}) \\ &\mu_{\boldsymbol{y}^*|\boldsymbol{y}} = \mu_{\boldsymbol{y}^*} + \Sigma_{y^*y}(\boldsymbol{\Sigma} + \sigma_\epsilon^2 I_n)^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y)' \\ &\Sigma_{\boldsymbol{y}^*|\boldsymbol{y}} = \Sigma_{y*y*} - \Sigma_{y^*y}(\boldsymbol{\Sigma} + \sigma_\epsilon^2 I_n)^{-1}\Sigma_{y^*y}' \end{aligned} \tag{2.20}$$

Note that for the locations in both sets, i.e. $\{\boldsymbol{s} : \boldsymbol{s} \in \mathcal{S} \cap \mathcal{S}^*\}$*, the kriging gives filtered values of the corresponding variables, filtering out the measurement error.*

Allowing regression coefficients to vary over space, i.e. moving from a fixed parameter vector $\boldsymbol{\beta}$ to spatially varying functions $\boldsymbol{\beta}(\cdot)$, provides a flexible extension of the spatial regression model (2.15) that can capture local heterogeneity in covariate effects. This is particularly useful when relationships between predictors and the response are not constant across the study region, enabling the model to better reflect spatial nonstationarity and improve both interpretation and predictive accuracy.

In particular, the idea is to move from $\boldsymbol{X}'(\cdot)\boldsymbol{\beta}$ to $\boldsymbol{X}'(\cdot)\boldsymbol{\beta}(\cdot)$, where $\boldsymbol{\beta}(\cdot) : \mathcal{D} \to \mathbb{R}^b$ is a function. This technique has been developed in the more general framework of functional data analysis (Ramsay and Silverman, 2007), where the aim is to infer the unknown underlying function $\boldsymbol{\beta}(\cdot)$ from noisy observations. A common approach is to represent $\boldsymbol{\beta}(\cdot)$ through an unknown linear combination of know basis functions. Typical choices of basis functions include radial basis functions (Martin, Buhmann, and Ablowitz, 2003), and spline basis functions (De Boor and De Boor, 1978). The construction of such bases usually requires a reference set of points, called knots, $\mathcal{V}_R = \{v_i\}_{i=1}^R \subset \mathcal{D}$, that span the domain. The finer the grid of knots, the greater the local flexibility of the resulting function. Within this framework, the SRE model with spatially varying coefficients becomes

$$y(\boldsymbol{s}) - \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\Psi}(\boldsymbol{s})\boldsymbol{\beta} = z(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \quad \text{with } \boldsymbol{s} \in \mathcal{D}, \tag{2.21}$$

where $\boldsymbol{x}(\boldsymbol{s})$ is the $(b \times 1)$ vector of fixed-effect covariates, and $\boldsymbol{\Psi}(\cdot) = \mathrm{blockdiag}(\boldsymbol{\psi}_1(\cdot), \ldots, \boldsymbol{\psi}_b(\cdot))$ is the $(b \times bR)$ basis functions design matrix, where $\boldsymbol{\psi}_i(\cdot) = (\phi_{i,1}(\cdot), \ldots, \phi_{i,R}(\cdot))$ is the a set of known

basis function for i i-components for $i = 1, \dots, b$ with coefficients collected in the $(bR \times 1)$ vector $\boldsymbol{\beta} = \text{stack}(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_b)$ where $\boldsymbol{\beta}_i$ are the coefficients for the i-components. The operator blockdiag$(\cdot)$ is the block diagonal matrix operator, and stack$(\cdot)$ is the stacking operator. The corresponding data model in (2.21) follows directly, and estimation may be performed within either the GLS or maximum likelihood framework.

2.4.2 Multivariate spatial modelling

When the q-variate process is considered, the cokriging procedure is a natural extension of kriging. We are interested in making inferences on the $\boldsymbol{z}(\boldsymbol{s}) = (z_1(\mathbf{s}), \dots, z_q(\mathbf{s}))', \mathbf{s} \in \mathcal{D}$ given a finite number of observations $\boldsymbol{y}$. Specifically, a variable of interest is "cokriged" at a specific location from data about itself and about auxiliary variables in the neighbourhood (details on the definition of the neighbourhood will be discussed later).

Depending on how the measurements of the different variables are scattered in space, we distinguish between isotopic and heterotopic data sets. Let $\mathcal{S}_i$ for $i = 1, \dots, p$ be the set of observed locations for the i-process, with $|\mathcal{S}_i| = n_i$. If the data are measured for each variable at all sampling points, namely $\mathcal{S}_1 = \cdots = \mathcal{S}_p$, it is called an isotopic configuration. In contrast, the completely heterotopic configuration refers to the case where each variable is observed only at distinct, non-overlapping locations, that means $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$ for $i = 1, \dots, p$. A less restricted case is the partially heterotopic setting, in which only some variables share some sites such that $\mathcal{S}_i \cap \mathcal{S}_j = \mathcal{S}_{ij}$ for some i and j.

The most informative scenario is the isotopic configuration, while entirely heterotopic data poses a problem for inferring the cross variogram or covariance model. Indeed, empirical cross-variogram (2.8) cannot be computed for entirely heterotopic data. With partially heterotopic data, it is advisable, whenever possible, to infer the cross variogram or the covariance function model based on the isotopic subset of the data.

To model spatial multivariate processes, it is necessary to define both the direct covariance $\boldsymbol{c}_{ii}(\cdot)$ and the cross-covariance $\boldsymbol{c}_{ij}(\cdot)$ such that the multivariate covariance $\boldsymbol{C}(\cdot) = \text{Cov}[z_1(\cdot), \dots, z_p(\cdot)] = [c_{ij}(\cdot; \vartheta)]_{i,j=1}^{p}$. Two popular choices of multivariate covariance are the intrinsic correlation model and the LCM (2.10) discussed in Section 2.2.

Given the specification of the cross-covariance among the p processes, the SRE model (2.15) can be extended to allow the multivariate case,

$$\boldsymbol{y}(\boldsymbol{s}) - \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} = \boldsymbol{W}\boldsymbol{z}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}), \quad \text{with } \boldsymbol{s} \in \mathcal{D}, \tag{2.22}$$

where $\boldsymbol{y}(\boldsymbol{s}) = (y_1(\boldsymbol{s}), \ldots, y_p(\boldsymbol{s}))'$ is the p-dimensional vector of observations, $\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}$ is the deterministic mean structure with design matrix $\boldsymbol{X}(\boldsymbol{s}) = \text{blockdiag}(\boldsymbol{x_1}(\boldsymbol{s}), \ldots, \boldsymbol{x_p}(\boldsymbol{s}))$ where $\boldsymbol{x_i}(\boldsymbol{s}) = (x_{i,1}(\boldsymbol{s}), \ldots, x_{i,b_i}(\boldsymbol{s}))$ is a $(1 \times b_i)$ vector of the covariates for the i-process. The coefficient vector $\boldsymbol{\beta} = \text{stack}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$ where each $\boldsymbol{\beta}_i$, for $i = 1, \ldots, p$, is a column vector of dimension b_i corresponding to the regression covariates associated with $y_i(\boldsymbol{s})$. $\boldsymbol{W}$ is a $p \times q$ loading matrix that linearly projects the latent random effects $\boldsymbol{z}(\boldsymbol{s}, t)$ into the observation space. $\boldsymbol{z}(\cdot) = (z_1(\cdot), \ldots, z_p(\cdot))'$ is a zero-mean, intrinsically stationary p-variate process with $\text{Cov}[z_1(\cdot), \ldots, z_p(\cdot)] = \boldsymbol{C}(\cdot)$. $\boldsymbol{\epsilon}(\cdot) = (\epsilon_1, \ldots, \epsilon_p)'$ is a p-variate measurement error with $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for all $i, j = 1, \ldots, p$ and $\text{Cov}(\boldsymbol{z}(\cdot), \boldsymbol{\epsilon}(\cdot)) = 0$. From the model (2.22) the corresponding data model in follows directly, and estimation may be performed within either the GLS or maximum likelihood framework. If $\boldsymbol{z}(\boldsymbol{s})$ is a GP, than the cokriging estimator is $\text{E}(\boldsymbol{y}(\boldsymbol{s}_0)|\boldsymbol{y})$.

2.4.3 Spatio-temporal modelling

Various approaches have been proposed to model spatio-temporal processes (e.g. Cressie and Wikle (2015), for a review). If one considers time as an extra dimension, then traditional spatial statistics techniques can be applied Stein (1999). Let $\boldsymbol{y}(\boldsymbol{s}, t) = (y_1(\boldsymbol{s}, t), \ldots, y_p(\boldsymbol{s}, t))' \in \mathbb{R}^p$ be a p-dimensional spatio-temporal random field defined over a spatial domain $\mathcal{D}$ and discrete time steps $t = 1, \ldots, T$, where $\boldsymbol{s} \in \mathcal{D}$ indexes the spatial location. The SRE model (2.22) can be extended to account for the temporal correlations as,

$$\boldsymbol{y}(\boldsymbol{s}, t) - \boldsymbol{X}'(\boldsymbol{s}, t)\boldsymbol{\beta} = \boldsymbol{W}\boldsymbol{z}(\boldsymbol{s}, t) + \varepsilon(\boldsymbol{s}, t), \tag{2.23}$$

where $\boldsymbol{X}(\boldsymbol{s}, t) = \text{blockdiag}(\boldsymbol{x}_1(\boldsymbol{s}, t), \ldots, \boldsymbol{x}_p(\boldsymbol{s}, t))$ is a $b \times p$ design matrix, $\boldsymbol{x}_i(\boldsymbol{s}, t)$ is a column vector of dimension b_i associated with $y_i(\boldsymbol{s}, t)$ for $i = 1, \ldots p$, and $b = \sum_{i=1}^{p} b_i$. The measurement errors are IID and $\boldsymbol{\epsilon} \sim \text{N}(\boldsymbol{0}, \boldsymbol{R})$. $\{\boldsymbol{z}(\boldsymbol{s}, t) : \boldsymbol{s} \in \mathcal{D}, t \in \mathcal{T}\}$ is the q-variate latent multivariate field which models both spatial and temporal covariance by a spatio-temporal GP (Cressie and H.-C. Huang, 1999; Gneiting, 2002; Stein, 2005). A simple approach to construct valid spatio-temporal covariance functions is simply by taking the sum or the product of a purely spatial and a purely temporal covariance function, as in Eq. (2.11)–(2.12). However this approach has been limited by the fact that such covariance functions are often not realistic for complicated processes, and dimensionality can prohibit practical implementation.

In contrast, spatio-temporal processes can also be considered from the multiple time series perspective (e.g. Kyriakidis and Journel (1999)). That is, each spatial location is associated with a time series. Then, multivariate time series techniques can be transferred to the space-time problem. These

models describe the temporal evolution of spatial processes by specifying suitable dynamic equations. The dynamical statistical approach models how the current state depends on previous states through either mechanistic or probabilistic relationships. Hereinafter, with a slight abuse of notation, we will use the term spatio-temporal process to refer specifically to this dynamical statistical approach.

The best-known probabilistic dynamic model in a purely temporal context is the discrete-time Markov process (see Definition 12). Considering the model (2.23), we impose the Markovian assumption on the latent process $\boldsymbol{z}(\boldsymbol{s},t) = (z_1(\boldsymbol{s},t),\dots,z_q(\boldsymbol{s},t))$ which evolves over time according to the linear equation:

$$\boldsymbol{z}(\boldsymbol{s},t) = \boldsymbol{F}\boldsymbol{z}(\boldsymbol{s},t-1) + \boldsymbol{\eta}(\boldsymbol{s},t), \tag{2.24}$$

where $\boldsymbol{F}$ denotes a $q \times q$ transition matrix describing the system's dynamics. The vector $\boldsymbol{\eta}(\boldsymbol{s},t) = (\eta_1(\boldsymbol{s},t),\dots,\eta_q(\boldsymbol{s},t))$ represents a q-dimensional zero-mean GP which is temporally independent but spatially correlated. Its multivariate covariance is given by $\mathrm{Cov}(\boldsymbol{\eta}(\boldsymbol{s},t)) = [c_{ij}(\boldsymbol{s}-\boldsymbol{s}';\vartheta)]_{i,j=1}^{q} = \boldsymbol{Q}$. The process starts from an initial state $\boldsymbol{z}(\boldsymbol{s},0)$, which is a GP with mean $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$. To better highlight the dynamic formulation of the space-time process, Eq. (2.24) can be written as $\boldsymbol{z}_t(\boldsymbol{s}) = \boldsymbol{F}\boldsymbol{z}_{t-1}(\boldsymbol{s}) + \boldsymbol{\eta}_t(\boldsymbol{s})$, which essentially captures the dynamic transition from the state, i.e. the spatial process at time $t-1$, $z_{t-1}(\boldsymbol{s})$ to the state at time t, $z_t(\boldsymbol{s})$.

Equation (2.23) is referred to as the measurement equation, while Equation (2.24) is called the state equation, as it describes how the latent process evolves over time. Together, the measurement and state equations (2.23)–(2.24) define the linear state space model (SSM; Shumway, Stoffer, and Stoffer (2000)) with parameters:

$$\Pi = \{\boldsymbol{\beta}, \boldsymbol{F}, \boldsymbol{W}, \boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\}. \tag{2.25}$$

This formulation is particularly appealing due to its generality and flexibility, and is capable of capturing both observed and latent components of the time series. Given the observed data $\boldsymbol{y}_{1:T} = (\boldsymbol{y}_1',\dots,\boldsymbol{y}_T')'$ where $\boldsymbol{y}_t = (y(\boldsymbol{s}_1,t),\dots,y(\boldsymbol{s}_m,t))'$, we wish to make inference on the random variable $\boldsymbol{z}(\boldsymbol{s}_0,t)$, regardless of whether there is a datum $\boldsymbol{z}(\boldsymbol{s}_0,t)$ available or not. Inference on the random variable $\boldsymbol{z}(\boldsymbol{s}_0,t')$ where $t' \in \{1,2,\dots,T\}$, from data $\boldsymbol{y}_{1:T}$ is known as filtering. Inference on random variable $\boldsymbol{z}(\boldsymbol{s}_0,t')$, where $t' \in \{1,2,\dots,T\}$, from data $\boldsymbol{y}_{1:T}$ is known as smoothing; and inference on the random variable $\boldsymbol{z}(\boldsymbol{s}_0,t')$, where $t' \in \{T+1,T+2,\dots\}$, from data $\boldsymbol{y}_{1:T}$ is known as forecasting.

In the spatio-temporal context, the SSM model (2.23)-(2.24) is neither new nor the most general formulation. Nevertheless, such models have received considerable attention in the environmental literature in recent years (see, e.g., the review in Cressie and Wikle (2002)). In this framework, the eigenvalues of the transition matrix $\boldsymbol{F}$ control the temporal correlation of the process: larger

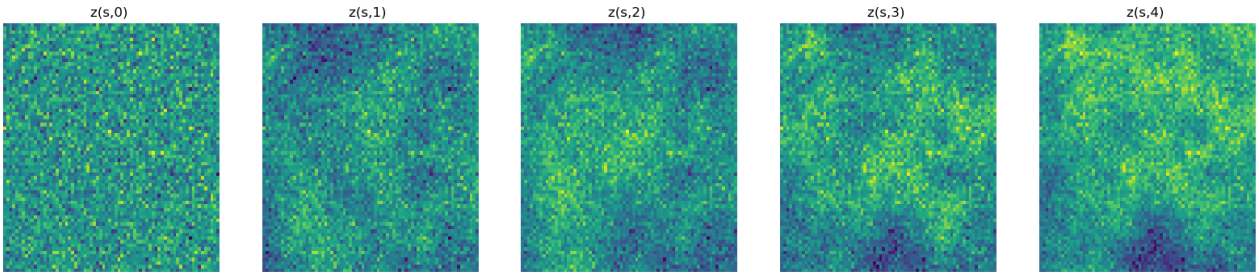


Figure 8: Simulation of the univariate state equation Eq. (2.24). Sample data $\{z(\boldsymbol{s}_1, t), \ldots, z(\boldsymbol{s}_n, t)\}$ for $t \in \{0, \ldots, 4\}$ are show from left to right. $z(\boldsymbol{s}, 0) \sim N(0, I_n)$, $f = 0.2$. The innovation covariance matrix $\boldsymbol{Q}$ is given by the stationary Matérn function parametrised by $\vartheta = \{\sigma^2 = 1, \nu = 1, \theta = 0.5\}$.

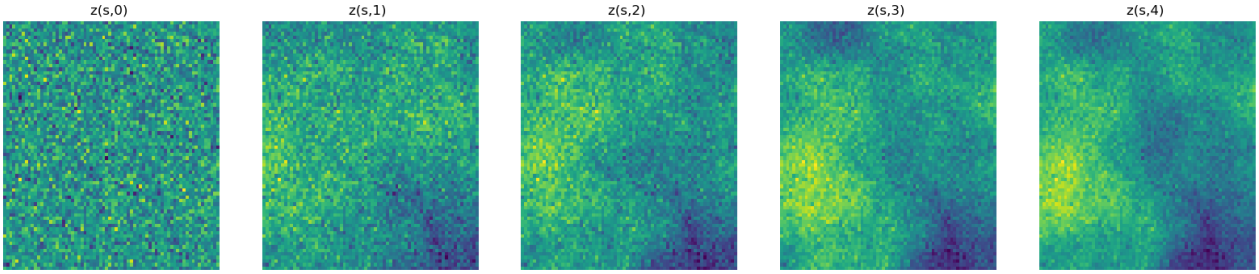


Figure 9: Simulation of the univariate state equation Eq. (2.24). Sample data $\{z(\boldsymbol{s}_1, t), \ldots, z(\boldsymbol{s}_n, t)\}$ for $t \in \{0, \ldots, 4\}$ are show from left to right. $z(\boldsymbol{s}, 0) \sim N(0, I_n)$, $f = 0.9$. The innovation covariance matrix $\boldsymbol{Q}$ is given by the stationary Matérn function parametrised by $\vartheta = \{\sigma^2 = 1, \nu = 1, \theta = 0.5\}$.

eigenvalues correspond to stronger correlation between the current and past states. Figures 8 and 9 illustrate a univariate simulation of the Markovian process $z(\boldsymbol{s}, t)$ for $t = 1, \ldots, 4$ with $\boldsymbol{F} = 0.2$ and $\boldsymbol{F} = 0.9$, respectively. The initial state is set as $z(\boldsymbol{s}, 0) \sim \mathrm{N}(0, I_n)$, and the innovation covariance function is given by the stationary Matérn function (2.1). As expected, the process with $\boldsymbol{F} = 0.9$ exhibits stronger temporal correlation.

The dynamical specification (2.23)-(2.24) of the spatio-temporal process is usually preferred for several reasons. First, it is usually derived from scientific knowledge about the phenomenon under study. Second, one can derive covariance models from the dynamical specification, and such models can be guaranteed to be valid (i.e., nonnegative-definite). Third, sequential updating allows rapid smoothing, filtering, and forecasting of the state from noisy and missing data observed at different times. For these reasons dynamical SSMs are a widely used framework for formulating Gaussian regression problems (Shumway, Stoffer, and Stoffer, 2000), particularly in spatio-temporal contexts; see, for instance, Xu and Wikle (2007), Cressie and Wikle (2002), and Fassò and Finazzi (2011) and Calculli et al. (2015) for extensions to multivariate heterotopic data.

Given the parameter set Π, inference on the unobserved latent states $z(\boldsymbol{s}, t)$ can be efficiently performed using the Kalman filter (KF) and the Kalman smoother (KS; Hamilton (2020)). Specifically, the KS yields the state estimate $\boldsymbol{z}_t^T = \mathrm{E}_\Pi(\boldsymbol{z}_t|\boldsymbol{Y})$ and the corresponding uncertainty $\boldsymbol{P}_t^T = \mathrm{Var}_\Pi(\boldsymbol{z}_t|\boldsymbol{Y})$.

The notations $\mathrm{E}_\Pi(\cdot)$ and $\mathrm{Var}_\Pi(\cdot)$ mean the expected value and the variance given the known set of parameters Π and hence the distribution from which the data is sampled. As $\boldsymbol{\varepsilon}(\boldsymbol{s}, t)$ and $\boldsymbol{\eta}(\boldsymbol{s}, t)$ are Gaussian, the inference on $\boldsymbol{z}(\boldsymbol{s}, t)$ using the KF results in the best linear estimator in the minimum mean-square-error sense (Humpherys, Redd, and West, 2012). In the more common case where Π is unknown, the Expectation-Maximisation (EM) algorithm is typically employed for parameter estimation (Shumway, Stoffer, and Stoffer, 2000).

The EM algorithm is based on the complete-data likelihood function $L_\Pi(\boldsymbol{y}, \boldsymbol{z})$ (or equivalently on the complete-data log-likelihood function $\log L_\Pi(\boldsymbol{y}, \boldsymbol{z})$, since the logarithm is monotone). At each iteration $k = 1, 2, \ldots$, the EM algorithm consists of an expectation (E) step and a maximisation (M) step. Given the current values of the parameters $\Pi^{(k)}$, the E-step computes the expected value of the complete likelihood function conditional on the observation $\boldsymbol{y}$, said:

$$Q(\Pi \mid \Pi^{(k)}) = \mathrm{E}_{\Pi^{(k)}}[\log L(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{y}, \boldsymbol{X}; \Pi)]. \tag{2.26}$$

At the M-step, the next value of Π, namely $\Pi^{(k+1)}$ is chosen so that $Q(\Pi^{(k+1)} \mid \Pi^{(k)}) \geq Q(\Pi \mid \Pi^{(k)})$ for each k. The maximisation step is equivalent to the maximisation of $Q(\Pi \mid \Pi^{(k)})$:

$$\Pi^{(k+1)} = \arg\max_{\Pi \in \mathcal{P}} Q(\Pi \mid \Pi^{(k)}), \tag{2.27}$$

where the maximisation is done on a suitable compact set $\mathcal{P}$ which contains Π as an interior point.

The EM algorithm works to improve $Q(\Pi \mid \Pi^{(k)})$ rather than directly improving the observed log-likelihood $\log L(\boldsymbol{y}; \Pi)$. It is possible to show that improvements to the former imply improvements to the latter (Dempster, Laird, and Rubin, 1977). Indeed, for each iteration k this quantity holds $\log L(\boldsymbol{y}; \Pi) - \log L(\boldsymbol{y}; \Pi^{(k)}) \geq Q(\Pi \mid \Pi^{(k)}) - Q(\Pi^{(k)} \mid \Pi^{(k)})$. In words, the maximisation at the M-step chose Π, said $\Pi^{(k+1)}$, to improve $Q(\Pi \mid \Pi^{(k)})$ which causes $\log L(\boldsymbol{y}; \Pi)$ to improve as well.

EM is especially useful when the likelihood is an exponential family: the E-step becomes the sum of expectations of sufficient statistics, and the M-step involves maximising a linear function. Here, the problem is formulated as a discrete SSM; thus, these sufficient statistics can be calculated by the KS (Amisigo and Van De Giesen, 2005).

Due to the high dimensionality, the measurement error and process covariance matrices typically involve too many parameters to estimate directly, making the EM approach of limited utility. In such high-dimensional settings, the convergence of the EM algorithm is also likely to be problematic. For this reason, the matrices in Π must be suitably parametrised. Shumway, Stoffer, and Stoffer (2000) discusses modifications to the EM algorithm to accommodate fully restricted parameter matrices. However, it is not clear how they account for partially restricted or parameterised model matrices in

the SSM framework. One approach to deal with partially restricted parameter matrices is to set initial parameters (in the EM algorithm) to agree with the known values. Then, in the M-step, only those parameters that require estimation are updated, so that the fixed parameters do not change (Shumway, Stoffer, and Stoffer, 2000). Alternatively, one can update all parameters, but then immediately impute the known values for the fixed parameters. Although these approaches are relatively easy to implement, it is not clear that they give the maximum likelihood estimates under the state-space model assumptions. Another approach (Xu and Wikle, 2007) is to develop general EM algorithms to account directly for the restricted or partially restricted model matrices. However, the fact that we prefer closed-form update formulas does put a limit on our choices for parametrisations. In general, the EM algorithm is useful for spatio-temporal SSM that are effectively parametrised by relatively few parameters. Through the thesis, and in particular in Chapters 5 and 6, some of the possible parametrisations will be considered.

Chapter 3

Fundamentals of discrete-space processes

In Section 2, the spatial domain $\mathcal{D}$ was defined as a continuous region. However, situations also arise where data are defined on discrete domains, such as lattices (regular grids) or, more generally, on any collection of countably many spatial locations. The aim of the following chapter is therefore twofold: first, to formalise the concept of a discrete spatial domain, and second, to present useful models for handling data defined under this assumption. Datasets defined on discrete domains are also referred to as areal datasets. In what follows, we focus on univariate spatial modelling, although the motivating examples of this dissertation concern spatio-temporal multivariate modelling.

Formally, a univariate discrete-space process is a collection of random variables $\{z(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$, where $\mathcal{D}$ is a countable index set. The collection $\mathcal{D}$ is called a lattice, which may be spatially regular or irregular, and is equipped with a neighbourhood structure. More generally, a lattice can be represented as a graph, where the sites correspond to vertices and the edges define the neighbourhood structure specifying which elements of the index set $\mathcal{D}$ are considered neighbours of one another.

Definition 14. *Given the spatial lattice $\mathcal{D}$, the neighbourhood structure of a site $\boldsymbol{s} \in \mathcal{D}$ is the collection of indices $\mathcal{N}(\boldsymbol{s}) \subseteq \mathcal{D} \setminus \{\boldsymbol{s}\}$. The set $\mathcal{N}(s)$ is the set of indices that are neighbours of index $\boldsymbol{s}$.*

Definition 14 provides a definition of neighbourhood without explicitly addressing how to construct the set $\mathcal{N}(\mathbf{s})$. In practice, there are several ways to define neighbourhood structures; in spatial statistics, distance-based definitions are often particularly appropriate, as they naturally reflect spatial proximity. Common examples include the first-order neighbourhood (rook's move), the second-order neighbourhood (queen's move), distance-based neighbourhoods, and k-nearest neighbours (k-NN). Given a lattice $\mathcal{D} \subset \mathbb{R}^2$, the first-order neighbourhood of a site $\mathbf{s} \in \mathcal{D}$, denoted $\mathcal{N}_1(\mathbf{s})$, is the set of sites that share an edge with $\mathbf{s}$ (i.e., the cells immediately above, below, to the left, and to the right). The second-order neighbourhood, denoted $\mathcal{N}_2(\mathbf{s})$, is the set of sites that share either an edge or a

corner with $\mathbf{s}$, which in two dimensions corresponds to up to eight surrounding sites. Let $d : \mathcal{D} \to \mathbb{R}^+$ be a distance function on $\mathcal{D}$. The distance-based neighbourhood of a site $\mathbf{s}$ with threshold $r > 0$ is defined as $\mathcal{N}_r(\mathbf{s}) = \{\mathbf{s}' \in \mathcal{D} \setminus \{\mathbf{s}\} : d(\mathbf{s}, \mathbf{s}') \leq r\}$. Finally, for a fixed integer $k \geq 1$, the k-nearest neighbours of $\mathbf{s}$ are defined as the k sites in $\mathcal{D} \setminus \{\mathbf{s}\}$ with the smallest distances $d(\mathbf{s}, \mathbf{s}')$. Figure 10 illustrates first-order and second-order neighbourhoods, while Figure 11 shows distance-based and k-nearest neighbours.

Having established the concepts of lattice and neighbourhood structures, in the next section, we introduces two fundamental models for discrete-space data: the simultaneous autoregressive (SAR) and conditional autoregressive (CAR) models.

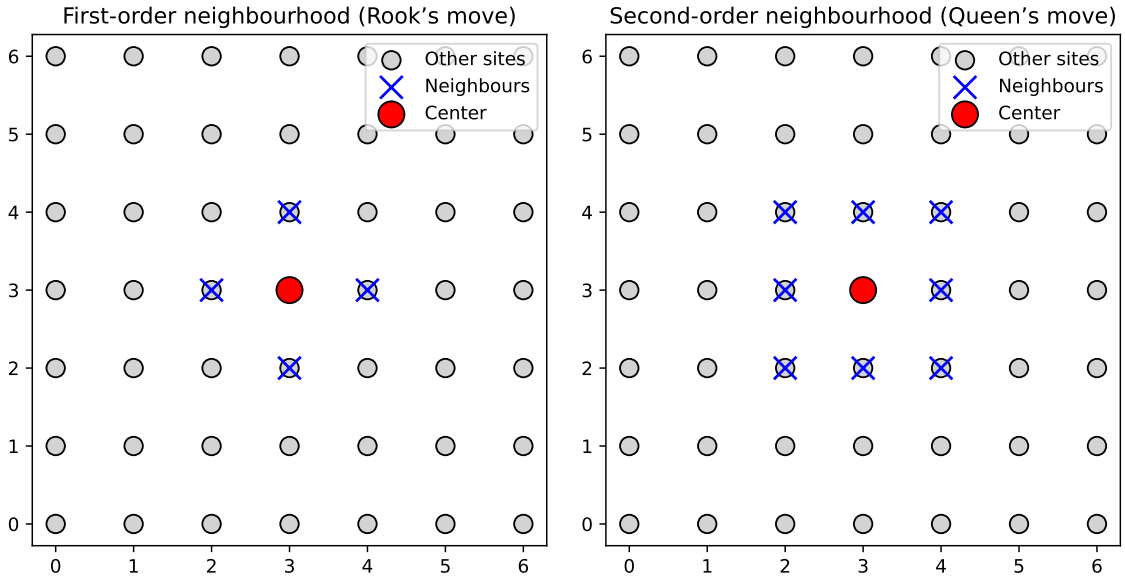
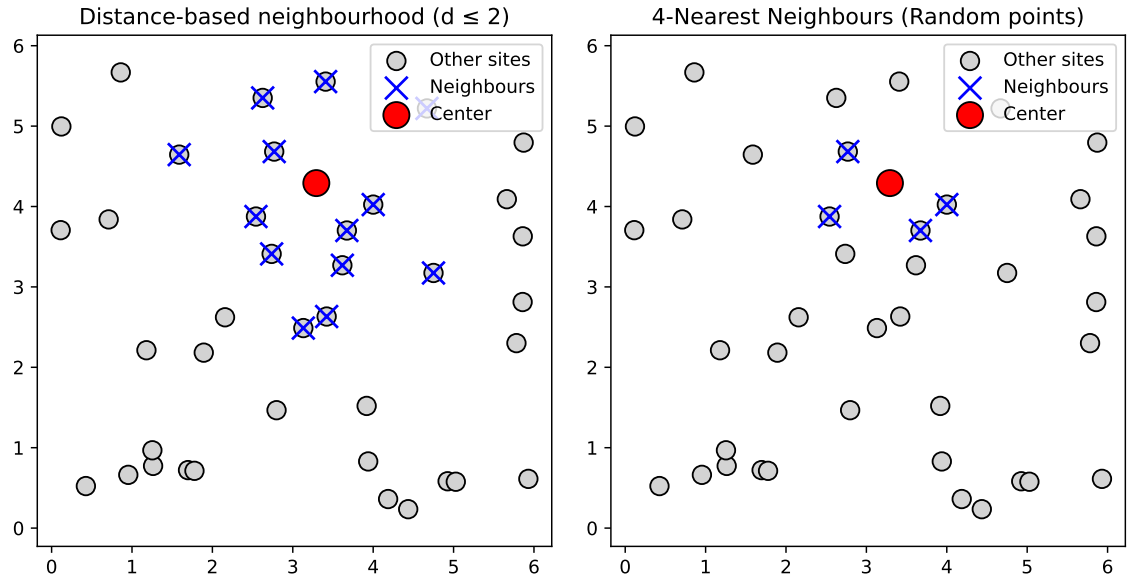


Figure 10: Neighbourhood structures for regular lattice data in $\mathbb{R}^2$: first-order neighbourhood (rook's move) on the left, and second-order neighbourhood (queen's move) on the right.

3.1 Simultaneous autoregressive (SAR) and conditional autoregressive (CAR)

In this section, we introduce the simultaneous autoregressive (SAR) and conditional autoregressive (CAR) models, two principal frameworks for modelling discrete-space data. These approaches are theoretically distinguished by their formulation: the SAR model specifies the joint distribution of all sites directly, whereas the CAR model is defined through the set of conditional distributions of each site given its neighbours. Let $\mathcal{D} = \{(x_i, y_j) : i = 1, \ldots, n\}$ be the discrete lattice.

We start by considering the SAR model, which was introduced by Whittle (1954). Suppose that



(a) (b)

Figure 11: Neighbourhood structures for irregular lattice data in $\mathbb{R}^2$: distance-based neighbourhoods, on the left, and k-nearest neighbours on the right.

the process $\{\epsilon(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ is a GP of IID random variables. Then, the SAR model defines the process $\{z(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ by specifying the linear system

$$(\boldsymbol{I}_n - \boldsymbol{B})(\boldsymbol{z} - \boldsymbol{X}'\boldsymbol{\beta}) = \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{z} = (z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_n))'$ is a realisation of the process $z(\boldsymbol{s})$ at the location set $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$. $\boldsymbol{X}'\boldsymbol{\beta}$ is the fixed effect term, $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \boldsymbol{\Lambda})$ with diagonal covariance matrix, $\boldsymbol{B} = [b_{ij}]_{i,j=1}^{n}$ is a matrix of spatial dependence, such that $b_{ij} > 0$ if $z(\boldsymbol{s}_i)$ is thought to depend positively with $z(\boldsymbol{s}_j)$, and if $z(\boldsymbol{s}_i)$ is thought to not depend on $z(\boldsymbol{s}_j)$ then $b_{ij} = 0$. It is suppose that $b_{ii} = 0$ for $i = \{1, \ldots, n\}$. The SAR model (3.1) holds assuming $(\boldsymbol{I}_n - \boldsymbol{B})$ non singular. Since $\boldsymbol{z}$ is a linear combination of $\boldsymbol{\epsilon}$, then $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{X}'\boldsymbol{\beta}, (\boldsymbol{I}_n - \boldsymbol{B})^{-1}\boldsymbol{\Lambda}(\boldsymbol{I}_n - \boldsymbol{B}')^{-1})$. The model's parameter set is $\Pi = \{\boldsymbol{\beta}, \mathrm{diag}(\boldsymbol{\Lambda}), \boldsymbol{B}\}$ and can be estimated through the MLE, as show in Example 4.

Example 4. *(SAR model estimation) This example presents the maximum likelihood estimation (MLE) for the model in* (3.1) *and assuming a parametric form for* $\boldsymbol{B}$ *given by* $\boldsymbol{B} = \lambda\boldsymbol{W}$*, and* $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma_\epsilon^2\boldsymbol{I}_n)$*. The matrix* $\boldsymbol{W} = [w_{ij}]$ *contains values computed from spatial proximity measures. A basic choice is* $w_{ij} = 1$ *if* $s_i \in \mathcal{N}(s_j)$*, and* $w_{ij} = 0$ *otherwise. With this specification, the SAR model can be written as* $(\boldsymbol{I}_n - \lambda\boldsymbol{W})(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\epsilon}$*. The parameter set* $\Pi = \{\boldsymbol{\beta}, \lambda, \sigma_\epsilon^2\}$ *can then be estimated using the ML methods.*

The development of the SAR model (3.1) starts by specifying a joint model for the data. Another approach is to start with conditional models instead, i.e., the conditional distribution of $z(\boldsymbol{s})$ given

its neighbours $\mathcal{N}(\boldsymbol{s})$ with $\boldsymbol{s} \in \mathcal{D}$. This led to the so-called conditional autoregressive (CAR) model, which was introduced by Besag (1974). We begin by recalling the Markov property in Definition 12: a stochastic process $\{z(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ defined on the lattice $\mathcal{D}$ satisfies the Markov property if Eq. (3.2) holds for each $\boldsymbol{s} \in \mathcal{D}$.

$$\mathrm{P}(z(\boldsymbol{s})|\{z(\boldsymbol{u}) : \boldsymbol{u} \in \mathcal{D})\}) = \mathrm{P}(z(\boldsymbol{s})|\{z(\boldsymbol{u}) : \boldsymbol{u} \in \mathcal{N}(\boldsymbol{s})\}) \tag{3.2}$$

Basically, the Markov property in (3.2) states that a variable is conditionally independent of all other variables given its neighbours,

$$z(\boldsymbol{s}) \perp \{z(\boldsymbol{u}) : \boldsymbol{u} \in \mathcal{D} \backslash \{\boldsymbol{s}\}\} | \{z(\boldsymbol{u}) : \boldsymbol{u} \in \mathcal{N}(\boldsymbol{s})\} \tag{3.3}$$

where $\perp$ indicates stochastic independence. $\mathrm{P}(z(\boldsymbol{s})|\{z(\boldsymbol{u}) : \boldsymbol{u} \in \mathcal{D})\})$ is the so-called local specification, i.e., the outcome on the $\boldsymbol{s} \in \mathcal{D}$ depends only on its set of neighbours. If Eq. (3.2) holds, then $z(\boldsymbol{s})$ is said Markov random field. The Markov property in Eq. (3.2) induce a factorisation from which it is possible to define the joint distribution of $z(\boldsymbol{s})$. For Gaussian processes, Besag (1974) showed that the factorisation yields the joint distribution $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{X}'\boldsymbol{\beta}, (\boldsymbol{I}_n - \boldsymbol{C})^{-1}\boldsymbol{M})$ where $(\boldsymbol{I}_n - \boldsymbol{C})$ is not singular and $(\boldsymbol{I}_n - \boldsymbol{C})^{-1}\boldsymbol{M}$ is a valid covariance matrix with $\boldsymbol{M} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_n^2)$.

One reason for the popularity of the CAR models is that conditional specifications can be preferable for estimation and model interpretation, and in fact, any SAR model can be expressed as a CAR model but not vice versa (see Cressie and Wikle (2015), for details). A disadvantage with the CAR models of Besag (1974) and Besag and Kooperberg (1995) was that while they have been used for both lattices and spatially motivated graphs, they were only spatially consistent for regular lattices, which limited their applicability. However, Lindgren, Håvard Rue, and Lindström (2011) derived the CAR models in a new way that both removes the lattice constraint and allows for the construction of spatially consistent non-stationary CAR-like models. The method is based on re-formulating the problem in terms of stochastic partial differential equations (SPDEs) in combination with the Finite Element Method (FEM) from numerical analysis. We return to this method in Section 6.4.

3.2 Gaussian Markov Random Field

We have already introduced Markov random fields in the previous section. The Gaussian Markov random fields (GMRFs; see Havard Rue and Held (2005), for extensive details) arise when CAR models are defined for Gaussian processes. This section focuses on GMRFs for spatial modelling, which are particularly relevant to this PhD thesis, as they provide an efficient representation of Gaussian

processes.

The Markov property (3.3) implies that if $(z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n))' \sim \mathrm{N}(0, \mathbf{Q}^{-1})$, then the precision matrix $\mathbf{Q}$ satisfies $Q_{ij} = 0$ if and only if $\mathbf{s}_j \notin \mathcal{N}(\mathbf{s}_i) \cup \mathbf{s}_i$. The neighbourhood of $\boldsymbol{s}_i$ typically consists of all points that, in some sense, are close to s_i. In theory, there are no restrictions on the size of the neighbourhood, and one could, for example, have $\mathcal{N}(\boldsymbol{s}_i) = \mathcal{D} \setminus \boldsymbol{s}_i$, which shows that any multivariate normal distribution with a symmetric positive definite covariance matrix is a GMRF and vice versa. However, the advantages of the Markov assumption naturally occur when the neighbourhood is small. The sparse precision matrix is the main reason why using GMRFs is largely considered in spatial analysis.

Lindgren, Håvard Rue, and Lindström (2011) derived a method for explicit, and computationally efficient, continuous Markov representations of Gaussian Matérn fields. As previously mentioned, the method can be used to redefine the CAR models, removing the lattice constraint and allowing for non-stationary extensions. However, the most important implication of the work in Lindgren, Håvard Rue, and Lindström (2011) is that it provides a spatially consistent method for approximating continuous-space GP using GMRFs, which enables the use of sparse matrix techniques for GMRF inference, see Figure 12. In particular, Lindgren, Håvard Rue, and Lindström (2011) provides the explicit link between GP parameters and GMRF parameters. The goal is to approximate a GP with covariance matrix Σ using a GMRF with precision matrix Q such that Q^{-1} approximates Σ in a suitable norm. A GMRF representation of a Matérn GP can be efficiently constructed using the SPDE approach, in combination with the FEM over a triangulated domain. Figure 12 shows a comparison of precision matrices from two spatial models: the GMRF precision matrix obtained from the SPDE discretisation and Matérn GP. The GMRF representation of a continuous-space GP will be discussed in detail in Chapter 6, where we apply the SPDE method within the SSM framework. Additional details on the GMRF representation through the SPDE approach are provided in Appendix B.3.

Example 5. *(Advantages of sparse matrix) In Gaussian models, likelihood evaluation involves inversion of the covariance matrix, which has a complexity of $O(n^3)$, see Example 2. As a result, the likelihood computation becomes infeasible even for moderate n. By parametrising the model through the precision matrix and thus avoiding the direct inverse operation, sparsity can be exploited since many entries are exactly zero, reflecting conditional independencies. Sparse matrix algorithms such as Cholesky factorisation then scale faster, often close to linear in n, depending on the sparsity pattern or the bandwidth. This reduces both memory requirements and computational time, making likelihood-based inference practical for high-dimensional spatial models.*

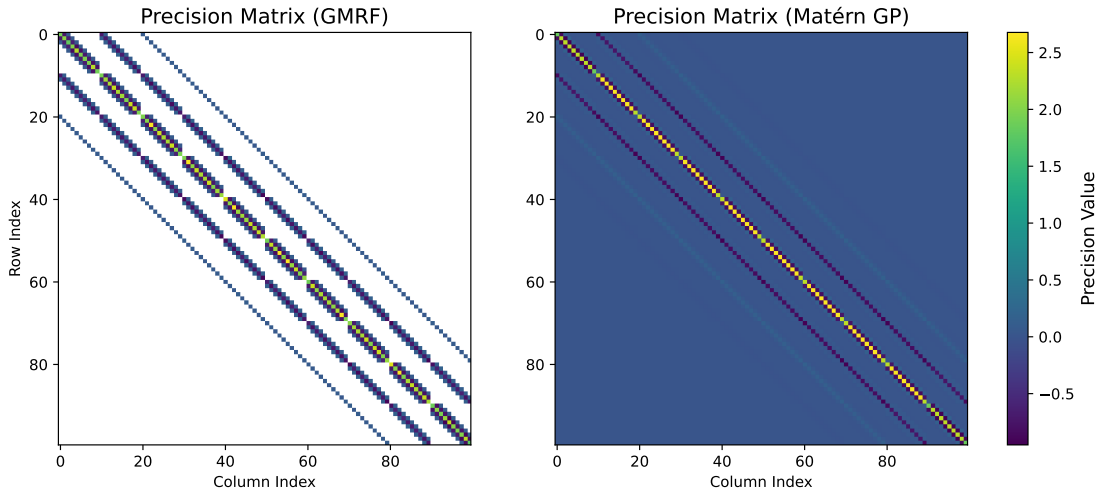


Figure 12: Comparison of precision matrices from two spatial models. Left: Gaussian Markov Random Field (GMRF) precision matrix obtained from the SPDE discretisation. Right: Matérn Gaussian Process (GP) precision matrix. Zero entries are shown in white, highlighting the sparsity pattern of the GMRF compared to the dense structure of the Matérn GP.

Chapter 4

Geostatistical dataset harmonisation: a dataset on livestock, meteorology and air quality in the Lombardy region, Italy

Section 1.1 highlighted the public relevance of monitoring the particulate matter (PM) in Lombardy region, situated in the heart of the Po Valley in Northern Italy, which is known to be highly polluted. Fine particulate matter ($PM_{2.5}$) has been identified as the most hazardous air pollutant (European Environmental Agency, 2022). Information about the air quality dynamics is essential for decision-makers to effectively mitigate adverse effects.

To support studies quantifying the role of the agricultural and livestock sectors on the Lombardy air quality, this chapter presents the development of a spatio-temporal dataset for the Lombardy air quality modelling. The dataset was developed within the framework of the AgrImOnIA project (*Agriculture Impact on Italian Air*), which aims to evaluate the impact of the livestock sector on air quality in Lombardy and to enable comparative studies across European regions. The resulting dataset, referred to as the Agrimonia dataset, is openly accessible via the Zenodo repository (Fassò, Rodeschini, et al., 2023b). The Agrimonia dataset provides harmonised information on emissions (including ammonia), agricultural activities, and air pollution in a single structure, thereby addressing the challenges associated with combining spatio-temporal data from multiple sources. By relying, whenever possible, on data available at the European level, the dataset is intended to serve as a standardised reference for interdisciplinary research and for interregional comparisons across Europe.

To make spatial predictions, it is necessary to have covariates defined across the entire domain. We address this challenge by constructing a gridded representation of the Agrimonia dataset, referred to as the Agrimonia Gridded Dataset (AGC), which enables consistent spatial alignment between air

pollution data and explanatory variables.

The balance of this chapter is organised as follows. Section 4.1 provides an overview of explanatory variables useful to model the air quality. Section 4.2 introduces the data sources used to build the Agrimonia dataset. The transformations applied to harmonise the dataset, as well as the methodologies used to impute missing data and handle negative values, are detailed in Section 4.3. In Section 4.4, we detail the gridded representation of the Agrimonia dataset, while further details are provided in Appendix A. Much of the material and methodological framework in this chapter is based on Fassò, Rodeschini, et al. (2023a), which serves as the main reference for the analyses and approaches presented here.

4.1 Overview

Several studies have found that weather is a crucial factor in air quality modelling (Bertaccini, Dukic, and Ignaccolo, 2012; Ignaccolo, Mateu, and Giraldo, 2014; Merk and Otto, 2020; Stuart K Grange, Jörg Sintermann, and Christoph Hueglin, 2023; Chang and Zou, 2022). Changes in weather conditions, such as temperature, precipitation, and wind speed and direction, can affect atmospheric stability and turbulence, which can influence the transport and deposition of pollutants. For instance, temperature and boundary layer height are usually negatively related to air pollutant concentrations. Similarly, we typically observe reduced PM concentrations during periods with increased precipitation or wind speed. On the contrary, the direction and size of the effect of the relative humidity are still debated, but it undoubtedly affects the $PM_{2.5}$ concentrations (L. Zhang et al., 2017).

We considered agricultural influences, which appear to impact air quality (e.g., Thunis et al., 2021; Lovarelli, Conti, et al., 2020). We consider information about the average density of pigs and cattle per municipality (expressed as animals per km^2). Including livestock data is essential to capture the impact of ammonia (NH_3) emissions on air quality, as livestock farming is the major source of NH_3 emissions (up to 95%). Therefore, including LI data in air quality modelling can help better understand and mitigate livestock's impact on air pollution levels. Note that emissions refer to the amount of pollutants released into the atmosphere from sources such as heating systems, thermal power, or livestock activities. Concentrations, instead, indicate the level of pollutant actually measured in the air at a given place and time. While emissions depend on the environmental conditions at the source, concentrations also reflect atmospheric processes such as dispersion, transport, chemical reactions, and deposition.

In light of these reasons, to support quantitative analyses of the contribution of the agricultural and

livestock sectors to regional air quality, the Agrimonia dataset integrates information on air quality, meteorology, emissions, livestock, and land and soil use at a daily temporal resolution for the period 2016–2021. Because geostatistical methods can exploit information from neighbouring territories to improve predictive accuracy near borders (Cressie and Wikle, 2015), an area surrounding the Lombardy region is considered by applying a 0.3° buffer to the regional boundaries, as shown in Figure 13. The resulting neighbouring area intersects multiple adjacent regions. The dataset includes satellite data, model output and in situ measurements with different spatial and temporal resolutions from national and international agencies. Therefore, to combine the different datasets, a processing step is necessary.

4.2 Source data description

The data presented in this chapter relate to five dimensions: air quality (AQ), weather and climate (WE), pollutant emissions (EM), livestock (LI), and land and soil characteristics (LA). The various data sources used to construct the Agrimonia dataset are summarised in Table 1 and are described in the following subsections, with details on spatio-temporal resolution and data availability.

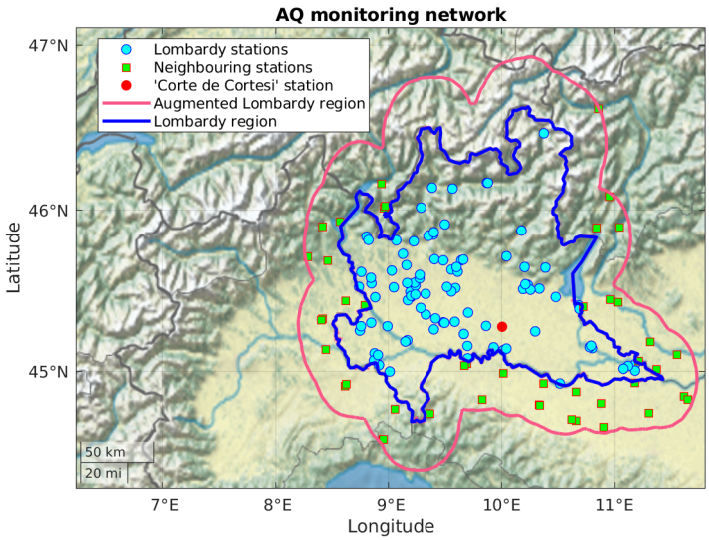


Figure 13: AQ network of $S = 141$ stations in the augmented Lombardy region (pink boundaries): $S_1 = 93$ stations are inside the Lombardy region (blue boundaries) and $S_2 = 48$ in the 0.3° buffer defining the neighbouring area between blue and pink boundaries. The station named 'Corte de Cortesi' is marked as a red circle and used as a reference.

Dimension	Source	Spatial coverage	Temporal resolution
Air quality (AQ)	EEA – Air pollution section European Environmental Agency (EEA) - Air pollution section (2021)	Europe	Daily / Hourly / Bi-hourly
	ARPA Lombardy – Air quality section	Lombardy	Daily / Hourly / Bi-hourly
Weather (WE)	Copernicus Climate Change Service (ERA5) Hersbach et al. (2018) and Muñoz Sabater (2021)	Europe	Hourly
Emission (EM)	Copernicus Atmosphere Monitoring Service (CAMS) Inness et al. (2019)	Global	Monthly
Livestock (LI)	National Data Bank (BDN) of the Zootechnical Registry Italian Ministry of Health (2021)	Italy	Biannual
Land cover (LA)	Copernicus Land Monitoring Service (CLMS) European Union (2018)	Europe	Only 2018
	Lombardy Region Agriculture Information System (SIARL) Lombardy Region, Sistema Informativo Agricoltura Regione Lombardia (SIARL) (2019)	Lombardy	Annual
	Copernicus Climate Change Service (ERA5) Muñoz Sabater (2021)	Europe	Hourly

Table 1: Input sources dataset used to build the Agrimonia dataset.

Air Quality

The AQ data are pollutant concentrations ($\mu g/m^3$) sampled at $S = 141$ ground-level monitoring stations, irregularly located over the augmented Lombardy region, as shown in Figure 13. For the Lombardy area, the AQ data of 93 stations are retrieved by the environmental protection agency of the Lombardy region (ARPA Lombardy, hereinafter ARPA), while outside of Lombardy, data are obtained by the European Environment Agency (EEA) from another 48 stations. Inside Lombardy, the variables listed in Table 2, namely $PM_{2.5}$, PM_{10}, NO_2, NO_x, CO, SO_2 and NH_3 with a daily temporal resolution come from the open data system of ARPA and are validated under EEA protocols. For the neighbouring EEA stations, data are either daily or hourly and cover all the above variables but NH_3.

The data used for the Lombardy neighbouring areas comes from the open-access service of the EEA. Table 2 summarises the pollutants selected, their sources and the number of sensors available for each pollutant.

Pollutant	Description	Source	Temporal resolution	Number of sensors
PM_{10}	Particulate matter with an aerodynamic diameter of less than 10 µm	ARPA Lombardy / EEA	Daily / hourly	107
$PM_{2.5}$	Particulate matter with an aerodynamic diameter of less than 2.5 µm	ARPA Lombardy / EEA	Daily / hourly / bi-hourly	54
CO	Carbon monoxide	ARPA Lombardy / EEA	Hourly	56
NH_3	Ammonia	ARPA Lombardy	Daily / hourly	10
NO_x	Nitrogen oxides	ARPA Lombardy / EEA	Hourly	110
NO_2	Nitrogen dioxide	ARPA Lombardy / EEA	Hourly	136
SO_2	Sulphur dioxide	ARPA Lombardy / EEA	Hourly / bi-hourly	40

Table 2: AQ pollutants concentrations data sources, descriptions, sampling frequency and number of sensors (or stations) per pollutant, which gives a total of 513 sensors throughout extended Lombardy.

Weather

Meteorological data are obtained from the Copernicus Climate Change Service through the ERA5 datasets containing the numerical model output computed by the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 is the fifth-generation ECMWF reanalysis of the global climate for the past decades. The reanalysis combines model data with observations from across the world into a globally complete and consistent dataset using the laws of atmospheric science. The ERA5 datasets used here are ERA5-Single level (Hersbach et al., 2018) and ERA5-Land (Muñoz Sabater, 2021). ERA5-Single level provides hourly estimates for various atmospheric and land-surface quantities with a regular grid scheme at various atmospheric levels. ERA5-Land provides near-surface variables over several decades at an enhanced resolution compared to the ERA5-Single level. ERA5-Single level and ERA5-Land datasets are provided on a regular latitude/longitude grid of 0.25×0.25 degrees and 0.1×0.1 degrees, respectively. Table 3 summarises the WE variables selected from ERA5 datasets, their sources, descriptions and units.

We complete the weather variables by calculating relative humidity (RH). Using the temperature

(T) and dew point temperature (T_{dew}), we compute the RH by using the August-Roche-Magnus approximation formula (Alduchov and Eskridge, 1996) in Eq. (4.1):

$$RH = 100 \times \exp\left(\frac{17.625 \times T_{dew}}{243.04 + T_{dew}} - \frac{17.625 \times T}{243.04 + T}\right). \tag{4.1}$$

Dataset	Variable	Description	Unit
ERA5 Land	10 m u-component of wind	Eastward component of the wind at 10 metres altitude	m/s
	10 m v-component of wind	Northward component of the wind at 10 metres altitude	m/s
	2 m dewpoint temperature	Temperature to which the air, at 2 metres above the surface of the Earth, would have to be cooled for saturation to occur	K
	2 m temperature	Temperature of air at 2 metres above the surface of land, sea or inland waters	K
	Total precipitation	Accumulated liquid and frozen water, comprising rain and snow that falls to the Earth's surface	m
	Surface net solar radiation	Amount of solar radiation that reaches a horizontal plane at the surface minus the amount reflected by the Earth's surface	J/m^2
ERA5 Single Level	100 m u-component of wind	Eastward component of the wind at 100 metres altitude	m/s
	100 m v-component of wind	Northward component of the wind at 100 metres altitude	m/s
	Boundary layer height	Depth of air next to the Earth's surface that is most affected by the resistance to the transfer of momentum, heat or moisture across the surface	m
	Surface pressure	Pressure (force per unit area) of the atmosphere at the surface of land	Pa
	Precipitation type	Type of precipitation on the Earth's surface. Values of precipitation type are: no precipitation (0), rain (1), freezing rain (3), snow (5), wet snow (6), mixture of rain and snow (7), ice pellets (8)	$Categ.$

Table 3: WE variables selected from ERA5 datasets. $Categ.$ stands for Categorical.

Emissions

The Copernicus Atmosphere Monitoring Service (CAMS) implemented by the ECMWF is one of the most recent global databases covering anthropogenic source emissions. CAMS datasets are compiled

emission inventories for many atmospheric compounds developed for the years 2000-2020 (Granier et al., 2019; Inness et al., 2019). These inventories are based on a combination of existing datasets and new information, describing anthropogenic emissions from fossil fuel use on land, natural emissions from vegetation, soil and more. The anthropogenic emissions on land are further separated into specific activity sectors (e.g. traffic, agriculture). Pollutant emissions data are provided by the CAMS-anthropogenic emissions dataset, which contains monthly global anthropogenic and natural emissions from 36 sources with a spatial resolution of (0.1×0.1) degrees. Table 4 summarises the emission variables selected, their origins, descriptions and units.

Variable	Emission origin	Description
NH_3	Livestock sector	Emissions of NH_3 originating from the livestock sector for manure management
	Agriculture soils	Emissions of NH_3 originating from agriculture soils
	Waste burning	Emissions of NH_3 originating from agriculture waste burning
	Total	Total emissions of NH_3 across all sectors
NO_x	Road transportation	Emissions of NO_x from on road transportation
	Total	Emissions of NO_x across all sectors
SO_2	Total	Emissions of SO_2 across all sectors

Table 4: EM variables selected from the CAMS-anthropogenic emissions dataset. Spatial resolution of (0.1× 0.1) degrees. Unit is $kg/(m^2 s)$.

Livestock

The Agrimonia dataset focuses on data on pigs and bovines because of their role in the NH_3 emission. Indeed, for Lombardy, about 87% of the annual NH_3 emissions are estimated to come from livestock management of these species (ARPA Lombardia Settore Monitoraggi Ambientali, INEMAR, 2022). The related emission pattern has a remarkable spatial and temporal variability in the study region.

The number of livestock per municipality is obtained by the Italian National Data Bank of the Zootechnical Registry (BDN) (Italian Ministry of Health, 2021), deriving from the livestock census in Italy. The BDN dataset is managed by the Italian Ministry of Health and represents the official source of data on livestock, both for the control authorities and users. The BDN data are updated every six months and aggregated at the municipality level. Figure 14 shows the number of swine and bovines in the augmented Lombardy region. The absolute number of livestock is particularly high in the Southeastern areas.

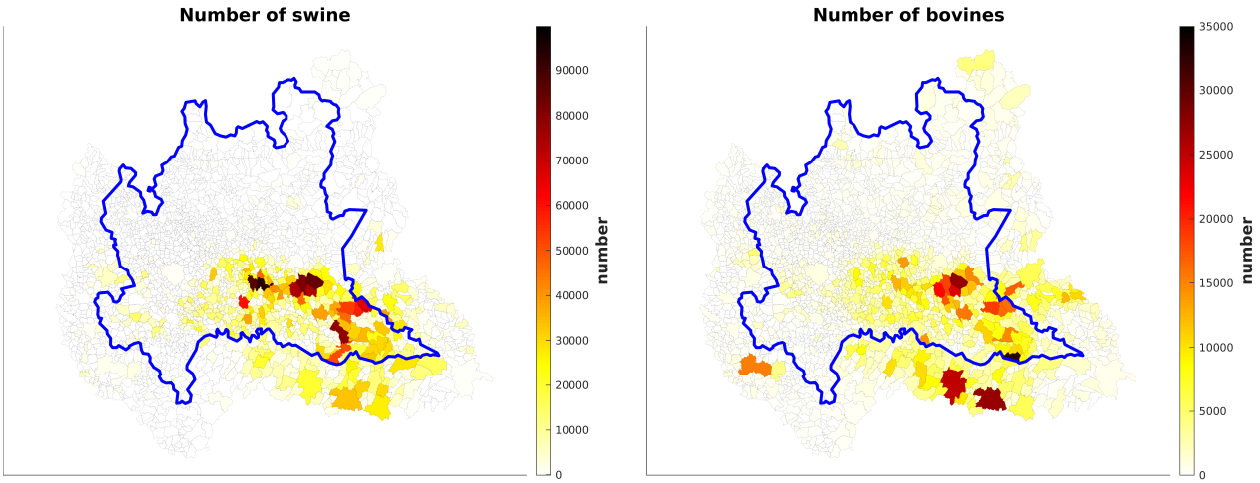


Figure 14: Number of swine (left) and bovines (right) in the augmented Lombardy region and neighbouring area, aggregated at the municipal level on 31 December 2021.

Land

The land cover, land and soil use are potentially important factors to assess the agricultural impact on air quality (Nenes et al., 2020). To characterise the land, we add the land cover, land use and soil use variables, summarised in Table 5. Land cover refers to the physical and biological features observed on the Earth's surface, such as forests, croplands, urban areas, or water bodies, whereas land use describes the way humans manage and utilise these surfaces, for example, for agriculture, residential, or industrial purposes. In addition, soil use provides complementary information on the cultivation type.

For the land cover, we consider high and low vegetation indices obtained by the ERA5-Land dataset (Muñoz Sabater, 2021) introduced before. For land use, we consider the Corine Land Cover (CLC; Büttner et al. (2021)) dataset. It is available for 2018 and consists of an inventory of land use in 44 classes within a minimum mapping unit of 25 hectares, and it is built using satellite images. Finally, we consider information on soil use and cultivation type provided by the SIARL dataset (Lombardy Region, Sistema Informativo Agricoltura Regione Lombardia (SIARL), 2019). It provides the classification of agricultural soil use in 21 classes at yearly frequency, and is available until 2019.

4.3 Data harmonisation and processing

The previous section discussed the input data's different spatial and temporal resolutions. We introduce here the methods used to harmonise the data before merging them into the Agrimonia dataset. We also consider missing data imputation and some variable transformations.

Dataset	Variable	Description	**Unit**
ERA5-Land (land cover)	Low vegetation index	One-half of the total green leaf area per unit horizontal ground surface area for low vegetation type	m^2/m^2
	High vegetation index	One-half of the total green leaf area per unit horizontal ground surface area for high vegetation type	m^2/m^2
CLC (land use)	Third-level of land use	Classification of land use with 44 classes	$Categ.$
SIARL (soil use)	Soil cultivation type	Classification of soil use with 21 classes	$Categ.$

Table 5: LA variables selected from the ERA5, CLC and SIARL datasets. $Categ.$ stands for Categorical.

Air quality

The AQ data, from ARPA and EEA networks, have both short and prolonged periods with stations turned off for maintenance, instrument calibration or other reasons. In some cases, measurements are taken at irregular intervals, or sampling policies change, and daily, bi-hourly and hourly AQ measurements are present in the network, see Table 2.

Since the presence of several consecutive missing values in a day may introduce a bias in the daily average, we implemented the missing imputation procedure at the hourly/bi-hourly frequency, which is summarised by the algorithm 1. For each time series, missing values are then imputed using a state-space model (Durbin and Koopman, 2012) and the relative Kalman smoother (Harvey, 1990), providing an estimate of the missing data and their uncertainty. Then uncertainty associated with the hourly estimate is propagated to the daily average, thus providing daily uncertainties due to the missing data imputation. Table 6 shows the missing data percentage by year for the different pollutants remaining after harmonisation.

Weather

This section describes the harmonisation process for the WE variables (see Table 7). The data about WE come from ERA5 datasets and are given by hourly reanalysis estimates in a regular grid format, and there are no problems with missing and negative values. As shown in Figure 15, the AQ station is misaligned with weather data. To associate weather time series to each AQ station, we use the inverse distance weighted (IDW) interpolation algorithm (Cressie and Kornak, 2003), considering the four nearest grid cells. After the hourly time series is created using the IDW approach for each station, we convert the temporal resolution from hourly to daily using different ensemble functions according to

Algorithm 1: AQ missing data imputation and daily averaging

Input: Hourly/bi-hourly time series $\{y_t\}_{t=1}^{T}$
Output: Daily series $\{\bar{y}_d\}$ with associated uncertainties $\{\sigma_d\}$

Preprocessing;
foreach $y_t \in \{y_1, \ldots, y_T\}$ **do**
    if y_t *not validated* ***or*** $y_t < 0$ **then**
        $y_t \leftarrow$ NaN;
Impute missing values using Kalman smoother;
$\hat{y}_t, \hat{\sigma}_t \leftarrow \text{SSM}(y_t)$;
Daily Aggregation;
foreach *day* d **do**
    if *number of consecutive hourly missing values* > 6 **then**
        $\bar{y}_d \leftarrow$ NaN;
    else
        $\bar{y}_d \leftarrow \dfrac{1}{n_d} \sum_{t \in d} \hat{y}_t$;
        $\sigma_d \leftarrow \sqrt{1/24^2 \sum_{t \in d} \hat{\sigma}_t^2}$;
return $\{\bar{y}_d, \sigma_d\}$

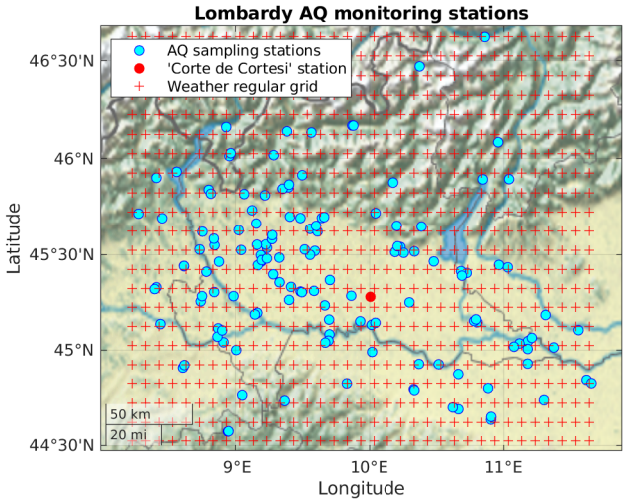


Figure 15: The irregularly located AQ monitoring stations (cyan circles) and the weather grid centres (red '+' symbols). The station named 'Corte de Cortesi' is marked as a red circle. .

the variable type (Cameletti, Ignaccolo, and Bande, 2011). Table 7 lists the weather variables in the Agrimonia dataset.

Emissions

The data about EM are from the CAMS datasets with a monthly temporal resolution and on a regular grid. As done for the WE variables, we performed a two-step transformation process to create daily

Variable name	Description	% 2016	% 2017	% 2018	% 2019	% 2020	% 2021
AQ_pm_10	Particulate matter with an aerodynamic diameter of less than 10 μg	6	6.3	4.7	5.7	4.4	11.6
AQ_pm_2.5	Particulate matter with an aerodynamic diameter of less than 2.5 μg	15	12.7	10.6	10.4	9.9	31
AQ_co	Carbon monoxide	8.3	8.4	6.2	6.8	9.7	13.8
AQ_nh3	Ammonia	33.8	22.7	29.7	28.7	16.3	18.5
AQ_nox	Several oxides of nitrogen	23.9	24.7	6.2	5.4	16.1	23
AQ_no2	Nitrogen dioxide	8.6	15.2	14.1	13	12	14
AQ_so2	Sulphur dioxide	9.6	7.6	5.6	5.5	3.2	7.2

Table 6: Missing data of AQ variables in the Agrimonia dataset. All AQ variables included in the Agrimonia dataset are harmonised to the daily time resolution through the Kalman smoother and the daily mean. The columns starting with '%' show the percentage of missing data after harmonisation for each pollutant grouped by year. Note that these percentages refer to the sensor numbers listed in Table 2 only.

emission time series. In the first step, we use the same IDW approach described for WE variables to compute emission values related to each monitoring station with monthly resolution. In the second step, we use spline interpolation techniques to convert the series to the same daily temporal resolution. To avoid oscillations, overshoots, edge effects and negative values, we use piecewise cubic Hermite interpolating polynomials (Fritsch and Carlson (1980); PCHIP). As discussed in more detail in the Technical Validation Section, this method interpolates the data smoothly, while retaining the data's shape and monotonicity.

Livestock

The data related to the livestock sector are retrieved from the BDN dataset, which provides the number of bovines and swine aggregated at the municipality level. The BDN dataset is updated every six months, in June and in December. As a result, for each municipality, a time series of 12 values is available. Each AQ station is associated with the time series of the municipality to which it belongs, see Figure 16 for swine and bovines. The PCHIP interpolation is used to increase the temporal resolution from biannual to daily. Subsequently, the municipal animal density is calculated by dividing the animal count by the area of the station municipality (expressed in km^2). In this way, we obtain the

Variable name	Description	**Temporal aggregation function**	**Units**
WE_temp_2m	Temperature of air at 2 m	Daily mean	$°C$
WE_wind_speed_10_mean	Mean intensity of the wind speed at 10 m	Daily mean	m/s
WE_wind_speed_10_max	Max intensity of the wind speed at 10 m	Daily max	m/s
WE_direction_10m	Wind direction at 10 m	Daily mode	$Categ.$
WE_tot_precipitation	The accumulated liquid and frozen water that falls to the Earth's surface	Daily sum	m
WE_precipitation_t	The type of precipitation at the time t	Daily mode	$Categ.$
WE_surface_pressure	Atmosphere pressure at the surface of land	Daily mean	Pa
WE_solar_radiation	Solar radiation at the surface minus the amount reflected by the Earth's surface	Daily max	J/m^2
WE_wind_speed_100_mean	Mean intensity of the wind speed at 100 m	Daily mean	m/s
WE_wind_speed_100_max	Max intensity of the wind speed at 100 m	Daily max	m/s
WE_direction_100m	Wind direction at 100 m	Daily mode	$Categ.$
WE_blh_layer_max	The maximum depth of the boundary layer	Daily max	m
WE_blh_layer_min	The minimum depth depth of the boundary layer	Daily min	m
WE_rh_min	Maximum relative humidity	Daily min	%
WE_rh_mean	Mean relative humidity	Daily mean	%
WE_rh_max	Minimum relative humidity	Daily max	%

Table 7: WE variables included in the Agrimonia dataset. Wind speed is calculated as the Euclidean norm of the wind vector with u- and v- components; the wind direction is discretised using the classical 8-wind rose: North (N), North-east (NE), East (E), South-east (SE), South (S), South-west (SW), West (W), North-west (NW); the temperature is converted from Kelvin to Celsius degrees. $Categ.$ stands for Categorical

Variable name	Description
EM_nh3_livestock_mm	Emissions of NH_3 originating from the livestock sector for the manure management
EM_nh3_agr_soils	Emissions of NH_3 originating from agriculture soils
EM_nh3_agr_waste_burn	Emissions of NH_3 originating from the burning of agricultural waste
EM_nh3_sum	Total emissions of NH_3 across all sectors (anthropogenic)
EM_nox_traffic	Emissions of NO_x from the on-road transportation
EM_nox_sum	Emissions of NO_x across all sectors (anthropogenic)
EM_no2_sum	Total emissions of SO_2 across all sectors (anthropogenic)

Table 8: Pollutant EM variables (mg/m^2) present in the Agrimonia dataset with daily temporal resolution. For each EM variable, we use the IDW function as a spatial transformation and the PCHIP interpolation function to transform from monthly to daily temporal resolution.

daily time series of swine and bovine density for each monitoring station. See Table 9 for a summary of the livestock variables in the Agrimonia dataset.

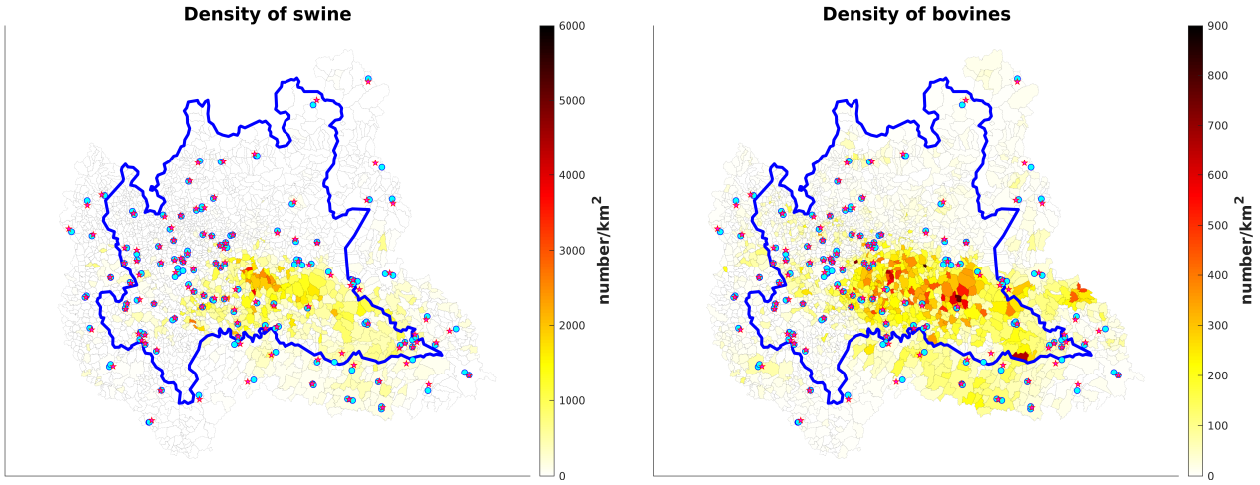


Figure 16: Swine (left) and bovines (right) density over the augmented Lombardy region on 31 December 2021. The AQ stations (cyan circle) are spread randomly over the studied area. Each station is associated with the municipality centroid to which it belongs (red start). Based on this, stations in the same municipality share the same municipality centroid so they have the same livestock time series.

Land

The data related to land cover and land and soil use are given by the ERA5-Land, CLC and SIARL datasets, respectively, describing the land and soil over time. Considering that high and low vegetation indices from ERA5-Land have a daily resolution over a spatial regular grid, we use the same IDW approach described for WE variables to create the daily time series associated with each AQ station. Information about land use is relatively stable over time. For the CLC dataset, we take the 2018 data

Variable name	Description
LI_pigs	Municipal density of swine related to AQ stations
LI_bovine	Municipal density of bovines related to AQ stations

Table 9: LI variables in the Agrimonia dataset with daily temporal resolution. Information on the number of swine and bovines is expressed as a density with respect to the municipal area: $number/km^2$. More details on the transformation process can be found in the metadata file named 'Metadata_Agrimonia.csv' available with the Agrimonia dataset Fassò, Rodeschini, et al. (2023a).

and keep it constant for the period from 2016 to 2021. For the SIARL dataset, the values are annual until 2019 (for more details on land use and land cover, see the Technical validation Section). The CLC provides categorical data in polygons while SIARL does so on a regular grid. In both cases, each AQ station is associated with the polygon or the cell to which it belongs. In this way, we obtain daily piecewise constant functions for land cover and soil use associated with each AQ station.

Variable name	Description	**Spatial transformation**	**Temporal transformation**	**Unit**
LA_hvi	One-half of the total green leaf area per unit horizontal ground surface area for high vegetation type	IDW	None	m^2/m^2
LA_lvi	One-half of the total green leaf area per unit horizontal ground surface area for low vegetation type	IDW	None	m^2/m^2
LA_land_use	CORINE Land Cover - Land use across 44 sectors	None	None	$Categ.$
LA_soil_use	SIARL Lombardy - Lombardy soil use across 21 sectors	None	None	$Categ.$

Table 10: LA variables in the Agrimonia dataset with daily time resolution. More details on the transformation process and label for categorical variables can be found in the metadata files named 'Metadata_Agrimonia.csv', 'Metadata_LA_CORINE_labels.csv' and 'Metadata_LA_SIARL_labels.csv', respectively, provided with the Agrimonia dataset Fassò, Rodeschini, et al. (2023a). $Categ.$ stands for Categorical.

4.4 Gridded dataset

The Agrimonia project aims to assess the impact of NH_3 emissions on air quality across the spatial domain. Since the set of covariate functions $\boldsymbol{x}(\cdot) : \mathcal{D} \to \mathbb{R}^b$ must be defined over the entire domain $\mathcal{D}$, it is necessary to construct covariates at the locations where predictions are to be made. However, explanatory variables are often available on different spatial supports, for example, at the

administrative-unit level, that is, on an irregular spatial lattice. This gives rise to the so-called change of support problem, where variables measured on different supports (e.g., irregular polygons versus point- or grid-based observations) cannot be directly compared or integrated.

In this section, we address this problem by constructing a gridded representation of the study variables, enabling consistent spatial alignment of air pollution and explanatory variables (e.g livestock density data). A precipice known as a change of support (Gelfand, Zhu, and Carlin, 2001). Specifically, we describe the construction of the Agrimonia Grid Covariates (AGC) dataset, which is defined on a regular grid of resolution $0.1^\circ \times 0.1^\circ$ (approximately 10×10 km), resulting in 1053 regular cells. The temporal coverage spans 2016–2021, with daily resolution. The AGC contains daily values of the main variables included in the Agrimonia dataset, covering the Lombardy region and surrounding areas (see Figure 15). This gridded dataset provides the foundation for (i) estimating pollutant concentrations at unknown spatial locations, and (ii) conducting scenario analyses to evaluate air quality control strategies and ammonia emission reduction policies. All the explanatory variables are harmonised to this common spatio-temporal support using the same transformation and aggregation techniques applied in the Agrimonia dataset, ensuring internal consistency across variables and facilitating joint modelling of environmental and agricultural factors.

Weather

Weather variables are provided with different spatial resolutions, see Table 1. To associate the weather variable with the cells of the AGC dataset, we used the Inverse Distance Weighted (IDW) interpolation using the 4 nearest cells, where the distances are computed from the centres of cells. If the distance is 0, cells share the same value.

Emission

Emission data are provided on a monthly temporal basis. Then they are harmonised daily using the PCHIP cubic spline. Since the centre of the CAMS cells coincides with the centre of the AGC dataset, no spatial transformations are needed. Figure 17 shows the seasonal maps for the NH_3 emission. The seasonal behaviour of the NH_3 emissions is determined by seasonal practices of agricultural activities responsible for strong variations in the NH3 emissions level. In particular, the NH_3 emissions are higher in the spring and summer periods because these periods coincide with widespread application of manure and synthetic fertilisers. Elevated temperatures in those months enhance ammonia volatilisation, increasing emissions. In contrast, winter sees much lower emissions

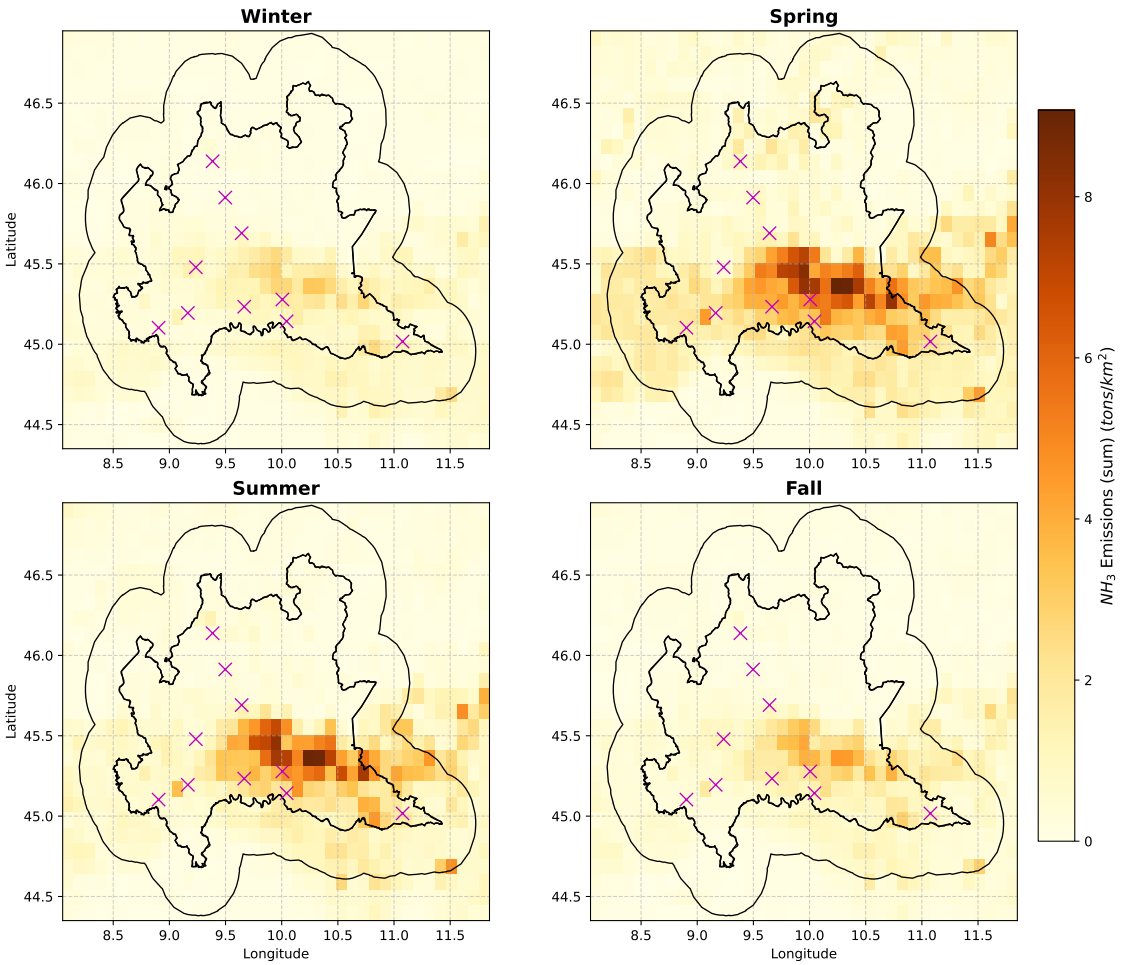


Figure 17: Season variability of the NH_3 emission in $(tons/km^2)$ in 2020. From left to right, from top to bottom, the seasons Winter, Spring, Summer and Fall are shown. NH_3 air quality network is depicted by magenta crosses.

due to restricted spreading practices and cooler conditions that suppress volatilisation (Colombi et al., 2024). It is worth noting that the air quality monitoring network is not specifically designed to monitor ammonia emissions. Indeed, most monitoring stations are located in areas with low NH_3 emission levels, and none are located where they reach the highest peaks, as shown in Figure 17. This raises potential issues of preferential spatial sampling bias and possibly results in underestimating the effect of the NH_3 concentration on the AQ. However, areas presenting high levels of ammonia emissions are characterised by high concentrations of PM_{10} and $PM_{2.5}$.

Livestock

The livestock variables include the density of pigs and bovines at the municipal level with a semi-annual temporal resolution. In the first stage, the data are harmonised from a semi-annual temporal basis to a daily temporal resolution, using the PCHIP interpolation spline. Since the data are associated with the centroid of the municipality, the data have to be reshaped from the area level to the AGC grid.

Let us consider the AGC grid output grid (0.1×0.1) such that $\mathcal{D} = \cup_i^K \mathcal{A}_i$ where $\mathcal{A}_i$ is the area associate with the i-cell, and $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ for each $i \neq j$ and $i, j = \{1, \dots, K\}$ where K is the total number of cells. Let $x(\mathcal{A}_i)$, or simply x_i, denotes the density level for the area $\mathcal{A}_i$. It can not be measured directly as the livestock density is provided only at the municipal level, says $\{x(\boldsymbol{s}_1), \dots, x(\boldsymbol{s}_n)\}$ with centroids at coordinates $\{\boldsymbol{s}_1, \dots, \boldsymbol{s}_n\}$. $\{x(\boldsymbol{s}_1), \dots, x(\boldsymbol{s}_n)\}$ is a realisation of the latent stochastic process $x(\boldsymbol{s})$ with $\boldsymbol{s} \in \mathcal{D}$ representing the true livestock density, which is continuous in space. The average density over the area $\mathcal{A}_i$ would be given by

$$x_i = \int_{s \in \mathcal{A}_i} x(\boldsymbol{s}) w(\boldsymbol{s}) d\boldsymbol{s}, \tag{4.2}$$

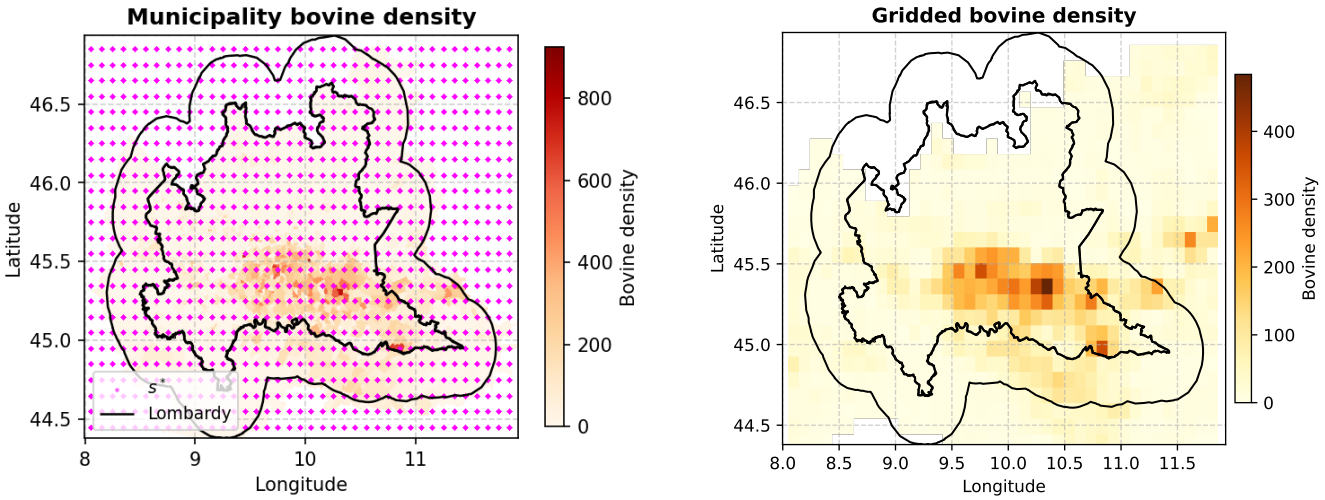
where $w(\boldsymbol{s})$ is a weight for a generic spatial point $s \in \mathcal{A}_i$ such that $\int_{s \in \mathcal{A}_i} w(\boldsymbol{s}) d\boldsymbol{s} = 1$. The stochastic integral (4.2) can be computed through Monte Carlo integration using a set of additional points, denoted by $\boldsymbol{s}^*$, which are the centroids of a regular grid covering the region of interest (see e.g. Lee and Shaddick (2010), G. Huang, Lee, and M. Scott (2015), Gelfand, Zhu, and Carlin (2001)). With this approach the density value x_i for area $\mathcal{A}_i$ is estimated through the following weighted mean:

$$x_i = \sum_{\boldsymbol{s} \in \mathcal{N}_i} x(\boldsymbol{s}) w(\boldsymbol{s}), \tag{4.3}$$

where $\mathcal{N}_i = \{\boldsymbol{s}^* : \boldsymbol{s}^* \cap \mathcal{A}_i = \boldsymbol{s}^*\}$ is the set of points inside area $\mathcal{A}_i$. $\boldsymbol{s}^* \in \mathcal{D}$ are drawn from a regular lattice of (0.01×0.01) degrees, ten times lower than the AGC grid, such that each $\mathcal{A}_i$ is discretised with 100 points, says $|\mathcal{N}_i| = 100$. The estimated mean values $\hat{x}(\boldsymbol{s}^*)$ are obtained as point-extraction from the municipal values. To compute the livestock density x_i associated with the area $\mathcal{A}_i$, we consider the simple mean with $w(\boldsymbol{s}) = 1/|\mathcal{N}_i|$. Figure 18 shows the change of support for the bovine density variable obtained by the discretisation (4.3). A more sophisticated approach can be adopted; for instance, by using a model-based framework, where the livestock density can be represented through an appropriate spatial model. Then estimate the quantity $\hat{x}(\boldsymbol{s}^*) = \mathrm{E}[x(s^*)|\{x(\boldsymbol{s}_1), \dots, x(\boldsymbol{s}_n)\}]$ can be computed using such model.

Land cover and Altitude

Land variables are provided from the Copernicus Land Monitoring Service (CLMS) as a shape file with a polygon associated with categorical variables which identify the type of land. To associate a categorical value indicating the type of land (urban, forest, sea, etc.) to the AGC cells $\mathcal{A}_i$, we apply a similar approach used for the livestock data based on discretisation (4.3). Since the expected outcome is a categorical value, the mode is used to get the most probable realisation among $s(\boldsymbol{s}^*)$. Future



(a) (b)

Figure 18: Left: bovine density at the municipal level. The magenta dots represent the centroids of the AGC regular grid (0.1×0.1) degrees, while the Lombardy boundaries are depicted by the solid black line. Right: resulting bovine density organised on a regular grid, computed by the discretisation of Eq. (4.2).

developments could adopt more sophisticated approaches, such as considering the distribution of each category as an indicator of the land type within each cell instead of just keeping the most probable one. To associate an altitude value with each cell of the AGC dataset, we used the Digital Elevation Model (DTM). To this end, the discretisation in Eq. (4.3) was used.

Chapter 5

Case study of analysis of livestock-related PM pollution based on spatio-temporal model

This chapter presents different spatio-temporal modelling approaches useful for relating air quality (AQ) and livestock activity. Section 1.1 has already discussed the public importance of modelling air quality for decision-makers, as well as the associated references. Moreover, Chapter 4 details the Agrimonia dataset, which includes information about livestock, meteorology, and air quality in the Lombardy region, Italy. Building on these motivations and data resources, this chapter explores how different spatio-temporal modelling strategies can be employed to link livestock activity with observed air quality pollution. Such approaches are crucial for capturing the interactions between agricultural practices, meteorological conditions, and air quality outcomes. In the last decades, the development of spatio-temporal models of air quality has been extensive, and several approaches, both statistical and based on machine learning, have been proposed to handle the spatial and temporal correlation. Statistical and machine learning models can provide valuable insights into air quality behaviour, including the identification of pollution sources, the factors influencing its variability, and the forecasting of future pollution levels under different scenarios, such as changes in emissions or weather patterns.

Furthermore, the combination of different modelling techniques may enhance results: by leveraging the strengths of each approach, decision-makers can gain a more comprehensive understanding of air pollution and make more informed choices for mitigation strategies. For this reason, part of this chapter is devoted to comparing three statistical and machine learning models with varying degrees of flexibility to explain the daily $PM_{2.5}$ concentrations in the Lombardy region. More precisely, we consider three predictive models with an increasing degree of flexibility: hidden dynamic geostatistical models (HDGM), generalised additive mixed models (GAMM), and the random forest spatio-temporal

kriging models (RFSTK). All three models employed in this study have been specifically developed to handle spatio-temporal data.

Building on the insights gained from the model comparison, we then develop a multivariate heteroskedastic spatio-temporal model designed to estimate $PM_{2.5}$ concentration surfaces from a finite set of observations. In particular, we propose to model heteroskedasticity in two main frameworks: *(i)*: time-varying unstructured flexible error variance, *(ii)* stochastic functions characterised by a lower number of degrees of freedom.

The proposed model is further applied in a scenario analysis to estimate the potential reductions in $PM_{2.5}$ concentrations resulting from decreases in ammonia (NH_3) emissions, thereby assessing the sensitivity of fine particulate matter to changes in ammonia. In addition, we demonstrate the flexibility of the proposed heteroskedastic framework by extending it to a multivariate setting that jointly models PM_{10} and $PM_{2.5}$, highlighting how skedastic functions can capture varying levels of uncertainty across pollutants.

The rest of this chapter is organised as follows. In Section 5.1, a preliminary analysis of $PM_{2.5}$ is done. In Section 5.2, we compare three statistical and machine learning models with varying degrees of flexibility to explain the daily $PM_{2.5}$ concentrations. In Section 5.3 we propose a heteroskedastic spatio-temporal model used to estimate the impact of NH_3 emissions on the concentration of $PM_{2.5}$ in the Lombardy region. Finally, Section 5.4 discusses the scenario analysis implementation, details the results and provides maps of the $PM_{2.5}$ change. Most of the material and methodological framework discussed in this chapter is drawn from Otto, Fusta Moro, et al. (2024) and Rodeschini, Fassò, et al. (2024), which provides the main reference for the analyses and approaches presented here.

5.1 Preliminary analysis

Section 4.1 already details the importance of the weather and emissions to explain the $PM_{2.5}$ concentrations. Based on that discussion, here we consider the period between 2016 and 2020 and the following variables: $PM_{2.5}$ measured at 45 ground stations (30 of which belong to the Lombardy region, while the remaining 15 are located in the neighbouring area), wind speed (average wind speed at 100 m), temperature (air temperature at 2 m), relative humidity (RH), boundary layer height (BLH, maximum daily air mixing layer height), high vegetation index (HVI, high vegetation abundance), sulphur dioxide emissions (SO_2), total ammonia emissions from agriculture (NH_3) and nitrogen oxide emissions (NO_x). In addition, we create new binary variables: Rain = 1 if the total daily precipitation exceeds the threshold of 1 mm, Urban = 1 if land use is classified as urban and a new categorical vari-

able Season which takes four categories: winter, spring, summer and autumn. Table 11 summarises the variables used in this study along with the main statistics.

Name	min	mean	max	std
$PM_{2.5}$ $[\mu g/m^3]$	1.00	21.58	182.00	16.63
Wind speed $[m/s]$	0.56	2.59	11.93	1.33
Temperature $[C°]$	-11.94	13.44	32.88	7.98
RH [%]	25.87	74.34	99.10	12.26
BLH $[m]$	27.38	1043.91	4420	556.77
HVI $[m^2/m^2]$	0.86	2.23	5.03	0.70
SO_2 $[mg/(m^2day)]$	0.05	3.34	45.06	5.53
NH_3 $[mg/(m^2day)]$	0.13	11.55	72.03	12.00
NO_x $[mg/(m^2day)]$	0.75	14.06	118.60	16.73
Rain	0	0.02	1	0.15
Urban	0	0.75	1	0.43

Table 11: Variables selected from the Agrimonia dataset (Fassò, Rodeschini, et al., 2023a) and main descriptive statistics. The Rain and Urban variables are binary. Season is not included because it is a categorical variable.

It is well known that air quality data are characterised by a seasonal pattern (Lonati and Cernuschi, 2020; Otto, Fusta Moro, et al., 2024) as show in Figure 1. Not surprisingly, there is a clear seasonality with higher concentrations in winter due to meteorological conditions resulting in reduced air circulation. The median $PM_{2.5}$ concentrations range between $10\mu gm^{-3}$ and $40\mu gm^{-3}$ across the year. Thus, throughout the year, the median concentration was beyond the threshold of $5\mu gm^{-3}$ considered hazardous by the World Health Organisation guidelines (WHO, 2021).

To explore a possible spatio-temporal correlation, we estimate a spatio-temporal variogram $\gamma(h, \tau)$ based on the sample variance of observations within certain distance ranges in space and time h and τ (see, e.g., Cressie and Wikle, 2015). Smaller values of the variogram for smaller distances indicate (short-term) statistical dependence. The spatio-temporal variogram of observed $PM_{2.5}$ concentrations is depicted in Figure 19. As expected, the variogram identifies an apparent correlation of the $PM_{2.5}$ concentrations across time and space. More precisely, the values of the variogram for the first temporal lags indicate a pronounced temporal dependence within the first 5-6 days, i.e., approximately one week. Furthermore, we observe a noticeable spatial dependence since the variogram increases with increasing spatial distances. It is important to note that this variation still includes spatial and temporal seasonalities and variations caused by exogenous factors.

The temporal dynamics were further analysed through the partial autocorrelation function (PACF), which highlights the first lag (1 day) as the most significant autocorrelated component, indicating that a model with first-order Markovian dynamics is appropriate. The autocorrelation at lag one is close to 0.8 for all stations and suggests that $PM_{2.5}$ is relatively stable in the atmosphere.

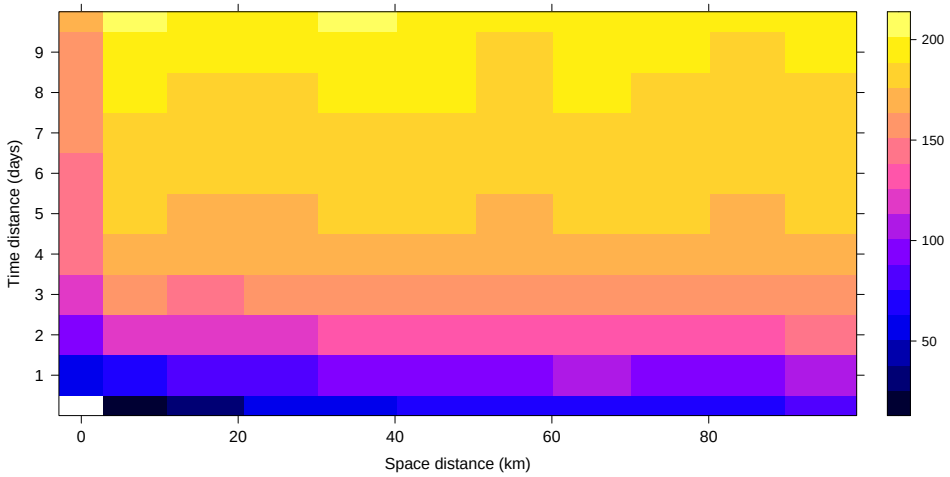


Figure 19: Spatio-temporal variogram computed on $PM_{2.5}$ daily observations at 45 stations from 2016 to 2020.

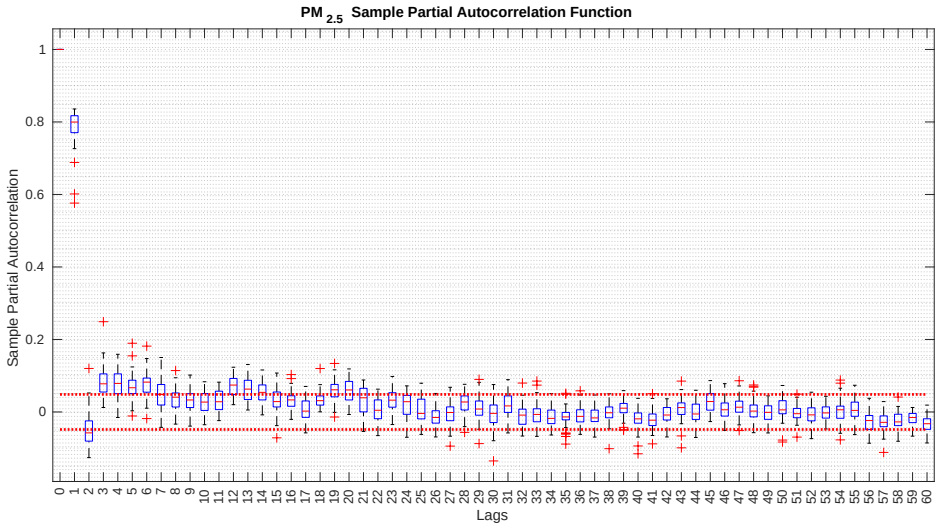


Figure 20: Partial Auto-Correlation Function (PACF) calculated by stations for the first 60 lags (days) and summarised through box plots. The i-th box plot summarises the partial autocorrelation calculated for each station at i-th lag.

The NO_x emissions follow human activities and, during the cold months, reach high levels of emissions due to endothermic-powered vehicles and domestic heating systems. Vice versa, the seasonal behaviour of the NH_3 emissions is determined by seasonal practices of agricultural activities responsible for strong variations in the NH_3 emissions level. As shown in Figure 17, the NH_3 emissions are higher in the spring and summer periods. Table 12 summarises the daily temperature average, emissions flow and $PM_{2.5}$ by season showing the composition regimes in the atmosphere. A special focus is devoted to NH_3 emissions. The spatial distribution of annual NH_3 emissions in 2020 is shown in Figure 17. The annual NH_3 emissions in 2020 aggregated by province are summarised in Table 13. In particular, the provinces of Brescia, Cremona, and Mantua are characterised by higher emissions. As discussed in Chapter 4 the air quality monitoring network is not specifically designed to monitor

Season	Temperature	NH_3	NO_x	SO_2	$PM_{2.5}$
Winter	3.88	5.21	19.97	4.11	35.87
Spring	12.67	17.45	11.33	2.98	16.54
Summer	23.23	14.71	10.45	2.80	12.19
Autumn	13.83	8.70	14.59	3.46	21.62

Table 12: Seasonal average of daily Temperature $[C^{\circ}]$, NH_3, NO_x and SO_2 emissions $[mg/m^2]$ and $PM_{2.5}$ concentrations $[\mu g/m^3]$.

ammonia emissions. Indeed, most monitoring stations are located in areas with low NH_3 emission levels, and none are located where they reach the highest peaks, as shown in Figure 17.

Province	Province code	NH_3 $[t/year]$
Bergamo	BG	7546
Brescia	BS	29524
Como	CO	1517
Cremona	CR	16892
Lecco	LC	615
Lodi	LO	6552
Mantua	MN	20336
Milan	MI	14013
Monza and Brianza	MB	714
Pavia	PV	8199
Sondrio	SO	4601
Varese	VA	1052

Table 13: Total NH_3 emissions by province in 2020.

5.2 Spatio-temporal modelling of $PM_{2.5}$ concentrations in Lombardy a comparative study

When measurements come from several spatial sites, it is convenient to account also the spatial correlation. Indeed, the spatio-temporal variogram shown in Figure 19 highlights spatio-temporal correlation structure among $PM_{2.5}$ concentrations. In our research to better understand which model better explains the relationship between predictor variables and daily $PM_{2.5}$, we compare three different classes of models: a geostatistical model, a semiparametric model and a machine learning technique. More precisely, hidden dynamic geostatistical models (HDGM), generalised additive mixed models (GAMM), and random forest spatio-temporal kriging (RFSTK) were utilised to describe the relationships between a large set of predictors and $PM_{2.5}$ concentrations. All three models employed in this study have been specifically developed to handle spatio-temporal data. HDGM incorporates a latent

variable to capture spatio-temporal dependence, while external factors are included linearly within the model. Conversely, GAMM allows for the nonlinear impact of exogenous predictors, which are estimated using splines. It incorporates spatio-temporal dependence by utilising a smoothing spline for spatial variation and a first-order autoregressive process for temporal dependence. Lastly, RFSTK employs a random forest (RF) to model the nonlinear effects of predictors and then a spatio-temporal kriging model to account for the possible spatio-temporal dependence.

These models are frequently applied in diverse areas. First, HDGM has been primarily used for air pollution studies (see, e.g Calculli et al. 2015; Finazzi, E. M. Scott, and Fassò 2013; Yaqiong Wang, Finazzi, and Fassò 2021). Notably, there are further applications in other fields, such as modelling bike-sharing data or coastal profiles (see Piter, Otto, and Alkhatib, 2022; Otto, Piter, and Gijsman, 2021). HDGM is a linear mixed effects model with a specific structure of the random effects capturing the spatio-temporal dynamics of environmental data, which are widely applied in diverse areas (see, e.g., Jiang and Nguyen 2007 for an overview). Second, GAMM has been employed in various areas, such as ecology (Knape, 2016; Kneib, Knauer, and Küchenhoff, 2011), psychology (Bono, Alarcón, and Blanca, 2021), economics (Fahrmeir and Lang, 2001) and psycholinguistics (Baayen et al., 2017). Third, needless to say, models based on decision trees have demonstrated their effectiveness in capturing complex patterns in different fields (see, e.g., Belgiu and Drăguţ 2016 for an overview in remote sensing, or Qi 2012 for bioinformatics), particularly in combination with kriging approaches, e.g., in environmental (Sekulić et al., 2020; Chen et al., 2019; Guo et al., 2015), or air pollution studies (Y. Liu, Cao, et al., 2018; Y. Liu, Zhao, et al., 2019). Furthermore, an interesting new approach is to use deep neural networks for the prediction and interpolation of spatial data (see Nag, Sun, and Reich, 2023; Daw and Wikle, 2023).

Focusing on Northern Italy, previous studies have successfully applied several statistical models to analyse air pollution scenarios, including generalised additive models (Bertaccini, Dukic, and Ignaccolo, 2012), Bayesian hierarchical models based on the stochastic partial differential equation approach (Cameletti, Lindgren, et al., 2013; Fioravanti et al., 2021), or random forests (Stafoggia et al., 2019). In a comparative study for Northern Italy, Cameletti, Ignaccolo, and Bande (2011) studied the effectiveness of different statistical models in a Bayesian framework. Machine learning algorithms, including random forests, are adept at capturing nonlinearities and interactions. Still, when applied to air quality modelling, the spatio-temporal nature of the phenomenon is often ignored (see, e.g., Fox, Ver Hoef, and Olsen (2020)). Consequently, the model's performance deteriorates, with worse outcomes than those obtained from Kriging, considered standard for modelling spatio-temporal phenomena. Kriging shows better results than random forest in Lombardy (Fusta Moro, Salis, et al.,

2022) and in the USA (Berrocal et al., 2020). On the other hand, machine learning algorithms outperform classical models if spatio-temporal dependence is not considered at all (Kulkarni et al., 2022). Lu, Cavieres, and Moraga (2023) compared geostatistical and machine learning models for NO_2 concentrations in Germany. Despite the limited number of studies comparing geostatistical and machine learning models, this subject is gaining increasing interest because the comparison provides valuable insights into the dynamics of the process, as we will illustrate below.

5.2.1 Spatio-temporal statistical models and machine learning techniques

From a theoretical perspective, consider the $PM_{2.5}$ concentrations as a spatio-temporal stochastic process $\{z(\boldsymbol{s},t) : \boldsymbol{s} \in \mathcal{D}, t = 1,2,\ldots,T\}$, as in Definition 9, where $\mathcal{D}$ is the spatial domain that contains a set of locations $\{\boldsymbol{s}_i : i = 1,\ldots,n\}$ (ground-level measurement stations) and the discrete temporal domain is $t = 1,\ldots,T$ (daily). Furthermore, we assume that $z(\boldsymbol{s},t)$ may be influenced by external variables. Using the SRE model (2.15), we can express the generic univariate spatio-temporal process as:

$$y(\boldsymbol{s},t) - S(\boldsymbol{s},t) = z(\boldsymbol{s},t) + \varepsilon(\boldsymbol{s},t), \tag{5.1}$$

where $S(\boldsymbol{s},t)$ are the large-scale components, $z(\boldsymbol{s},t)$ are small-scale spatio-temporal random effects, and $\epsilon(\boldsymbol{s},t)$ is the measurement error which is assumed to be a white noise process. All models, used for the comparison, share the same set of regressors, including weather conditions and livestock densities. For stability issues, the response variable is standardised using the logarithmic scale and the design matrix was normalised to have zero mean and unit variance. Based on an extensive review of the air pollution modelling literature, we carefully selected key weather variables, which are included in the fixed-effects term $S(\boldsymbol{s},t)$. Using the Wilkinson notation, $S(\boldsymbol{s},t)$ can be expressed as: $log(PM_{2.5}) = 1 + Altitude + Wind\_speed + Temperature + Precipitation + RH + BLH + LVI + HVI + Pigs + Bovine + Month.$

Hidden dynamic geostatistical model

The first model selected is the HDGM, which serves as the comparative analysis's starting point or baseline method. It is a widely applied geostatistical model, first considered by H.-C. Huang and Cressie (1996) as an extension of classical mixed-effects models for univariate spatio-temporal data. Calculli et al. (2015) extended the HDGM to multivariate data. This modelling approach proved particularly useful for air quality modelling (e.g., Fassò and Finazzi, 2011; Finazzi, E. M. Scott, and Fassò, 2013), as the comparative study of Cameletti, Ignaccolo, and Bande (2011) confirmed.

The HDGM is a two-level hierarchical model. The hierarchy is constructed by putting together two conditional submodels. At the first level, the observation variability is modelled by the measurement equation, which is essentially given by a regression component, a stochastic latent variable and a white noise error. The latent variable is defined at the second level of the hierarchy; it handles the spatio-temporal correlation through a Markovian process. The innovation term is a zero-mean Gaussian process with a spatial covariance function. Under these conditions, the HDGM can be cast into the SSM in Eq. (2.23)-(2.24).

Let $y(\boldsymbol{s}, t)$ be the response variable observed at site $\boldsymbol{s} \in \mathbb{S}^2$, where $\mathbb{S}^2$ is the surface of the sphere embedded in $\mathbb{R}^3$, and discrete time $t = 1, \dots, T$. The univariate HDGM is defined as follows:

$$\begin{aligned} y(\boldsymbol{s}, t) &= \boldsymbol{x}'(\boldsymbol{s}, t)\boldsymbol{\beta} + \alpha z(\boldsymbol{s}, t) + \epsilon(\boldsymbol{s}, t), \\ z(\boldsymbol{s}, t) &= f z(\boldsymbol{s}, t-1) + \eta(\boldsymbol{s}, t). \end{aligned} \tag{5.2}$$

where $S(\boldsymbol{s}, t) = \boldsymbol{x}'(\boldsymbol{s}, t)\boldsymbol{\beta}$ is the regression term with fixed effect coefficients $\boldsymbol{\beta}$; $\boldsymbol{x}(\boldsymbol{s}, t)$ is a $b \times 1$ vector of covariates that accounts for all exogenous effects; α is a scale parameter of the latent variable; the measurement error $\epsilon(\boldsymbol{s}) \sim N(0, \sigma^2_\epsilon)$ is independent in space and time; $z(\boldsymbol{s}, t)$ is a unit-variance Markovian scalar process ruled by the transition coefficient f; the innovation term $\eta(\boldsymbol{s}, t)$ is a zero-mean Gaussian process, $GP(0, \rho(\|\boldsymbol{s} - \boldsymbol{s}'\|; \theta))$, independent in time where ρ is a valid spatial correlation function, and $\|\boldsymbol{s} - \boldsymbol{s}'\|$ is the geodetic distance between $\boldsymbol{s}$ and $\boldsymbol{s}' \in \mathbb{S}^2$. The model parameters set is $\Pi = \{\boldsymbol{\beta}, f, \theta, \alpha, \sigma^2_\epsilon\}$, which is estimated using the maximum likelihood (ML) estimation through the EM algorithm. In this setting, the matrices are parameterised in such a way that the model depends on only a few parameters and still preserves the closed-form update formulas (Calculli et al., 2015), which makes the EM algorithm an efficient method for obtaining the MLE.

Generalised additive mixed model

Compared to generalised additive models (GAM, Hastie and Tibshirani 1987), generalised additive mixed models (GAMM) include a random-effects component to describe correlated response variables, such as time series, spatial or spatio-temporal data. It extends the HDGM by allowing for linear and nonlinear regressive effects in a GAM fashion, i.e., the response variable linearly depends on smooth functions of the predictors. To be precise, the large-scale components are given by

$$S(s, t) = \mathbf{x}'(s, t)\boldsymbol{\beta} + \sum_{j=1}^{b} \alpha_j(x_j(\boldsymbol{s}, t)), \tag{5.3}$$

with $\boldsymbol{x}'(s, t)\boldsymbol{\beta}$ being a linear parametric regression term of the first j-covariates with a parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_b)'$, including the intercept term β_0. Moreover, $\sum_{j=1}^{b} \alpha_j(x_j(\boldsymbol{s}, t)))$ is an

additive term with nonlinear functions $\alpha_j(\cdot) : \mathbb{R} \to \mathbb{R}$. $\alpha_j(\cdot)$ can be estimated along with the other model coefficients, e.g., as regression splines or penalised splines (Fahrmeir, Kneib, and Lang, 2004).

The small-scale effects of the GAMM are specified as a first-order autoregressive model for the temporal dependence and a smooth spatial surface for the spatial dependence, that is,

$$z(\boldsymbol{s}, t) \quad = \quad g(y(\boldsymbol{s}, t-1) - S(\boldsymbol{s}, t-1) - \eta(\boldsymbol{s}))\,, \tag{5.4}$$

where g is the parameter representing the temporal dependence, where zero indicates no temporal correlation. The spatial dependence is modelled as a smooth surface $\eta(\boldsymbol{s})$, which follows a GP with an exponential covariance function with the range parameter θ (Handcock and Wallis, 1994). This structure is identical to the spatial term of the random effects model in HDGM as given by equation (5.2). In our case, θ is estimated as proposed by Kammann and Wand (2003). The model estimation is computationally implemented in the package `mgcv` available in R (Wood, 2017).

Random forest spatio-temporal kriging

For the third approach, RFSTK, we increase the flexibility of the model in the large-scale component by considering a random forest (RF; Breiman (2001)) algorithm. In other words, the third hybrid model combines an RF for the large-scale component $S(\boldsymbol{s}, t)$ with that of a spatio-temporal kriging model for $z(\boldsymbol{s}, t)$. The idea traces back to the combination of random forests and kriging, the so-called random forest residual kriging, which, even if only considering spatial dependence, showed promising results compared to RF alone (e.g. L. Wang, Wu, and H.-B. Liu, 2019; Viscarra Rossel, Richard Webster, and Kidd, 2014). For spatio-temporal data, RFSTK has been considered to model air quality, again showing good performances (see Zhan et al., 2018; Shao et al., 2020).

RFs are constructed as an ensemble of multiple decision trees, making them highly versatile and robust for a wide range of predictive tasks. In a regression problem, the prediction of the large-scale effect is obtained by averaging across the predictions of n_{tree} decision trees. Each of these decision trees is trained or estimated from independent bootstrap samples $\boldsymbol{y}_j^*$ of the input data. These bootstrap samples are created through random resampling without replacement (recommended for dependent variables, see Strobl et al. 2008) from the original dataset. The predictions of these individual decision trees $\hat{y}_j(\boldsymbol{s}_i)$ are then averaged to produce the final prediction. This ensemble approach helps to improve the robustness and generalisation of the model, reducing the risk of overfitting. Thus, the large-scale model is given by

$$S(\boldsymbol{s}_i, t) = \frac{1}{n_{\text{tree}}} \sum_{j=1}^{n_{\text{tree}}} \hat{y}_j(\boldsymbol{s}_i), \tag{5.5}$$

for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. $\hat{y}_j(\boldsymbol{s}_i)$ is the prediction of the j-th decision tree at location $\boldsymbol{s}_i$. It is essential to emphasise that for regression trees within the random forest framework, the averaging process should be conducted for each region of interest in the covariate space. This means that the model considers the different regions of the input space and provides predictions tailored to the characteristics of each region. Random forests excel at handling complex, nonlinear relationships and are widely used in various applications, including classification and regression tasks, as well as feature selection and data exploration.

Geostatistical model versus machine learning

The geostatistical models allow for inference on the parameters. Thus, it is possible to understand if a covariate is helpful to explain the response variable. In particular, HDGM and GAMM use the MLE approach to estimate the model parameters, as well as the associated uncertainty. The GAMM model is more flexible than HDGM, allowing for nonlinear regressive effects, estimated as the predictors' smooth functions. The RF model is a well-known solution to handle the non-linear relationship, allowing both non-linear and full interaction regression terms. However, machine learning techniques are usually referred to as "black-box" because of their lack of interpretability. Furthermore, several approaches have been proposed in the literature to investigate the variable importance, the regression function, and the uncertainty associated with the prediction. However, when the number of predictors and the number of tuning hyperparameters (i.e., number of trees, number of random variables chosen for the split, number of observations within final leaves) increase, the computation burden becomes unfeasible.

How to compare statistical and machine learning models

The three different models are comparable in some commonalities, but are not entirely comparable due to different definitions. To compare the magnitude of the regressors among the models, we compare the estimated coefficients of fixed effects in HDGM and GAMM to the Variable Importance Factor (VIF) calculated within the RF. The VIF is based on the mean decrease accuracy algorithm, which ranks the regressors according to the prediction performance changes after the permutation of the selected regressor. On the other hand, to identify the relationship between the target variable and the regressors, we compare the slope of HDGM fixed effects to the spline of GAMM and the Partial Dependence Plot (PDP) within RF. In particular, the PDP simulates the relationships between predictor variables and predicted responses in the trained dataset. The PDP of the predicted responses is calculated on a subset of predictor variables X^s by marginalising the other variables X^c (complementary set). Assuming

	HDGM		GAMM		RFSTK	
	LS	FM	LS	FM	LS	FM
RMSE [μgm^{-3}]	11.92	1.814	11.456	8.468	8.245	5.361
MAE [μgm^{-3}]	0.068	−0.002	−0.002	−0.003	−0.129	−0.019
R^2	0.469	0.988	0.509	0.732	0.746	0.893

Table 14: In-sample performance of the three models assessed by RMSE (in μgm^{-3}), and the adjusted coefficient of determination R^2. The fit of the only large-scale component (LS) is compared to the full model (FM).

that each observation is equally likely, and the dependence between X^s and X^c and the interactions of X^c and X^s in responses is not strong, the partial dependence can be approximated by using observed predictor data as follows:

$$f^s(X^s) = \frac{1}{N}\sum_{i=1}^{N} f(X^s, X_i^c), \tag{5.6}$$

where N is the number of observations and $X_i = (X_i^s, X_i^c)$ is the i-th observation. In other words, PDPs show how the predicted outcome of the changes as a single predictor variable is varied while all other variables are held constant. Then, we will perform residual diagnostic checking to determine which model best removes the spatial and temporal correlation. In the end, for the assessment of prediction performance, we will use the Leave-One-Station-Out Cross-Validation (LOSO-CV) scheme with a focus on the behaviour of test error and checking for overfitting issues.

5.2.2 Comparative study results

To highlight the difference between the model fits, we compared the in-sample (i.e. using the entire data set) predictive performance of all the models, separately for the large-scale component (LS) and the full model (FM), including the space-time effects. The comparison was based on the root mean squared errors (RMSE), mean absolute errors (MAE), and the coefficient of determination R^2. The results are reported in Table 14. HDGM generally had better prediction capabilities and lower computational costs than the other models when considering the full model. For comparison, the RFSTK model had relatively good prediction capabilities but was computationally intensive, particularly when including the spatio-temporal kriging.

The residual distributions of the full model (FM) are symmetric and slightly leptokurtic, which means there is a greater chance of extreme values than a normal distribution, indicating that all models are less reliable at predicting extreme events. This is not surprising as they are designed to predict the mean level of the distribution and not for modelling the extremes.

The model uncertainties across time, shown in Figure 21, where the standard deviation of the

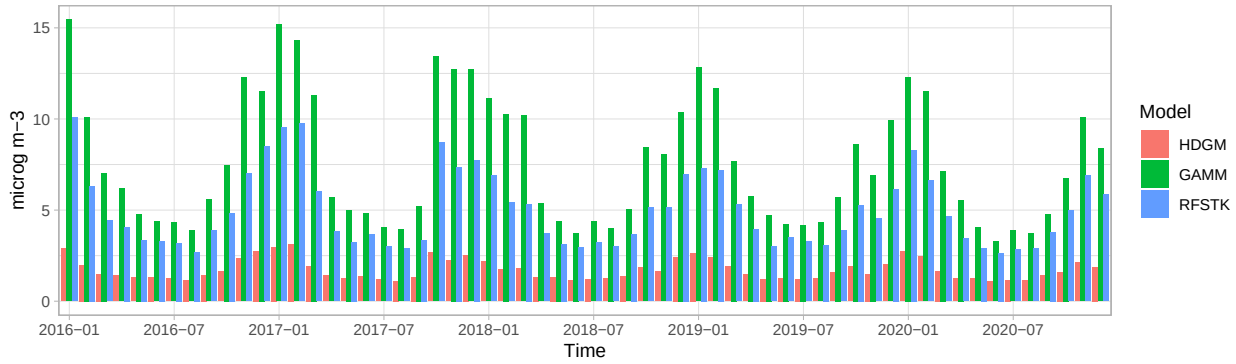


Figure 21: In-sample residual diagnostics for the three models. Standard deviation of residuals grouped by months for each model.

full model residuals is depicted for each month, reveal that all three models have varying uncertainty throughout the year. More precisely, the PM concentrations in the winter periods could be less accurately predicted than in the summer periods. Therefore, when the models are implemented for forecasting or scenario analysis, it is recommended to use a heteroscedastic model to not underestimate the prediction accuracy in the winter months (or overestimate for the summer period). For instance, spatio-temporal stochastic volatility models could be estimated for the residual process, as demonstrated in Otto et al. (2023) for much simpler mean models.

Furthermore, we investigated the spatial and temporal dependence of the residuals, estimating temporal autocorrelation functions (ACF) and spatio-temporal variograms. Different patterns were observed across the three models. While HDGM shows a small negative correlation at the beginning, indicating a slight overestimation of the temporal dependence, the spatial correlation is satisfactorily captured. On the contrary, GAMM leads to significantly lower temporal correlations in the residuals but does not capture the spatial dependence. The RFSTK is characterised by a more pronounced positive autocorrelation, consistent with the model specification that does not consider an autoregressive term. The spatio-temporal correlation was clearly captured.

In general, the informativeness of the in-sample results can be questionable due to the potential sensitivity of models to the training data or overfitting. To address this issue, we evaluated prediction performances within a cross-validation scheme. Three factors: randomness of partition, mutual independence of test errors, and independence between the training set and test set, are crucial considerations in the context of cross-validation. We employed the leave-one-station-out cross-validation (LOSOCV) scheme, which is a variation of the commonly used leave-one-out cross-validation (LOSOCV) approach applied in the spatio-temporal framework (e.g., Meyer et al. (2018) and Nowak and Welsh (2020)). For this method, a complete time series of a single station withheld is not used in the model's training but is used to evaluate the model's prediction performance. In

this way, the validation blocks are sufficiently large not to destroy the spatio-temporal dependence. In general, utilising a LOOCV provides an assessment of model performance but may overlook the temporal correlation of errors. By adopting a LOSOCV scheme, we gain the ability to examine the autocorrelation of errors at each station. This approach highlights distinct behaviours among stations, influenced, for instance, by factors like atmospheric stability, thereby accounting for temporal dependence. All stations within the Lombardy region were used for validation, except for the station "Moggio." It is located in the mountains with unique climatic conditions that are not well-represented by all other stations. After implementing the LOSOCV approach, we obtained prediction results from the full models for the 31 stations included in the validation process. It is worth noting that we applied the identical cross-validation scheme for each model so that the results are directly comparable.

The prediction performances assessed in the LOSOCV scheme in terms of mean squared errors (MSE), RMSE, MAE, and R^2 are summarised in Table 15. HDGM is confirmed to be the best model, but compared to the in-sample residuals in Table 14, the uncertainty is on a realistic level with an RMSE comparable to the other models. That is, the overfitting in the in-sample data did not affect the generalisation ability of the HDGM. This could be due to the linear structure in the large-scale component. While a more flexible model (e.g., random forest or artificial neural networks) could produce extremely bad predictions in areas of insufficient training data or overfitting, the linear structure of the HDGM regression term prevents us from obtaining such extreme predictions. Generally, we observe satisfactory prediction performances for all three model alternatives, with RFSTK and the GAMM approach being in second and third place, respectively. The substantial difference between the in-sample residuals from the model trained on the entire dataset and the errors from the LOSOCV scheme highlights the importance of validating prediction uncertainty through a cross-validation scheme, which accounts for the spatio-temporal nature of the data. Interestingly, the GAMM obtained a similar fit in terms of the coefficient of determination in both the in-sample case and the cross-validation. Thus, we would not overestimate the prediction capabilities when only looking at the in-sample fit. In contrast to the other models, the GAMM incorporates a relatively 'weaker' small-scale structure, capturing spatio-temporal dependence with a smooth spatial surface, serving as a relatively simple constant intercept. This design prevents overfitting and, thereby, results in similar training and testing performance.

Eventually, we compare the prediction performances for each station separately because we observed that the order of the best-fitting model is not homogeneous across space. Comparing the prediction performances for each station separately, we note that where there is a clear local variability, the HDGM tend to have a higher RMSE compared with the GAMM. The GAMM model, which

	MSE	RMSE	MAE	R^2
HDGM	35.373	5.948	4.376	0.879
GAMM	78.042	8.834	6.239	0.733
RFSTK	53.099	7.286	5.119	0.819

Table 15: Prediction performance indices evaluated with the LOSOCV scheme [$\mu g m^{-3}$]

does not include strong, time-varying spatial interactions, was able to capture this difference in air quality better than the HDGM and RFSTK models. In another situation where there is a clear spatial correlation among observations, the RFSTK model performed the worst while HDGM showed the best performance. The reason may lie in its poor ability to capture temporal dependence well. The concentrations of $PM_{2.5}$ at this station are very stable over time, and the RFSTK model does not fully capture this stability. Indeed, RFSTK underestimates $PM_{2.5}$ concentrations, especially in the winter periods, when the air circulation is at its lowest limit and temporal stability is the highest. These results highlight the need to select the model according to the local conditions carefully. The best model for one location may not be the best for another. Moreover, model averaging could additionally improve the predictions.

A comprehensive interpretation of each estimated model is detailed below, in the Section 5.2.3. What is interesting here is the comparison among the three models. Intriguing observations arise when comparing the results of the three models. First, the temporal variation plays an essential role in all three models, as evidenced by the magnitude of the temporal dummy coefficients in HDGM and GAMM (with January as the worst month for air quality) and 'month' receiving the highest ranking in the VIF of RFSTK. This suggests that weather variables alone are insufficient for capturing all seasonal variability, even considering the most important weather variables. Interestingly, temperature, which is usually negatively associated with $PM_{2.5}$, was found to be positively related to $PM_{2.5}$ conditional on the month. This is due to the opposing effect of temperature and monthly indicator variables. Furthermore, surprisingly, the significance tests in both HDGM and GAMM on livestock densities suggested that bovine livestock density is not related to $PM_{2.5}$ concentrations, while pig density is.

In Figure 22, we depict partial dependence plots (PDPs) for selected covariates of all three models with the corresponding regressors in the hot (summer and spring) and cold (autumn and winter) months as yellow and purple contours, respectively. PDPs allow us to compare the relationships identified within the large-scale of each model and highlight the transition from linear to more complicated relationships. Contrary to the marginal effects, these PDPs account for the typical range of the other predictors. This is accomplished by associating a fixed value of a regressor across all observations with the mean of predicted $PM_{2.5}$ concentrations. The mean of prediction is calculated for different fixed

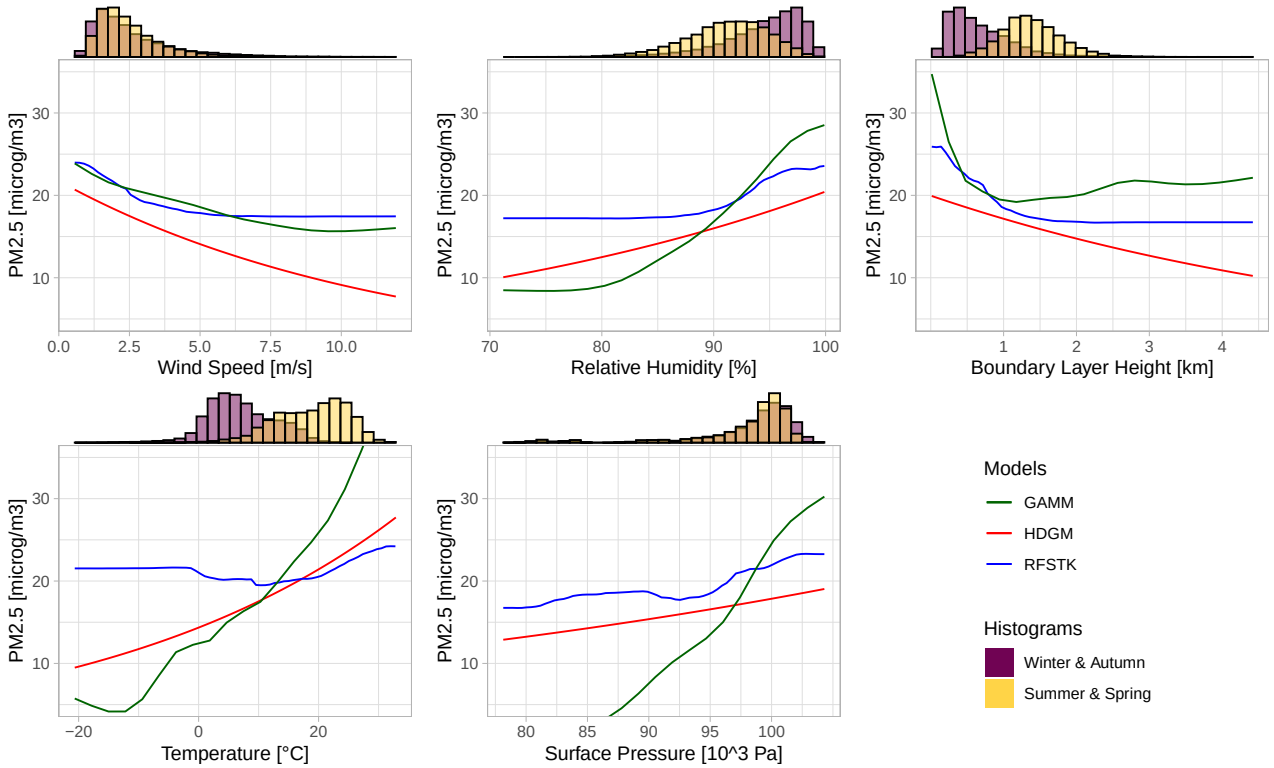


Figure 22: PDP calculated on the large scale of all three models. In yellow, the empirical histogram is for cold months, while in purple for hot months.

values of the regressor, typically moving from the minimum to the maximum on an equidistant grid. In other words, PDPs show how the predicted outcome of the changes as a single predictor variable is varied while all other variables are held constant. It is worth noting that all three models demonstrate similar trends, even though they have different levels of flexibility. For example, they all show that $PM_{2.5}$ concentrations are higher in cold periods. However, slight differences exist in the models' behaviour, especially for temperature. This suggests there may be nonlinear influences or interactions between temperature and other variables. For example, temperature may have a different impact on $PM_{2.5}$ concentrations in different altitudes or seasons. RFSTK can capture these interactions more effectively than the other two models, which is why its PDP is flatter. This suggests that RFSTK is better at capturing the complex relationships between $PM_{2.5}$ concentrations and other variables.

This finding has important implications for the development of air quality models. It suggests that machine learning (ML) techniques can be used to improve the performance and interpretability of classic geostatistical approaches, such as HDGM or GAMM. This is because ML techniques can identify nonlinearities and interactions that are difficult to identify using traditional methods. In the second step, the more interpretable models could include the nonlinear effects and interactions, e.g. GAMM. The comparison of PDPs also highlights the complementary nature of ML techniques and classic approaches. ML techniques are better at capturing complex relationships, while tradi-

tional approaches are easier to interpret and allow for straightforward uncertainty estimation. Thus, we advertise combining ML techniques and classic approaches for modelling and predicting PM concentrations.

5.2.3 Model comparison interpretation

In the following Section, we delve into the outcomes derived from our models, offering a comprehensive interpretation of each estimated model.

HDGM results

Table 16 summarises the estimated β parameters of the large-scale component of HDGM. Except for bovine density, all coefficients significantly differ from zero. The signs of the majority of the coefficients are consistent with our expectations: summer months are related to reductions of $PM_{2.5}$ concentrations (about -27/29 $\mu g/m^3$), every 1 m/s of wind speed is related to a decrease of 2 $\mu g/m^3$ of $PM_{2.5}$, and every 10 mm of precipitation are related to an expected decrease of 1.6 $\mu g/m^3$ of $PM_{2.5}$. Moreover, each degree Celsius increase in temperature is related to an increase of 0.5 $\mu g/m^3$ of $PM_{2.5}$, which seems counterintuitive at first glance. However, the temperature effect should be interpreted together with the monthly fixed effects. The relative humidity is positively related to $PM_{2.5}$, so high humidity levels (100%) are associated with an increase of 18 $\mu g/m^3$ with respect to extremely dry air. The maximum height of the boundary layer is negatively associated with $PM_{2.5}$; every increase of 1000 m is related to an expected decrease of 3 $\mu g/m^3$ of $PM_{2.5}$. Regarding the agricultural impact, we observe that the number of pigs in the territory is positively associated, and an increase of 1000 animals per km^2 corresponds to an expected increase of 5 $\mu g/m^3$ of $PM_{2.5}$. Both vegetation indices are negatively related to $PM_{2.5}$, while low vegetation (e.g. bushes) has a stronger effect than higher vegetation (e.g. trees). The small-scale effect of the HDGM is defined by a latent variable $z(\boldsymbol{s}, t)$, which has an autoregressive structure of order one and a Gaussian process with exponential covariance function. The spatial range parameter θ describes the decay of the exponential correlation function and is estimated to be equal to 0.79 °. Thus, there is a large spatial correlation (i.e., > 0.37) for surrounding stations in an area of 80 kilometres. The estimate of the temporal autoregressive parameter is equal to $\hat{g} = 0.72$. This indicates that the time series has low-frequency components with relatively gradual changes over time.

Variable	Coefficient	Std. Err.	t-value	p-value
(Intercept)	39.626	1.399	28.332	< 0.0001
February	-7.222	1.316	5.490	< 0.0001
March	-15.298	1.474	10.378	< 0.0001
April	-21.569	1.552	13.898	< 0.0001
May	-27.608	1.596	17.295	< 0.0001
June	-27.662	1.666	16.602	< 0.0001
July	-27.493	1.690	16.270	< 0.0001
August	-29.316	1.681	17.436	< 0.0001
September	-27.611	1.628	16.956	< 0.0001
October	-20.335	1.555	13.080	< 0.0001
November	-16.602	1.493	11.117	< 0.0001
December	-8.526	1.349	6.318	< 0.0001
Altitude	-0.007	0	20.180	< 0.0001
WE_wind_speed_100m_mean	-1.946	0.048	40.520	< 0.0001
WE_tot_precipitation	-159.416	7.145	22.310	< 0.0001
WE_temp_2m	0.505	0.037	13.595	< 0.0001
WE_rh_mean	0.186	0.008	24.188	< 0.0001
WE_blh_layer_max	-0.003	0	19.628	< 0.0001
LI_pigs_v2	0.005	0.001	9.288	< 0.0001
LI_bovine_v2	0.003	0.002	1.147	0.251
LA_lvi	-4.441	0.225	19.698	< 0.0001
LA_hvi	-0.494	0.141	3.507	< 0.0001

Table 16: Estimated coefficients of the large-scale component of the HDGM model

GAMM Results

The estimated coefficients of our second model, the GAMM, are presented in Table 17. In the first section of the table, the estimated coefficients for the linear part of the model (including the monthly fixed effects) are reported, while the second part summarises the effective degrees of freedom of the nonlinear effects as a measure of complexity/non-linearity. Compared to HDGM, the monthly fixed effects are slightly smaller, indicating that the seasonal variation is better captured by the weather variables, which enter the model nonlinearly. Complex nonlinear relationships with large degrees of freedom characterised the smooth terms of the penalised thin plate regression splines. All of them are significant except for the density of bovine. To illustrate the difference between the linear effects in the HDGM and the nonlinear effects in GAMM, we depict some selected regression splines in Figure 23 along with the estimated linear functions of the HDGM. These curves correspond to the marginal effect of variables, neglecting the spatio-temporal correlation (i.e., without the influence of neighbouring sites). In general, we observe a similar tendency for both models. An exceptional notice would be the height of the boundary layer, which has a negative effect for up to 500 kilometres, and afterwards, the effect changes to be positive. By contrast, the effect is negative for the HDGM, which

A: Linear effects	Coefficient	Std. Err.	t-value	p-value
(Intercept)	37.3052	0.4541	82.1536	< 0.0001
February	-7.2787	0.3559	-20.4496	< 0.0001
March	-11.6052	0.4169	-27.8335	< 0.0001
April	-14.6751	0.5173	-28.3679	< 0.0001
May	-20.5547	0.6097	-33.7147	< 0.0001
June	-22.0366	0.6978	-31.5800	< 0.0001
July	-23.2506	0.7287	-31.9076	< 0.0001
August	-24.5223	0.6813	-35.9908	< 0.0001
September	-22.7519	0.5743	-39.6143	< 0.0001
October	-18.9167	0.4687	-40.3605	< 0.0001
November	-15.8692	0.4005	-39.6225	< 0.0001
December	-11.4140	0.3541	-32.2379	< 0.0001
Altitude	-0.0023	0.0012	-1.9378	0.0526
B: Nonlinear effects	edf	Ref.df	F-value	p-value
WE_temp_2m	8.5851	8.5851	267.3673	< 0.0001
WE_tot_precipitation	7.2891	7.2891	210.3836	< 0.0001
WE_rh_mean	8.0060	8.0060	602.4805	< 0.0001
WE_wind_speed_100m_mean	6.5219	6.5219	214.0715	< 0.0001
WE_blh_layer_max	8.8552	8.8552	278.6934	< 0.0001
LI_pigs_v2	7.9422	7.9422	7.2958	< 0.0001
LI_bovine_v2	3.7865	3.7865	1.2020	0.2962
LA_hvi	8.4121	8.4121	23.1429	< 0.0001
LA_lvi	7.4114	7.4114	82.1857	< 0.0001
Longitude,Latitude	27.7758	32.0000	19.0023	< 0.0001

Table 17: Estimated coefficients of the large-scale component of the GAMM. Linear relationships (A) are identified by the β coefficients, while for nonlinear relationships (B), the complexity of the curve is described by the effective degrees of freedom (edf).

mimics the effect in the areas where most observations are located. The grey contour lines in Figure 23 additionally illustrate the estimated kernel density of the couple ($PM_{2.5}$, WE_regressor). We note that the confidence intervals around the fitted curves are smaller in areas with higher density.

The GAMM smooth spatial surface $\eta(\boldsymbol{s})$ is displayed in Figure 24. This smoothing spline $\eta(\boldsymbol{s})$ captures the spatial dependence and identifies correlated areas. Our study shows higher concentrations of $PM_{2.5}$ in the area of Como and the area of Brescia, while in the southwest, corresponding to the Ligurian border, lower concentrations. Furthermore, the estimated range parameter of the exponential covariance function is $\hat{\theta} = 1.16°$, which corresponds to approximately 110 km. Hence, it is in a similar range to the other two models. Furthermore, the autoregressive parameter is estimated as $\hat{g} = 0.67$, similar to HDGM and indicating a medium temporal persistence across one day. In this sense, the models show similar spatio-temporal dynamics as the HDGM.

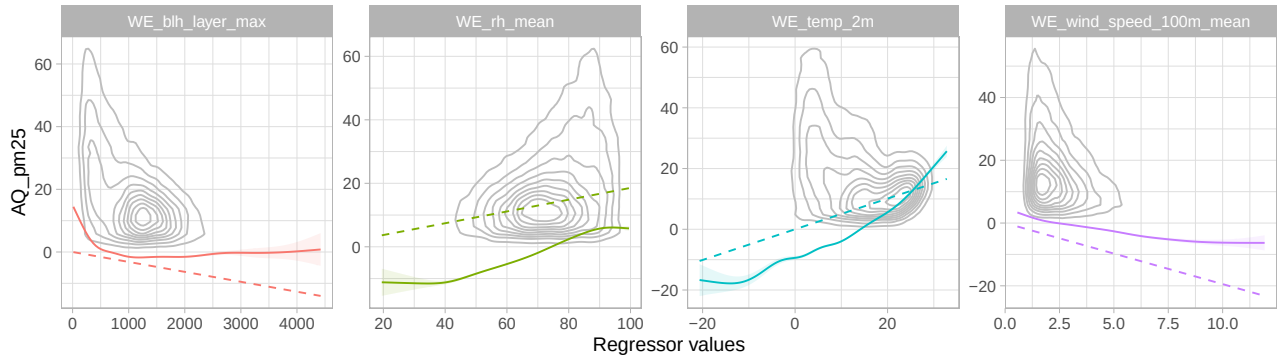


Figure 23: Regression splines (continuous lines), including their 95% confidence intervals (coloured shadows), and HDGM regression coefficients (dashed lines) for relevant weather regressors (i.e., boundary layer height, relative humidity, temperature, and wind speed). Grey contour lines represent the estimated two-dimensional kernel densities of the $PM_{2.5}$ concentrations and the corresponding regressors.

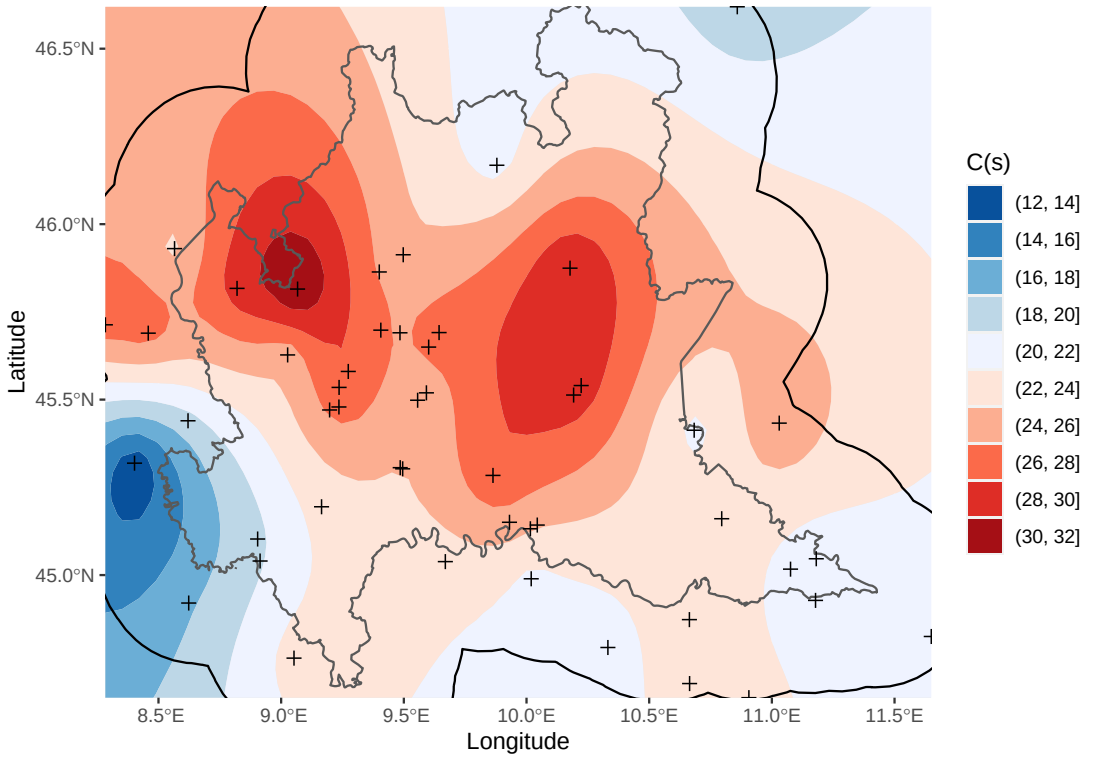


Figure 24: Estimated smoothing spline $\hat{\eta}(s)$ of GAMM, which corresponds to the $PM_{2.5}$ predicted on a regular grid using the large scale of GAMM, where all regressors are set to 0. Stations are marked with a black cross, Lombardy boundaries are shown in grey, and the black line marks the surrounding buffer zone.

RFSTK results

The interpretability of the RFSTK model can be challenging due to its intricate nature. However, the variable importance factor (VIF) can help to identify the most important variables. To determine the VIF, the mean decrease accuracy technique is employed, which assesses the variables' importance by measuring the increase in MSE (IncMSE) when the values of the regressors are permuted (Breiman, 2001). The results are depicted in Figure 25. Like the previous models, the random forest cannot fully capture the seasonality by the included (weather) regressors, as we can see by the high importance of the monthly effects. The most important variables after the month are the height of the boundary layer, the temperature and the (low) vegetation index. The bovine density is the least important variable, and indeed, it was not significantly different from zero for the other two models.

For the RFSTK, the spatio-temporal dependence is captured by considering a spatio-temporal Gaussian process with a separable exponential covariance function for space and time. The corresponding variogram model is fitted to the RF residuals. For our analysis, we obtained exponential decay parameters of $\hat{\theta}_t = 0.78$ days and $\hat{\theta}_s = 0.48°$ for time and space, respectively; the latter corresponds to approximately 50 km. The spatial range parameters are smaller than HDGM because the large scale (RF) explains more variation than the linear large-scale term of HDGM, as shown in Table 14.

5.3 Heteroskedastic spatio-temporal modelling of $PM_{2.5}$

Figure 21 in Section 5.2 shows that all three models have varying uncertainty throughout the year and highlights the need for a heteroskedastic model to not underestimate the prediction accuracy in the winter months. Furthermore, the hidden dynamic geostatistical model (HDGM) showed the best performance in the cross-validation study in $PM_{2.5}$ modelling in Lombardy with an average RMSE of 6.31 $\mu g/m^3$ since it has strong capabilities to handle space and temporal correlations (see Table 15).

Building on these insights, here we focus on developing a multivariate heteroskedastic spatio-temporal model designed to estimate $PM_{2.5}$ concentration surfaces from a finite set of observations. In particular, we propose to model heteroskedasticity in two main frameworks: *(i)*: time-varying unstructured flexible error variance, *(ii)* stochastic functions characterised by a lower number of degrees of freedom. We implement the two frameworks within a geostatistical approach by extending the HDGM to account for heteroskedasticity in the error terms. In both cases, this results in a lower-dimensional parameter dimension efficiently handled by the expectation-maximisation (EM) algorithm.

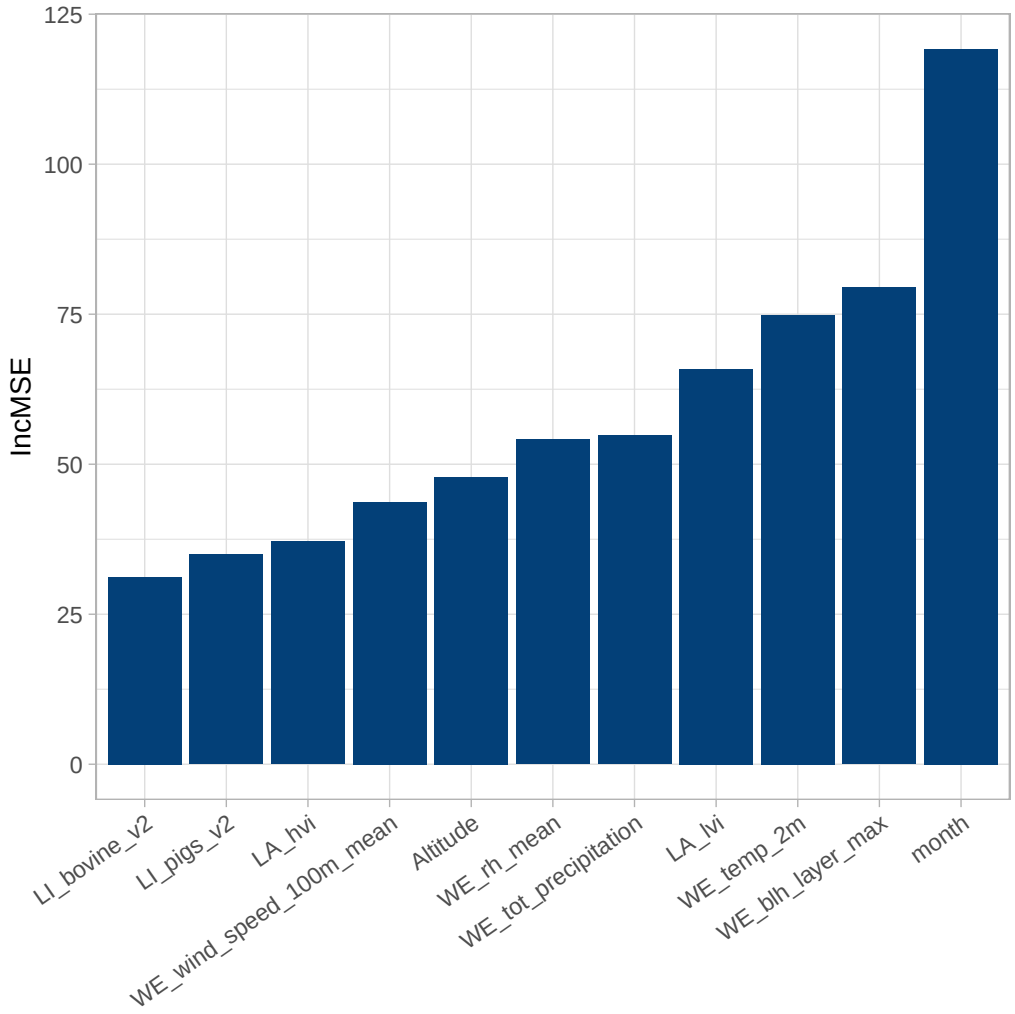


Figure 25: Variable importance factor (measured as IncMSE) for the 11 selected features in the large-scale component of the RFSTK.

In spatio-temporal models, heteroskedasticity has various facets and may refer to time, space, data heterogeneity or a mixture of the three. In addition, the skedastic function may be deterministic or stochastic. An example of data heterogeneity arises in data fusion problems where the data vector elements are obtained by different sensors or processes. See, e.g., Smith, Yan, and Cowles (2008). Considering the temporal dimension, we may model stochastic heteroskedasticity using the well-known approach introduced by the Nobel Prize winner Robert F. Engle Engle (1982). It is based on conditioning the error variance on the past and has resulted in a large suite of GARCH-like models developed in the last decades. See, e.g., Francq and Zakoian (2019). Also, the approach based on a deterministic skedastic function is used in environmental statistics, for example, a seasonal variance, see e.g. J. S. Benth, F. E. Benth, and Jalinskas (2007) and Fassò (2023). Considering the spatial dimension, deterministic spatial skedastic functions have been used extensively. For example, spatial skedastic functions are considered by Mur and Angulo (2009) in spatial econometrics models and by Hulshof and Umana (2023) in ecology. Engle and Bollerslev's generalised conditionally

heteroskedastic approach has been recently introduced to spatial econometrics by Otto, Schmid, and Garthoff (2018). For a review and developments, see Otto, Doğan, and Taşpınar (2023) and the references therein.

5.3.1 The multivariate heteroskedastic Hidden Dynamic Geostatistical Model

To understand the relationship between predictors and the response variable, taking into account spatial and temporal correlation, we propose a heteroskedastic extension of the HDGM, already introduced in Section 5.2.1. Let us consider the HDGM (5.2), which is a two-level hierarchical model. The hierarchy is constructed by putting together two conditional submodels. The heteroskedastic extension concerns only the measurement error $\boldsymbol{\epsilon}$; indeed, the observation variability is modelled by the measurement equation, which is essentially given by a regression component, a stochastic latent variable and an error with time-varying variance.

Let $\boldsymbol{y}(\boldsymbol{s}, t)$ be the p-variate response variable observed at site $\boldsymbol{s} \in \mathbb{S}^2$, where $\mathbb{S}^2$ is the surface of the sphere embedded in $\mathbb{R}^3$, and discrete time $t = 1, \dots, T$. The multivariate heteroskedastic HDGM is defined as follows:

$$\begin{aligned} \boldsymbol{y}(\boldsymbol{s}, t) &= \boldsymbol{x}'(\boldsymbol{s}, t)\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{z}(\boldsymbol{s}, t) + \boldsymbol{\epsilon}(\boldsymbol{s}, t), \\ \boldsymbol{z}(\boldsymbol{s}, t) &= \boldsymbol{F}\boldsymbol{z}(\boldsymbol{s}, t-1) + \boldsymbol{\eta}(\boldsymbol{s}, t). \end{aligned} \tag{5.7}$$

The heteroskedastic p-variate measurement error $\boldsymbol{\epsilon}(\boldsymbol{s}, t) = (\epsilon_1(\boldsymbol{s}, t), \dots, \epsilon_p(\boldsymbol{s}, t))$ where $\epsilon_i(\boldsymbol{s}, t) \sim N(\mathbf{0}, \boldsymbol{\sigma}_i^2(t))$ is independent in space and time with time-varying variance $\boldsymbol{\sigma}_i^2(t)$. Furthermore $\epsilon_i(\boldsymbol{s}, t) \perp \epsilon_j(\boldsymbol{s}, t)$ for all $i \neq j$ with $i, j = 1, \dots, p$. Considering the frameworks *(i)*, we propose to model time-varying variance using time-varying unstructured error variance, say $\boldsymbol{\sigma}_{\epsilon,i}^2 = (\sigma_{\epsilon,i}^2(0), \dots, \sigma_{\epsilon,i}^2(T))$. Specifically for the multivariate settings, we consider the following parametrisation: $\boldsymbol{F} = \text{diag}(f_1, \dots, f_p)$ and $\boldsymbol{A} = \text{diag}(a_1, \dots, a_p)$. The p-variate spatial innovation term $\boldsymbol{\eta}(\boldsymbol{s}, t) = (\eta_1(\boldsymbol{s}, t), \dots, \eta_p(\boldsymbol{s}, t))$ is a p-variate GP with zero mean and multivariate covariance function $\boldsymbol{C}(\boldsymbol{s})$ given by the LCM (2.10), that is $C(\boldsymbol{s}) = \boldsymbol{V}\rho(\boldsymbol{h}; \vartheta)$ where $\boldsymbol{V}$ is the coregionalisation matrix while $\rho(\boldsymbol{h}; \vartheta)$ is a correlation function parametrised by ϑ, and $\boldsymbol{h} = \|\boldsymbol{s} - \boldsymbol{s}'\|$. The model parameters set $\Pi = \{\boldsymbol{\beta}, \text{diag}(\boldsymbol{F}), \vartheta, \text{diag}(\boldsymbol{A}), \boldsymbol{V}, \boldsymbol{\sigma}_{\epsilon,1}^2, \dots, \boldsymbol{\sigma}_{\epsilon,p}^2\}$ is estimated using the maximum likelihood (ML) estimation through the EM algorithm.

Functional heteroskedastic HDGM

Estimating daily heteroskedastic variance through traditional structural models, like model (5.7), often leads to a large number of parameters, which, although updated in closed-form (as shown later

in (5.10)), can become computationally burdensome. To address this, a more flexible approach involves modelling the heteroskedasticity using a functional approach, where the variance is represented as a smooth function over time. Specifically, let the multivariate functional variance be denoted as $\boldsymbol{\sigma}^2_\epsilon(t) = (\sigma^2_{\epsilon,1}(t), \ldots, \sigma^2_{\epsilon,p}(t))$, and let $\boldsymbol{\Psi}_i(t) = (\psi_{1,i}(t), \ldots, \psi_{R,i}(t))$ be a $R \times 1$ where $\psi_{i,j}(t)$ is a known basis function $\psi_{i,j} : \mathbb{R}^+ \to \mathbb{R}$ evaluate at time t. The definition of such bases usually requires the choice of a reference set of points called knots $v_i \in \mathcal{V}_R$ for $i = 1, \ldots, R$ that span the considered domain $\mathcal{T}$. The finer the grid of knots, the more locally flexible the resulting function. The error variance for the i-components is modelled as

$$log(\sigma^2_{\epsilon,i}(t)) = \boldsymbol{\Psi}_i(t)\boldsymbol{c}_{\epsilon,i}. \tag{5.8}$$

This method offers interpretability and ensures a continuous variance structure over time. More compactly, Eq. (5.8) can be written as $log(\boldsymbol{\sigma}^2_\epsilon(t)) = \boldsymbol{\Phi}(t)\boldsymbol{c}_\epsilon$ where $\boldsymbol{\Phi}(t) = \text{blockdiag}(\boldsymbol{\Psi}_1(t), \ldots, \boldsymbol{\Psi}_p(t))$ and $\boldsymbol{c}_\epsilon = (\boldsymbol{c}'_{\epsilon,1}, \ldots, \boldsymbol{c}'_{\epsilon,p})'$ is the vector of unknown parameters.

The model parameters set is $\Pi = \{\boldsymbol{\beta}, \text{diag}(\boldsymbol{F}), \vartheta, \text{diag}(\boldsymbol{A}), \boldsymbol{V}, \boldsymbol{c}_{\epsilon,1}, \ldots, \boldsymbol{c}_{\epsilon,p}\}$, which is estimated using the maximum likelihood (ML) estimation through the EM algorithm. To simplify the notation, hereinafter we consider $\Psi_1 = \cdots = \Psi_p$. Different basis functions Φ_i, such as B-splines and Fourier series, can be used to capture the underlying variability. B-splines, in particular, are commonly applied due to their flexibility and local control. By using such basis functions, the model becomes more interpretable and efficient, while retaining the capacity to capture the variance dynamics. The HDGM relies on the Kalman smoother algorithm, which can handle a time-varying measurement error variance. Therefore, the heteroskedastic extension can be obtained straightforwardly by rewriting the EM formulas.

5.3.2 Matrix representation of heteroskedastic HDGM & estimation formulas

Using the standard notation, already introduced in Chapter 2 and considering the multivariate outcome $\boldsymbol{y}(\boldsymbol{s}, t) = (y_1(\boldsymbol{s}, t), \ldots, y_p(\boldsymbol{s}, t))$, let us suppose that the variable $y_i(\boldsymbol{s}, t)$ is observed at each spatial location $\mathcal{S}_i = \{\boldsymbol{s}_{1,i}, \ldots, \boldsymbol{s}_{m_i,i}\}$ for $i = 1, \ldots, p$, and let $m = \sum_{i=1}^p m_i$. Let $\boldsymbol{y}_t = \left(\boldsymbol{y}_1(\mathcal{S}_1, t), \ldots, \boldsymbol{y}_p(\mathcal{S}_p, t)\right)'$ be the $m \times 1$ vector of the response variable at time t. With these assumptions, the data model for (5.7) can be cast into the SSM where the observations at time t follow the equations:

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{z}_t + \boldsymbol{\epsilon}_t, \\ \boldsymbol{z}_t &= \boldsymbol{F}\boldsymbol{z}_{t-1} + \boldsymbol{\eta}_t, \end{aligned} \tag{5.9}$$

with $\boldsymbol{X}_t = \text{blockdiag}\left(\boldsymbol{X}_{1,t}, \ldots, \boldsymbol{X}_{p,t}\right)$ where $\boldsymbol{X}_{i,t} = \text{stack}(\boldsymbol{x}(\boldsymbol{s}_{i,1}, t), \ldots, \boldsymbol{x}(\boldsymbol{s}_{i,m_i}, t))$ and $x(\cdot, t) : \mathbb{R}^2 \to \mathbb{R}^b$ is the known function of the covariates. Vectors $\boldsymbol{z}_t$ and $\boldsymbol{\eta}_t$ are defined similarly

to $\boldsymbol{y}_t$. Moreover, considering the unstructured error variance, $\boldsymbol{\epsilon}_t$ is a p-dimensional random noise vector with independent components such that, $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\epsilon,t})$ where $\boldsymbol{\Sigma}_{\epsilon,t}$ is given by $\boldsymbol{\Sigma}_{\epsilon,t} = \text{blockdiag}(\boldsymbol{\Sigma}_1(t), \dots, \boldsymbol{\Sigma}_p(t))$ where $\Sigma_i(t) = \sigma^2_{\epsilon,i}(t)\boldsymbol{I}_{m_1}$ and $\boldsymbol{I}_{m_i}$ is the identity matrix of order m_i. The distribution of the latent variable at $t = 0$, $\boldsymbol{z}_0 \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. If all parameters Π are known, the unobserved temporal process $\boldsymbol{z}_t$ in the model (5.9) is estimated for each time t through the Kalman smoother technique, which is initialised with the condition $\boldsymbol{z}_0$. The Kalman smoother algorithm handles time-varying parameters in a natural way. Alongside the Kalman smoother output $\boldsymbol{z}_t^T$, $\boldsymbol{P}_t^t$ and $\boldsymbol{P}_{t,t-1}^T$, the quantities $\boldsymbol{S}_{10} = \sum_{t=1}^T \boldsymbol{z}_t^T(\boldsymbol{z}_{t-1}^T)' + \boldsymbol{P}_{t,t-1}^T$, $\boldsymbol{S}_{00} = \sum_{t=1}^T \boldsymbol{z}_{t-1}^T(\boldsymbol{z}_{t-1}^T)' + \boldsymbol{P}_{t-1}^T$ and $\boldsymbol{S}_{11} = \sum_{t=1}^T \boldsymbol{z}_t^T(\boldsymbol{z}_t^T)' + \boldsymbol{P}_t^T$ are introduced as the so-called EM second moments and used in the E-step to evaluate the sufficient statistics. As a result, starting with initial value $\Pi^{(0)}$, the updating formulas are:

$$\boldsymbol{\beta}^{(m+1)} = \Big[\sum_{t=1}^T (\boldsymbol{X}_t)'(\boldsymbol{\Sigma}_{\epsilon,t}^{(m)})^{-1}\boldsymbol{X}_t\Big]^{-1}\Big[\sum_{t=1}^T (\boldsymbol{X}_t)'(\boldsymbol{\Sigma}_{\epsilon,t}^{(m)})^{-1}(\boldsymbol{y}_t - \boldsymbol{A}^{(m)}\boldsymbol{z}_t^{T,(m)})\Big], \tag{5.10}$$

$$\alpha_i^{(m+1)} = \sum_{t=1}^T tr\Big[\boldsymbol{z}_{i,t}^{T,(m)}(\boldsymbol{\Sigma}_{i,t}^{(m)})^{-1}\Big(\boldsymbol{y}_{i,t} - \boldsymbol{X}_{i,t}\boldsymbol{\beta}_i^{(m)}\Big)\Big]\Big(\sum_{t=1}^T tr\Big[(\boldsymbol{\Sigma}_{i,t}^{(m)})^{-1}\Big(\boldsymbol{z}_{i,t}^{T,(m)}(\boldsymbol{z}_{i,t}^{T,(m)})' + \boldsymbol{P}_{i,t}^{T,(m)}\Big)\Big]\Big)^{-1}, \tag{5.11}$$

$$\boldsymbol{\mu}_0^{(m+1)} = \boldsymbol{z}_0^{T,(m)}, \tag{5.12}$$

$$\boldsymbol{\Sigma}_0^{(m+1)} = \boldsymbol{P}_0^{T,(m)}, \tag{5.13}$$

$$f_i^{(m+1)} = tr(\boldsymbol{S}_{i,10}^{(m)})tr(\boldsymbol{S}_{i,00}^{(m)})^{-1}, \tag{5.14}$$

$$\boldsymbol{\theta}^{(m+1)} = \underset{\theta}{\text{argmax}}\Big(Tlog|\boldsymbol{\Sigma}_\eta^{(m)}| + tr\Big[(\boldsymbol{\Sigma}_\eta^{(m)})^{-1}\Big(\boldsymbol{S}_{11}^{(m)} - \boldsymbol{S}_{10}^{(m)}\boldsymbol{F}^{(m)} - \boldsymbol{F}^{(m)}\boldsymbol{S}_{10}'^{(m)} + \boldsymbol{F}^{(m)}\boldsymbol{S}_{00}^{(m)}\boldsymbol{F}^{(m)}\Big)\Big]\Big), \tag{5.15}$$

$$(\sigma^2_{\epsilon,t})^{(m+1)} = \frac{1}{n}tr(\boldsymbol{\Lambda}_t^{(m)}), \tag{5.16}$$

where $\boldsymbol{\Lambda}_t^{(m)} = \text{E}_{\Pi^m}(\boldsymbol{e}_t|\boldsymbol{Y})\,\text{E}_{\Pi^m}(\boldsymbol{e}_t|\boldsymbol{Y})' + \text{Var}_{\Pi^m}(\boldsymbol{e}_t|\boldsymbol{Y})$ with $\text{E}_{\Pi^m}(\boldsymbol{e}_t|\boldsymbol{Y}) = \boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}^{(m)} - \boldsymbol{A}^{(m)}\boldsymbol{z}_t^T$ and $\text{Var}_{\Pi^m}(\boldsymbol{e}_t|\boldsymbol{Y}) = \boldsymbol{A}^{(m)}\boldsymbol{P}_t^T\boldsymbol{A}^{(m)}$.

Note that the updating formulas (5.10) to (5.16) are similar to those provided by standard EM estimation for the SSM, see Amisigo and Van De Giesen (2005). The main exception lies in the parameters of the measurement equation, which undergo modification due to the introduction of

time-varying variance. Specifically, Eq. (5.16) results in an average over space for each time t. The updating formulas for $\boldsymbol{\beta}$ and $\boldsymbol{A}$, in Eqs. (5.10) and (5.11), respectively, take into account the time-varying variance $\boldsymbol{\Sigma}_{\epsilon,t}$. The derivation of the updating formula Eq. (5.16) - (5.15) are detailed in Rodeschini, Fassò, et al. (2024).

In case we adopt the functional model, i.e. the case *(ii)*, we consider the assumption in Eq. (5.8) where the variance of the measurement error is given by

$$\boldsymbol{\Sigma}_{\epsilon,t} = \text{blockdiag}\left(\exp(\boldsymbol{\Psi_1}(t)\boldsymbol{c}_{\epsilon,1}\boldsymbol{I}_{m_1}, \ldots, \exp(\boldsymbol{\Psi_p}(t)\boldsymbol{c}_{\epsilon,p}\boldsymbol{I}_{m_p}\right). \tag{5.17}$$

The log-likelihood restricted to the vector $\boldsymbol{c}_\epsilon = (\boldsymbol{c}_{\epsilon,1}, \ldots, \boldsymbol{c}_{\epsilon,p})$ at the iteration m became $Q(\boldsymbol{c}_\epsilon, \boldsymbol{c}_\epsilon^{(m)}) = \sum_{t=1}^{T} log|\boldsymbol{\Sigma}_{\epsilon,t}^{(m)}| + \sum_{t=1}^{T} tr(\boldsymbol{\Sigma}_{\epsilon,t}^{(m)})^{-1}\boldsymbol{\Lambda}_t$. The MLE solution is obtained by maximise the $Q(\boldsymbol{c}_\epsilon, \boldsymbol{c}_\epsilon^{(m)})$ function, and in this case, the solutions are given by the root of the derivatives with to respect the $\boldsymbol{c}_\epsilon$ which are given by:

$$\boldsymbol{c}_{\epsilon,i} = \text{solve}\left\{\sum_{t=1}^{T} m_i \boldsymbol{\Psi}_i(t)\boldsymbol{c}_{\epsilon,i} - \text{tr}(\boldsymbol{\Psi}_i(t)\exp(\boldsymbol{\Psi}_i(t)\boldsymbol{c}_{\epsilon,i})\boldsymbol{I}_{m_i}\boldsymbol{\Lambda}_{i,t})\right\}, \tag{5.18}$$

where solve is a numerical procedure to find the root of the non-linear system in the form $g(\boldsymbol{x}) = \boldsymbol{0}$ where $g(\boldsymbol{x})$ is a scalar function. The algorithm used to solve the non-linear system in Eq. (5.18) is the trust-region-dogleg algorithm and is implemented in the 'fsolve' routine of MATLAB.

Prediction & block kriging

Given the maximum likelihood estimate $\hat{\boldsymbol{\Pi}}$, the plug-in predictions at new locations $\boldsymbol{s}_0 \in \mathcal{D}$ and times $t = 1, \ldots, T$ are defined as

$$\hat{\boldsymbol{y}}(\boldsymbol{s}_0, t) = \boldsymbol{X}(\boldsymbol{s}_0, t)\hat{\boldsymbol{\beta}} + \boldsymbol{A}\boldsymbol{z}_t^T(\boldsymbol{s}_0), \tag{5.19}$$

with associated prediction variance

$$\Sigma_{\hat{\boldsymbol{y}}}(\boldsymbol{s}_0, t) = \boldsymbol{A}\,\text{Var}\left[\hat{\boldsymbol{z}}_t^T(\boldsymbol{s}_0)\right]\boldsymbol{A}', \tag{5.20}$$

where $\boldsymbol{z}_t^T(\boldsymbol{s}_0)$ denotes the smoothed state process evaluated at location $\boldsymbol{s}_0$, and $\text{Var}[\boldsymbol{z}_t^T(\boldsymbol{s}_0)]$ is given by the corresponding Kalman smoother covariance $\boldsymbol{P}_t^T(\boldsymbol{s}_0)$. From a computational perspective, obtaining $\boldsymbol{z}_t^T(\boldsymbol{s}_0)$ requires running the Kalman filter and smoother using the estimated parameter set $\hat{\boldsymbol{\Pi}}$, by augmenting the state vector to account for the additional location $\boldsymbol{s}_0$. The computational complexity of this procedure is $O(m^3T)$ for each new site $\boldsymbol{s}_0$, where m is the dimension of the state vector. From a statistical standpoint, the estimate $\boldsymbol{z}_t^T(\boldsymbol{s}_0)$ can be interpreted as a kriging predictor, since the Kalman filter recursively computes conditional expectations under Gaussian assumptions,

thereby providing the best linear unbiased prediction at new space–time locations.

Given the pointwise predictions $z_t^T(\boldsymbol{s}_0)$ with $\boldsymbol{s}_0 \in \mathcal{D}$ and $t = 1, \dots, T$, one can make an inference on the underlying continuous spatial process. For visualisation purposes, it is often convenient to discretise the domain $\mathcal{D}$ into a collection of non-overlapping cells, for example, a regular grid of pixels. More generally, let us consider a partition of the domain into disjoint sets such that $\mathcal{D} = \bigcup_{i=1}^{K} \mathcal{A}_i$ with $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ for all $i \neq j$. This leads to the so-called *change of support* problem, where we are interested in predicting the process not at a point $\boldsymbol{s}_0$, but rather over a block $\mathcal{A}_i$. Formally, the block average of the process is defined as

$$z(\mathcal{A}_i) = \frac{1}{|\mathcal{A}_i|} \int_{\mathcal{A}_i} z(\boldsymbol{u})\, d\boldsymbol{u}, \quad |\mathcal{A}_i| > 0,$$

where $|\mathcal{A}_i|$ denotes area of the block. The support of $z(\mathcal{A}_i)$ is then the set $\mathcal{A}_i$. The optimal linear predictor of $z(\mathcal{A}_i)$, in the sense of minimising the mean squared prediction error, is given by

$$\mathrm{E}[z(\mathcal{A}_i) \mid \boldsymbol{z}] = \frac{1}{|\mathcal{A}_i|} \int_{\mathcal{A}_i} \mathrm{E}[z(\boldsymbol{u}) \mid \boldsymbol{z}]\, d\boldsymbol{u}. \tag{5.21}$$

Since this integral is typically intractable, it is approximated by a discrete average over a finite number of prediction points within $\mathcal{A}_i$:

$$\mathrm{E}[z(\mathcal{A}_i) \mid \boldsymbol{z}] \approx \frac{1}{R} \sum_{r=1}^{R} \mathrm{E}[z(\boldsymbol{s}_r) \mid \boldsymbol{z}], \qquad \boldsymbol{s}_r \in \mathcal{A}_i. \tag{5.22}$$

If $\mathcal{A}_i$ is sufficiently small, a common simplification is to approximate the block average by the conditional expectation at the centroid,

$$\mathrm{E}[z(\mathcal{A}_i) \mid \boldsymbol{z}] \approx \mathrm{E}[z(\boldsymbol{s}_r) \mid \boldsymbol{z}], \tag{5.23}$$

where $\boldsymbol{s}_r$ denotes the centroid of $\mathcal{A}_i$.

5.4 Scenario analysis of livestock-related $PM_{2.5}$ pollution based on a heteroskedastic spatio-temporal model

Several studies have shown that, within Europe, reducing NH_3 emissions is among the most effective strategies for mitigating $PM_{2.5}$ levels in both summer and winter (Megaritis et al., 2013). In this section, we aim to apply a statistical spatio-temporal model to conduct a "what-if" scenario analysis, evaluating the reduction in $PM_{2.5}$ concentrations resulting from different ammonia emission reduction

scenarios in the Lombardy region.

In particular, Megaritis et al. (2013) used five distinct control strategies, with a particular focus on testing a 50% reduction in gaseous emissions (SO_2, NH_3, NO_x and anthropogenic volatile organic compounds) to assess concentration sensitivity to emissions. The findings indicated that, in the majority of European regions, reducing NH_3 emissions during winter and summer periods proves more efficacious in lowering overall $PM_{2.5}$ levels compared with reductions in other gas precursors. Moreover, De Meij et al. (2009) showed that a targeted 50% reduction in ammonia emissions results in a decrease of total $PM_{2.5}$ levels by up to 2.4 $\mu g/m^3$ in the Lombardy region. Additionally, Pozzer et al. (2017) demonstrated that the nonlinear behaviour of the sulphate-nitrate-ammonia system influences the efficacy of $PM_{2.5}$ control strategies.

This section presents the 'what-if' scenario analysis approach and provides the formula to compute the impact uncertainty; in Section 5.4.1 we will formalise the scenarios considered and outline the scenario analysis to assess the $PM_{2.5}$ changes due to NH_3 emissions. Section 5.4.2 addresses the application of the heteroskedastic HDGM to the observed daily $PM_{2.5}$ in the Lombardy Region between 2016 and 2020. In our study, we are not interested in modelling and interpreting the skedastic function per se, but as a nuisance parameter needed to make correct inferences on spatial maps and aggregated results. For these reasons, we opt for a flexible, unstructured time-varying error variance applied to the HDGM. Section 5.4.3 discusses the scenario analysis implementation, details the results and provides maps of the $PM_{2.5}$ change. The results are summarised in Section 5.4.4.

5.4.1 Overview

We consider the 'what-if' scenario analysis approach: 'What would happen if the NH_3 reductions were fully implemented at time 0 (1st January 2016)?'. Under the 'what-if' approach we are interesting in comparing the predicted $PM_{2.5}$, $\hat{y}(\boldsymbol{s}, t)$, here based on the observed NH_3 emissions with a prediction, $\hat{y}^r(\boldsymbol{s}, t)$ where the NH_3 emissions are reduced by a factor r. The daily $PM_{2.5}$ change $\Delta_y(\boldsymbol{s}, t) = y(\boldsymbol{s}, t) - y^r(\boldsymbol{s}, t)$ and its variance are estimated by

$$\Delta_{\hat{y}}(\boldsymbol{s}, t) = \hat{y}(\boldsymbol{s}, t) - \hat{y}^r(\boldsymbol{s}, t)$$
$$\text{Var}(\Delta_{\hat{y}}(\boldsymbol{s}, t) - \Delta_y(\boldsymbol{s}, t)) = \Delta'_{\boldsymbol{x}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} \Delta_{\boldsymbol{x}} + \text{Var}(\epsilon_t - \epsilon_t^r)$$

where $\Delta_{\boldsymbol{x}} = \boldsymbol{x}(\boldsymbol{s}, t) - \boldsymbol{x}^r(\boldsymbol{s}, t)$; $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ is the estimate of the $\hat{\boldsymbol{\beta}}$ covariance matrix; ϵ_t and ϵ_t^r are the measurement errors of both predictions that are independent and with the same variance. The average of the daily $\Delta_{\hat{y}}(\boldsymbol{s}, t)$ and its variance can be extended over the block $\mathcal{A}_i \subset \mathcal{D}$ and time interval $\mathcal{I} \subset \{1, \ldots, T\}$ by considering the block average (5.21) and it's discretise Monte Carlo

computation (5.22). Specifically for this problem, the block average and its variance are given by

$$\bar{\Delta}_{\hat{y}}(\mathcal{A}_i, \mathcal{I}) = \frac{1}{|\mathcal{N}_i||\mathcal{I}|} \sum_{\boldsymbol{s}^* \in \mathcal{N}_i} \sum_{t \in \mathcal{I}} \Delta_{\hat{y}}(\boldsymbol{s}^*, t) \tag{5.24}$$

$$\mathrm{Var}[\bar{\Delta}_{\hat{y}}(\mathcal{A}_i, \mathcal{I}) - \bar{\Delta}_{y}(\mathcal{A}_i, \mathcal{I})] = \bar{\Delta}_{\boldsymbol{x}}(\mathcal{N}_i, \mathcal{I})' \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} \bar{\Delta}_{\boldsymbol{x}}(\mathcal{N}_i, \mathcal{I}) + \frac{2}{|\mathcal{I}|^2 |\mathcal{N}_i|^2} \sum_{t \in \mathcal{I}} \hat{\sigma}^2_{\epsilon,t} \tag{5.25}$$

where $\mathcal{N}_i$ is the set of support points used to compute the Monte Carlo integration, and $|\cdot|$ denotes the cardinality. $\hat{\sigma}^2_{\epsilon,t}$ is the estimate variance of the measurement error; and $\bar{\Delta}_{\boldsymbol{x}}(\mathcal{N}_i, \mathcal{I})$ is the average of the design matrix. Eq. (5.25) provides the standard deviation (std) used to assess the estimate change uncertainty in the next session.

5.4.2 Estimated model

Given the extensive literature on air quality in the Lombardy region, we rely on the following studies to select the covariates for the heteroskedastic HDGM (5.7). Furthermore, the comparison carried out in Section 5.2 provided detailed insights into the impact of weather conditions on $PM_{2.5}$ formation in Lombardy. Fassò (2023) explored the seasonality of pollutant concentration, revealing that the sensitivity of $PM_{2.5}$ to NH_3 and NO_x varies with season. Based on these findings, we adopt a univariate heteroskedastic HDGM with a regression term specified as follows:

$$\begin{aligned} PM_{2.5} \sim\ & (Intercept) + Windspeed + Temperature + RH + Rain + BLH + Urban + HVI \\ & + SO_2 + NO_x + NH_3 + (NO_x + NH_3) : Season \end{aligned} \tag{5.26}$$

while the spatial correlation function of the innovation term for the model (5.7) is defined by the exponential function as $\rho(\|\boldsymbol{s} - \boldsymbol{s}'\|; \theta)) = \exp\{-\|\boldsymbol{s} - \boldsymbol{s}'\|/\theta\}$.

To avoid numerical issues, the response variable and covariates are standardised. Hence, the estimated parameters in the current subsection refer to this standardised setup. Table 21 summarises the estimated coefficients $\boldsymbol{\beta}$ of the linear regression model used for the large-scale component. All coefficients except the intercept are statistically significant. As expected, the role of NH_3 is most prominent in the winter, when NH_3 plays a limiting role in the formation of $PM_{2.5}$. The estimated parameters of the latent variable and the uncertainty associated with them are $\hat{\theta} = 2.26°$ ($std < 0.06°$), $\hat{f} = 0.79$ ($std < 0.01$) and $\hat{\alpha} = 0.18$ ($std < 0.01$). Note that the latent process is stationary and that the magnitude of g indicates that the $z(\boldsymbol{s}, t)$ change smoothly over time. Finally, Figure 26 shows the daily $\hat{\sigma}^2_{\epsilon,t}$, which is higher in winter than in summer, highlighting the need for a heteroskedastic model.

We validate the model by considering its performance in cross-validation (CV). In particular, we

use the leave-one-station-out CV scheme (LOSOCV; Nowak and Welsh (2020)). To do this, we only use the 30 stations belonging to Lombardy in the LOSOCV procedure, while the remaining 15 stations are used in the training set only. As a result, the model in-sample root mean square error (RMSE) is $3.85 \mu g/m^3$ while the CV-RMSE is $5.91 \mu g/m^3$, which improves the CTM's performance (Veratti et al., 2023).

Name	β	std	$\lvert t \rvert$	p-value
(Intercept)	-0.02	0.05	0.46	0.64
Wind speed	-0.08	0.00	23.87	0
Temperature	-0.21	0.02	10.63	0
RH	0.09	0.01	17.39	0
Rain	-0.01	0.00	2.87	4.10E-03
BLH	-0.07	0.00	14.87	0
Urban	-0.08	0.01	15.24	0
HVI	-0.03	0.00	6.88	0
SO2	-0.05	0.01	9.13	0
NO_x	0.10	0.01	10.95	0
NO_x:Winter	0.03	0.01	3.35	7.96E-04
NO_x:Summer	-0.02	0.01	3.88	1.06E-04
NO_x:Spring	-0.02	0.01	4.06	4.96E-05
NH_3	0.11	0.01	8.96	0
NH_3:Winter	0.09	0.01	12.07	0
NH_3:Summer	-0.07	0.01	6.34	0
NH_3:Spring	-0.10	0.01	7.87	0

Table 18: Estimate of the fixed effect coefficients of the model (5.7). |t| is the absolute value of the t-statistic.

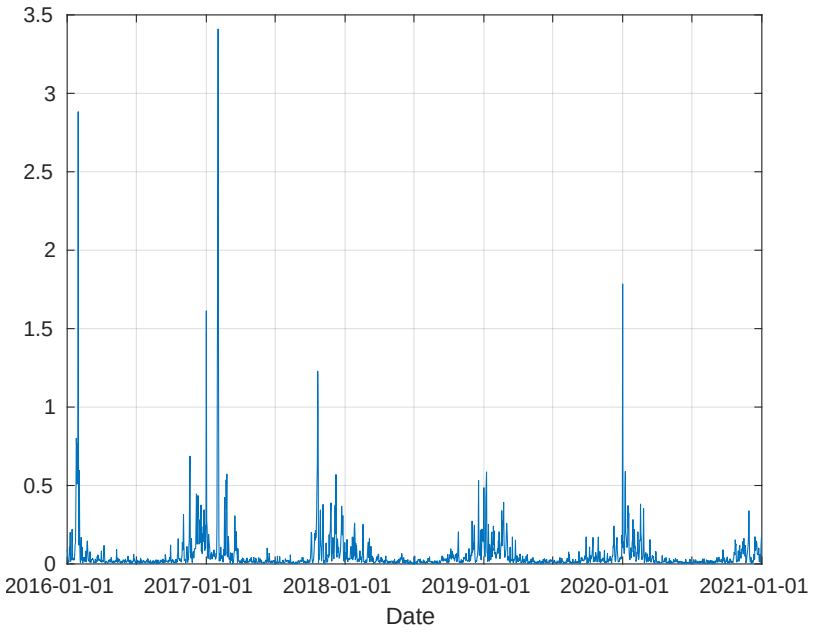


Figure 26: Estimated error variance, $\hat{\sigma}^2_{\epsilon,t}$.

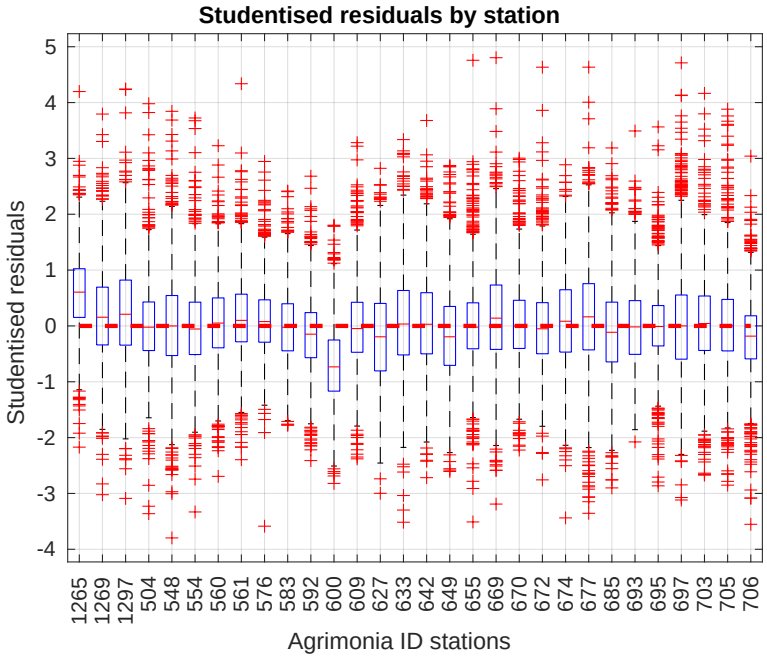


Figure 27: Studentised residuals grouped by station. The outliers (red cross) are identified by the Whisker length based on the first and third quartiles.

Since the variances of the residuals at different times t are different, in Figure 27, the distributions of the studentised residuals, computed as $e_t/\hat{\sigma}_{\epsilon,t}$, grouped by station are shown through boxplots. Their distributions are approximately centred around zero and moderately non-Gaussian. The residuals are further investigated through the autocorrelation function (ACF). We compute the ACF by station and then summarise the autocorrelation coefficients for all stations by using the boxplot representation, as shown in Figure 28. The ACF is generally not significant except at lag one, where there is a weak autocorrelation. This is not further considered because its influence in Eq. (5.25) is negligible.

5.4.3 Scenario analysis

According to Veratti et al. (2023), we propose two different scenarios for NH_3 emissions reduction: *(i)* scenario named PRIA that accounts for a reduction of 26% in NH_3 emissions; *(ii)* scenario named Strong that is characterised by a reduction of 50% in ammonia emissions. The PRIA scenario is based on the PRIA plan (in Italian '*Piano Regionale degli Interventi per la qualità dell'Aria*'), which identifies the actions needed to reduce ammonia emissions by 26% in Lombardy. We consider the 'what-if' scenario analysis approach: 'What would happen if the NH_3 reductions were fully implemented at time 0 (1st January 2016)?'. Considering the distribution of NH_3 emissions shown in Figure 17, we assess the scenario analysis only in non-forested areas below 640 m altitude. Since the response is log-standardised, we back-transform the results to the original units ($\mu g/m^3$).

For each scenario, thanks to the discretisation (5.22) and the approximation (5.23), we map the

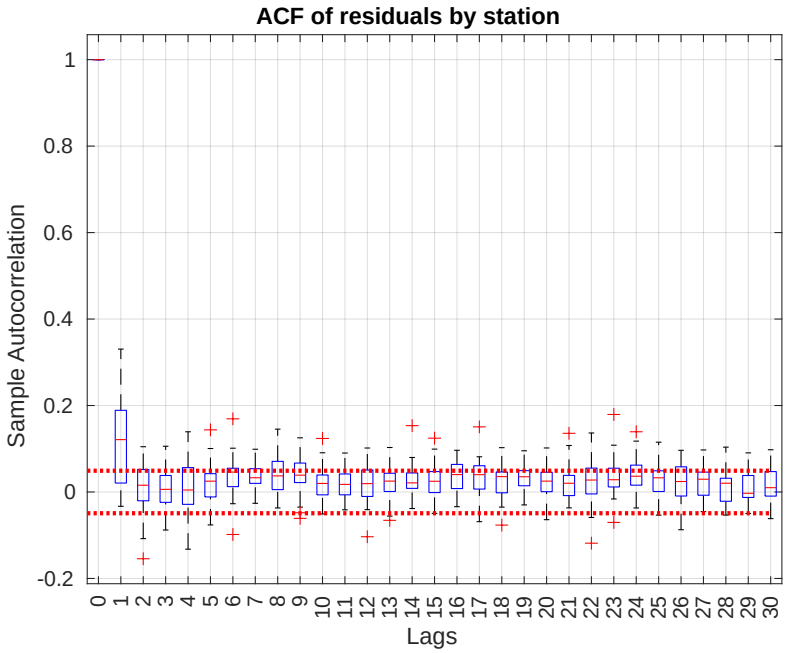
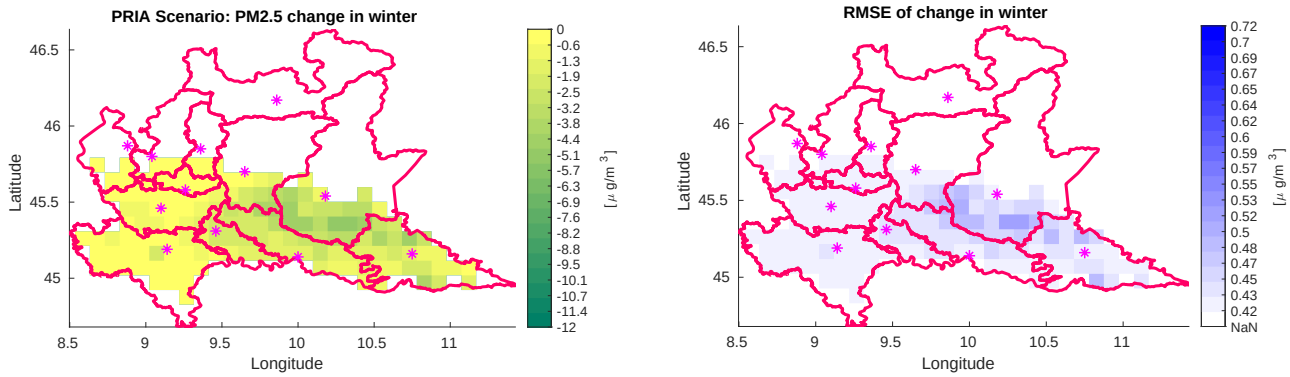


Figure 28: Autocorrelation function (ACF) computed by station residuals for the first 30 lags (days) and summarised through boxplots. The outliers (red cross) are identified by the Whisker length based on the first and third quartiles.

average concentration and the uncertainty of the $PM_{2.5}$ changes due to NH_3 reductions on a regular grid of $0.1° \times 0.1°$ over the studied area. We focus on winter (452 daily predictions for each pixel), which is characterised by high $PM_{2.5}$ concentrations (see Table 12) and the largest effect of NH_3 (see Table 21). For winter, Figure 29 depicts the average $PM_{2.5}$ reduction for the PRIA scenario and the associated uncertainty. Analogously, Figure 30 shows the average impact of the Strong scenario. Both scenarios show that, in winter, the main reduction effect is obtained in the southeast area of the region where the NH_3 emissions are the highest. In particular, in some areas, the average reduction of $PM_{2.5}$ concentrations is close to 6 $\mu g/m^3$. Considering the province's plain average, Brescia and Cremona, which are located in the south-east part of the region, show the highest $PM_{2.5}$ reductions. On the other hand, highly urbanised lands such as the metropolitan area of Milan, which is characterised by low levels of NH_3 emissions, do not benefit from the reduction of ammonia emissions. The more marked reduction of NH_3 emissions (Strong scenario) leads to more significant improvements in air quality, achieving local $PM_{2.5}$ reductions average close to 12 $\mu g/m^3$, reducing the risks associated with the health of the population.

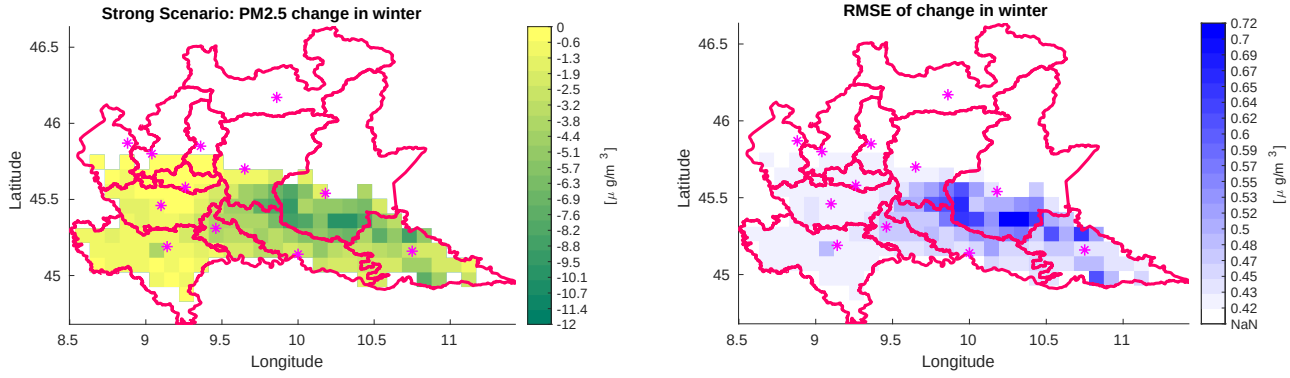
Furthermore, Figures 31 and 32 show the distribution of the winter daily $PM_{2.5}$ reduction aggregated by province and land use for the two scenarios considered. As expected, the largest reductions are obtained in Brescia, Cremona and Mantua provinces, which are characterised by large rural areas

and extensive livestock farming. It is also shown that metropolitan and hill areas are not affected by a significant reduction.



(a) (b)

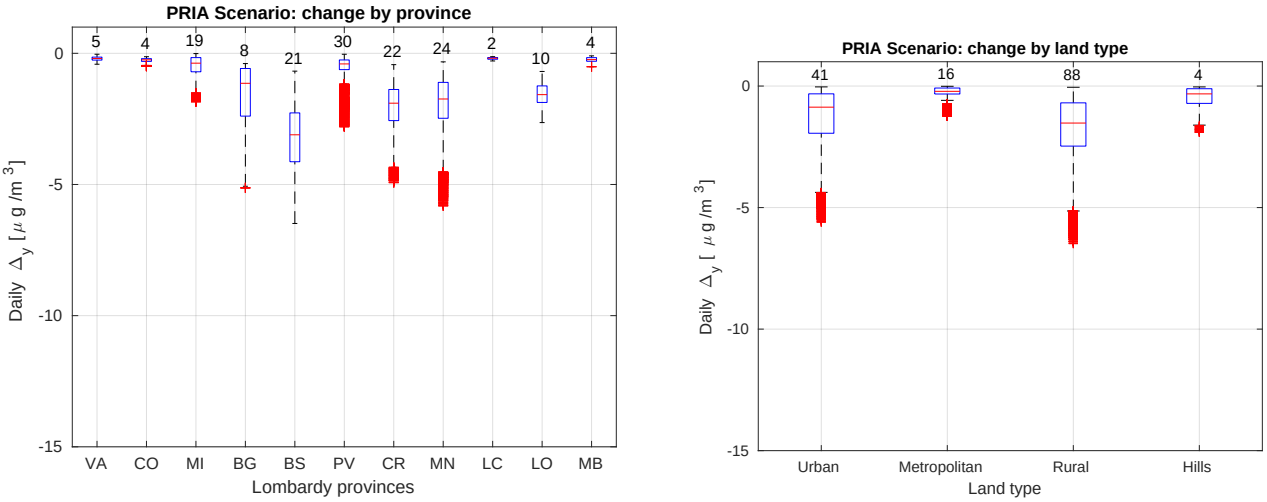
Figure 29: PRIA Scenario (-26%). Panel a: average $PM_{2.5}$ reduction in winter, $\bar{\Delta}_{\hat{y}}$ (452 daily observations for each pixel). Panel b: uncertainty associated with the average reduction (std($\bar{\Delta}_{\hat{y}}$)). The pink stars depict the provincial capitals listed in Table 13. The model is only evaluated in non-forested areas below 640 m altitude.



(a) (b)

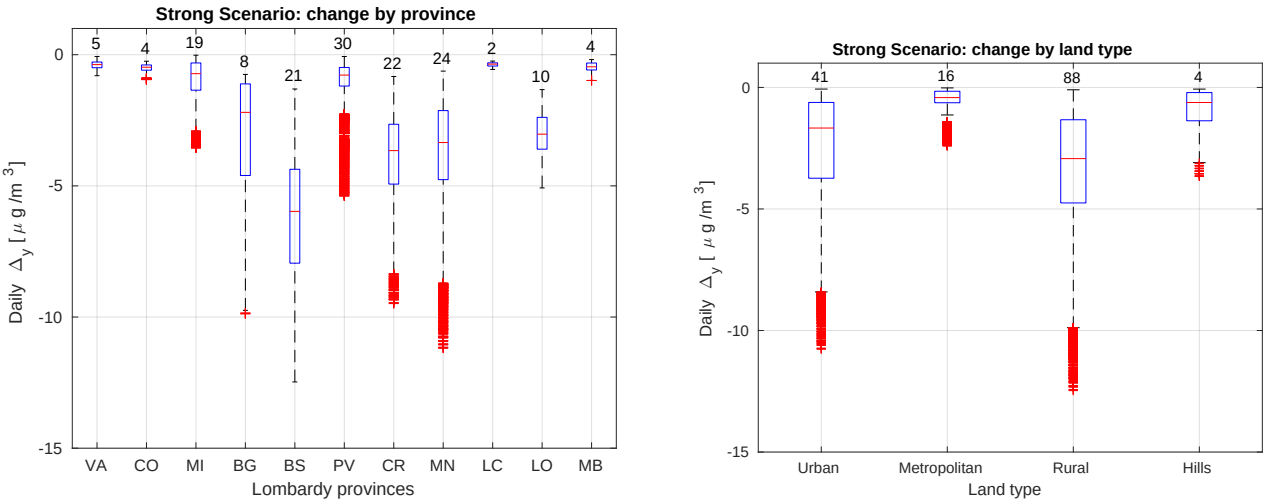
Figure 30: Strong Scenario (-50%). Panel a: winter $\bar{\Delta}_{\hat{y}}$ (452 daily observations for each pixel). Panel b: uncertainty associated with the average reduction (std($\bar{\Delta}_{\hat{y}}$)). The pink stars depict the provincial capitals listed in Table 13. The model is only evaluated in non-forested areas below 640 m altitude.

Table 19 summarises the average of $PM_{2.5}$ reductions aggregated by province in $\mu g/m^3$. Note that, because of the presence of the Prealps and Alps, the scenario analysis covers a limited surface of the Lecco, Varese and Como provinces. Finally, Table 20 shows the change aggregated by season. As expected, the largest effect is in winter, when NH_3 plays a limiting role in the formation of $PM_{2.5}$. We observe that, in spring, we have no effect because the increase is non-significant, being smaller than 0.1% and with a t-statistic not larger than 0.125. In winter, when the NH_3 emissions are less abundant, the overall estimated reduction over the plain areas for the PRIA scenario is 1.44 (std = 0.08), while for the Strong scenario is 2.76 (std = 0.16). These results can be compared with the $PM_{2.5}$ estimated



(a) (b)

Figure 31: Scenario PRIA (-26%). Boxplot of daily $\Delta_{\hat{y}}$ in winter (452 daily observations for each pixel). Panel a: grouped by province; panel b: grouped by land type. The number of pixels of each category is reported over the corresponding boxplot. The model is evaluated only in the non-forested area under 640 m altitude.



(a) (b)

Figure 32: Scenario Strong (-50%). Boxplot of daily $\Delta_{\hat{y}}$ in winter (452 daily observations for each pixel). Panel a: grouped by province; panel b: grouped by land type. The number of pixels of each category is reported over the corresponding boxplot. The model is evaluated only in the non-forested area under 640 m altitude.

average 39.49 $\mu g/m^3$. Hence, considering the PRIA scenario, the $PM_{2.5}$ overall reduction is close to 3.5%, while considering the Strong scenario, the $PM_{2.5}$ overall reduction is close to 7%.

5.4.4 Discussion of results

We assess the $PM_{2.5}$ sensitivity to the NH_3 emissions through a geostatistical approach by using scenario analysis. We consider the heteroskedastic HDMG (5.7); under the PRIA scenario, the overall $PM_{2.5}$ reduction is approximately 3.5%, while considering the Strong scenario, the overall $PM_{2.5}$ reduction is approximately 7%. The results are generally consistent with those obtained using

Province	$\mathcal{D}^*$	$\bar{y}$ $\mu g/m^3$	PRIA $\mu g/m^3$	Strong $\mu g/m^3$
VA	5	32.78	-0.19 (0.09)	-0.37 (0.09)
CO	4	36.26	-0.26 (0.11)	-0.50 (0.11)
MI	19	37.56	-0.47 (0.03)	-0.90 (0.06)
BG	8	40.60	-1.62 (0.11)	-3.11 (0.19)
BS	21	47.13	-3.14 (0.18)	-6.04 (0.34)
PV	30	36.10	-0.50 (0.03)	-0.96 (0.06)
CR	22	41.33	-2.04 (0.12)	-3.92 (0.22)
MN	24	39.64	-1.97 (0.11)	-3.80 (0.22)
LC	2	34.50	-0.19 (0.21)	-0.37 (0.21)
LO	10	38.41	-1.57 (0.10)	-3.02 (0.18)
MB	4	37.57	-0.24 (0.11)	-0.46 (0.11)
Overall	149	39.49	-1.44 (0.08)	-2.76 (0.16)

Table 19: Winter change by province and scenario. $\mathcal{D}^*$ is the number of pixels involved in the computation; $\bar{y}$ is the estimated average of the $PM_{2.5}$ concentrations; the columns headed PRIA and Strong show the average reduction and its uncertainty, in brackets, computed using Eq. (5.25) for both scenarios.

Season	$\bar{y}$ $\mu g/m^3$	PRIA $\mu g/m^3$	PRIA %	Strong $\mu g/m^3$	Strong %
Autumn	24.49	-0.66 (0.07)	-2.68	-1.26 (0.14)	-5.15
Spring	18.57	0.01 (0.08)	0.03	0.01 (0.15)	0.07
Summer	13.92	-0.25 (0.08)	-1.81	-0.49 (0.15)	-3.48
Winter	39.49	-1.44 (0.08)	-3.64	-2.76 (0.16)	-7.00
Overall	24.05	-0.58 (0.04)	-2.42	-1.12 (0.08)	-4.64

Table 20: $PM_{2.5}$ average change, $\bar{\Delta}_{\hat{y}}(s,t)$, by season and scenario. The number of pixels involved in the computation is $\mathcal{D}^* = 149$; $\bar{y}$ is the estimated average of the $PM_{2.5}$ concentrations; the columns headed PRIA and Strong show the average reduction and its uncertainty, in brackets, computed using Eq. (5.25) and the (%) of reductions for both scenarios.

chemical transport models. Considering the RMSE, our results are, in some cases, better than Veratti et al. (2023). This means that a detailed statistical model fitted to an extensive dataset may catch the main features of a chemical transport model. Because the computational burden is definitely lower, these results hint at the use of statistical emulators for policy impact assessment. Our results are also consistent with those in Fassò (2023), which are based on observed ammonia concentrations instead of inventory emissions. This means that the emission data are reliable for understanding the impact of livestock on air quality. The software for implementing the heteroskedastic HDGM developed in this study represents an updated version of the open-source D-STEM software Yaqiong Wang, Finazzi, and Fassò (2021). It can be accessed on the GitHub repository at `https://github.com/graspa-group/d-stem`.

The monitoring network spatial unbalance mentioned in the introduction and mapped in Figure

17 means that our model has not been trained where NH_3 emissions are very high. For this reason, we consider our results with caution and think that the impact computed may underestimate the true impact. Further research is needed to understand this. Because the Lombardy sampling bias is a consequence of the European Union (EU) rules not requiring NH_3, it is not easy to fill this gap for the EU. One possibility is to validate our approach using Swiss data S. K. Grange, J. Sintermann, and C. Hueglin (2023), which has good temporal and spatial coverage.

5.5 Multivariate Analysis

To demonstrate the model fitting using the functional heteroskedastic HDGM, we consider the multivariate analysis with $PM_{2.5}$ (number of available sensors, 45) and PM_{10} (number of available sensors, 101). Indeed the empirical cross-correlation between the $PM_{2.5}$ and PM_{10} in Lombardy is about 0.95. These pollutants are typically observed at partially different monitoring stations (heterotopic) and are known to be highly correlated. Hence, leveraging this correlation, the aim is to obtain maps of daily concentrations at the national level.

For both variables, we consider the regression term defined by Eq. (5.26). The heteroskedastic HDGM handles the cross-correlation among variable using the linear coregionalisation model (LCM) for the multivariate covariance of the innovation process $\boldsymbol{\eta}_t$, that is $\boldsymbol{\Sigma}_\eta = \boldsymbol{V}\rho(\|\boldsymbol{s} - \boldsymbol{s}'\|; \boldsymbol{\theta})$ where $\boldsymbol{V}$ is the $q \times q$ matrix of the coregionalisation and q is the state process dimension, while the function $\rho(\cdot, \boldsymbol{\theta}) : \mathbb{R} \to \mathbb{R}$ is a valid univariate correlation function.

The vector of the basis spline $\boldsymbol{\Psi}_i(t)$ is equal for each i-process and is defined using the cubic B-spline basis function. The domain $[1, T]$ is partitioned into 25 knots. Table 21 summarises the estimated coefficients $\boldsymbol{\beta}$ of the linear regression model used for the large-scale component. All coefficients except the intercept are statistically significant. As expected, the role of NH_3 is most prominent in the winter, when NH_3 plays a limiting role in the formation of PM. The estimated parameters of the latent variable are $\hat{\theta} = 0.6°$, $\hat{f}_1 = 0.77$, $\hat{f}_2 = 0.78$ and the model cross-correlation between $PM_{2.5}$ and PM_{10} is 0.91. Finally, Figures 33 and 34 show the daily $\hat{\sigma}_\epsilon(t)^2$, which is higher in winter than in summer, highlighting the need for a heteroskedastic model.

Name	β ($PM_{2.5}$)	β (PM_{10})
(Intercept)	24.049	34.946
Wind speed	-1.434	-1.91
Temperature	-0.33	-0.173
RH	0.131	0.127
Rain	-0.409	-0.54
BLH	-0.003	-0.004
Urban	-0.402	-1.63
Forest	**NAN**	-13.775
HVI	-2.247	-3.374
SO_2	-0.124	-0.028
NO_x	0.103	0.115
NO_x:Winter	0.063	0.081
NO_x:Summer	-0.068	-0.14
NO_x:Spring	-0.029	-0.069
NH_3	0.232	0.308
NH_3:Winter	0.829	1.041
NH_3:Summer	-0.186	-0.254
NH_3:Spring	-0.225	-0.291

Table 21: Estimate of the fixed effect coefficients β of the bi-variate model.

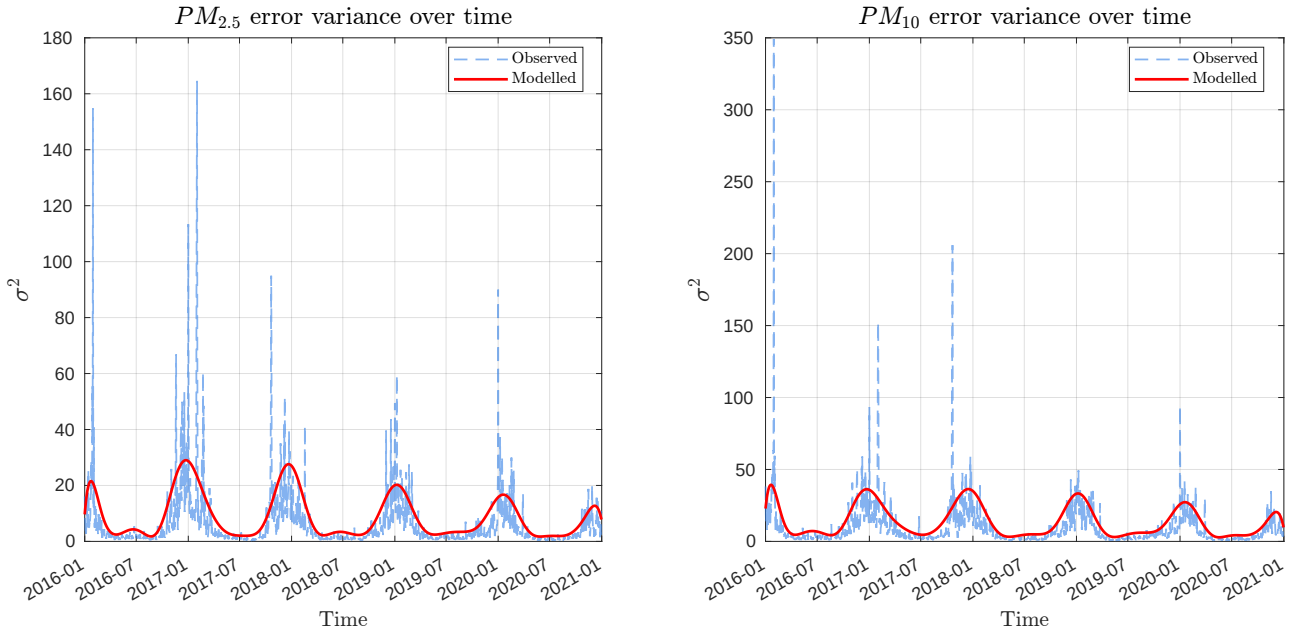


Figure 33: Panel (a): $\hat{\sigma}_\epsilon(t)^2$ for the $PM_{2.5}$ (blue curve) compared with the empirical daily variance of the residual (orange line). Panel (b): $\hat{\sigma}_\epsilon(t)^2$ for the PM_{10} (blue curve) compared with the empirical daily variance of the residual (orange line).

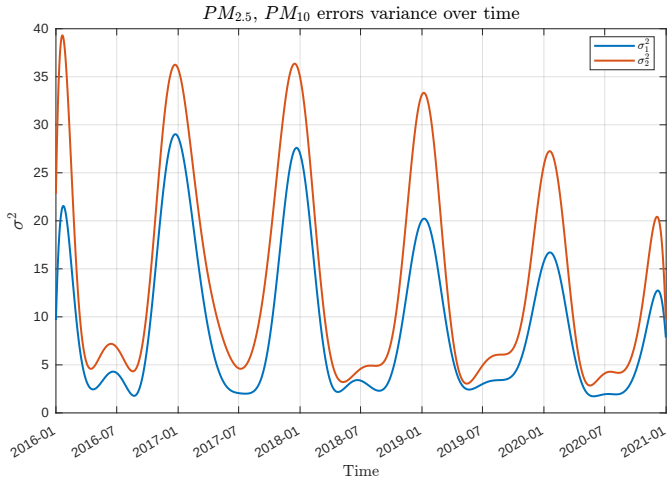


Figure 34: $\hat{\sigma}_\epsilon(t)^2$ for $PM_{2.5}$ (blue curve) and PM_{10} (orange curve).

Chapter 6

Multivariate low-rank State-Space Model with SPDE approach for high-dimensional data

With the widespread availability of remote sensing platforms, especially satellite-based systems, the volume and resolution of spatial data have increased dramatically, see for instance Datta et al. (2016) and Heaton et al. (2019), thereby significantly amplifying the computational demands associated with Gaussian process (GP) modelling. This challenge is known as the "Big n problem"(Jona Lasinio, Mastrantonio, and Pollice, 2013): the computational cost of inverting a covariance matrix scales poorly with the number of spatio-temporal observations $n = mT$, where m is the total number of locations and T is the number of time steps, since covariance matrices grow with complexity $O(m^2T^2)$ in storage and $O(m^3T^3)$ in inversion. These demands quickly exceed available resources under increasing data volume, rendering exact inference infeasible even with a moderately large n. Practical implications of inference on large spatio-temporal datasets are already introduced in Section 1.1.2.

In the context of GPs, a range of strategies has been developed to address the high computational cost associated with large datasets. Comprehensive reviews of these methods can be found in Heaton et al. (2019) and H. Liu et al. (2020). Three major classes of approximation techniques are commonly used: likelihood approximations, low-rank approximations, and Gaussian Markov Random Field (GMRF). A key drawback of using computationally efficient approximated models is the need to assess how closely the approximate predictors and their mean squared prediction errors match the true, theoretical values.

Given these considerations, through this chapter, we will introduce a novel multivariate Low-Rank State-Space Model (LR-SSM) designed for heterotopic observations. Our approach incorporates

the GMRF framework to significantly reduce the rank of the spatial covariance matrix, from m to $R \ll m$. This reduces the computational cost of the Kalman filter to $O(mR^2T)$, while preserving the physical interpretation of the original model. We prove theoretical results, including asymptotic properties and error bounds for the proposed low-rank model. The maximum likelihood estimation (MLE) of the model parameters is carried out via EM, which results in closed-form updates for most parameters. Extensive simulations validate both the accuracy and efficiency of the method. Although our discussion focuses on domains $\Omega \subset \mathbb{R}^2$, the approach extends naturally to higher-dimensional manifolds, and future studies may further adapt it to networks and graphs, highlighting the model's flexibility and broad applicability.

The remainder of this chapter is organised as follows. Section 6.1 details the big-n problem for the SSM model and will highlight the computational bottleneck. Section 6.2 details a range of strategies that have been developed to address the high computational cost associated with large datasets. Section 6.3 recalls briefly the formulation of the Gaussian regression problem using the SSM framework and presents the equations defining the proposed LR-SSM, tailored for large-scale multivariate spatio-temporal datasets. We also provide theoretical results and examine how the low-rank approximation affects inference. In Section 6.5, we detail the ML estimation procedure, with a focus on the low-rank representation derived via the GMRF approach. Section 6.6 presents numerical studies that illustrate the performance of our method and confirm the theoretical results. In Section 6.7, we present the application of the LR-SSM to bivariate air quality data. Most of the material and methodological framework discussed in this chapter is drawn from (Rodeschini, Tedesco, et al., 2025), which provides the main reference for the analyses and approaches presented here.

6.1 Big-n problem and computational details

An efficient approach to making inferences and predictions on spatio-temporal processes is to cast the space-time regression defined by the SRE model (2.15) into the SSM framework. Since it can handle both spatio-temporal random effects and fixed effects factors, the SSM in the spatio-temporal setting has been shown to have good performance and be competitive in explaining the daily $PM_{2.5}$ concentration. See, for instance, Table 15 in Chapter 5, where prediction performance indices among three models are compared in cross-validation. Furthermore, a heteroskedastic SSM is proposed in Section 5.3 to assess the $PM_{2.5}$ sensitivity to the NH_3.

To illustrate the computational advantage of using the Kalman filter (KF) on the SSM model (2.23)-(2.24) to address spatio-temporal regression, we compare the computational complexity of the

univariate SSM to that of an evaluation of $\mathrm{E}(y_t(\boldsymbol{s}_0) \mid y_1(\boldsymbol{s}), \ldots, y_T(\boldsymbol{s}))$ based on a generic joint Gaussian distribution of $(y_t(\boldsymbol{s}_0), y_1(\boldsymbol{s}), \ldots, y_T(\boldsymbol{s}))$. The computational complexity is calculated in terms of the number of observed time units T and the number of observed spatial locations m.

Given the observed data $\boldsymbol{y}_{1:T} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ where $\boldsymbol{y}_t = (y(\boldsymbol{s}_1, t), \ldots, y(\boldsymbol{s}_m, t))'$ is a realisation of the process $y(\boldsymbol{s}, t)$ at time t and location set $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m\}$, the calculation of $\mathrm{E}(y_t(\boldsymbol{s}_0) \mid \boldsymbol{y}_{1:T})$, under the Gaussianity assumption, is given by the conditional expected values of Gaussian distribution which requires inverting the covariance matrix $\mathrm{Cov}(\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_T^\top)$ of size $mT \times mT$, see the Example 3. The computational complexity is $O(m^3T^3)$, which results in a computation that will fail if m or T is too large. One way to reduce the computational complexity is to exploit the state-space formulation and implement vector-based Kalman filtering in the time dimension. Under the dynamical approach and the Markovian assumption, inference on $y(\boldsymbol{s}, t)$ simplifies to an iterative procedure for $t = \{1, \ldots, T\}$. This results in a complexity reduction from $O(m^3T^3)$ to $O(m^3T)$, but this computation will also fail if m is too large.

To address this issue, several approximation strategies have been proposed. For instance, covariance tapering has been applied to SSMs to induce sparsity in the spatial structure (Bevilacqua, Mateu, et al., 2010). Alternatively, low-rank approximations, such as those based on fixed rank kriging (FRK), have been applied within the SSM framework (Cressie, Shi, and Kang, 2010). More recently, Schmidt et al. (2023) proposed a method that reduces rank via a truncated singular value decomposition of the conditional covariance matrix, though this may come at the cost of losing physical interpretability of the latent process.

6.2 Efficient representations of continuous Gaussian Process

As already mentioned above, exact inference using covariance-based models is generally computationally expensive. The desire to utilise the SSM within computational constraints has led to a large number of approximation techniques, and a few of these are introduced in this section. While inference on the SSM using the Kalman filter scales linearly with time, it incurs a cubic cost in the number of spatial locations, which becomes prohibitive for large m. For this reason, we focus on spatial approximation techniques.

6.2.1 Likelihood approximations

Since any joint density can be written as a product of conditional densities based on some ordering of the observations, one way to lessen the computations is to condition on only some observations when

computing the conditional densities.

Vecchia (1988) suggested a simple approximation to the likelihood for spatial data. It reformulates the joint distribution of observations into a product of conditional distributions, where each observation conditions only on a subset of others. Using $\pi(\boldsymbol{y};\Pi)$ to indicate a generic density (possibly conditional), suppose that $\boldsymbol{y} = (y(\boldsymbol{s}_1),\dots,y(\boldsymbol{s}_m))$ are observation of the spatial process $y(\boldsymbol{s})$ has a joint density $\pi(\boldsymbol{y};\Pi)$ where Π is a set of unknown parameters. Partition $\boldsymbol{y}$ into subvectors $\boldsymbol{y}_i \subset \boldsymbol{y}$ of possibly different lengths for $i = 1,\dots,K$ and define $\boldsymbol{y}_{(i)} = (\boldsymbol{y}_1,\dots,\boldsymbol{y}_i)$. We always have

$$\pi(\boldsymbol{y};\Pi) = \pi(\boldsymbol{y}_1;\Pi)\prod_{j=2}^{K}\pi(\boldsymbol{y}_j|\boldsymbol{y}_{(j-1)};\Pi). \tag{6.1}$$

Vecchia (1988) noted that it may not be critical to condition on all components of $\boldsymbol{y}_{(j-1)}$ when calculating $\pi(\boldsymbol{y}_j|\boldsymbol{y}_{(j-1)};\Pi)$ and thereby reduce the computational effort. In particular, if, for $j = 1,\dots,K-1$, $\boldsymbol{s}_{(j)}$ is some subvector of $\boldsymbol{y}_{(j)}$, then

$$\pi(\boldsymbol{y};\Pi) \approx \pi(\boldsymbol{y}_1;\Pi)\prod_{j=2}^{K}\pi(\boldsymbol{y}_j|\boldsymbol{s}_{(j-1)};\Pi). \tag{6.2}$$

The approximation (6.2) is the general form for Vecchia's approximation to the likelihood. To implement this approximation, we need to order the observations in some manner and choose the conditioning vectors. Vecchia (1988) recommended choosing $\boldsymbol{s}_{(j-1)}$ to be made up of the m observations that are nearest to $\boldsymbol{y}_j$. This choice has the considerable virtue of simplicity. One might instead consider constructing conditioning sets that minimise the conditional prediction variance, thereby providing optimal predictors in the kriging sense. However, even if we could pick the conditioning vectors to minimise the prediction variance, such designs are not necessarily good choices for purposes of estimation (Stein, Chi, and Welty, 2004). Jones and Yiming Zhang (1997) discussed applications of Vecchia's approach to spatio-temporal processes. However, it does point out the difficulty in choosing conditioning sets when different co-ordinates are not commensurable.

Covariance tapering for univariate (Furrer, Genton, and D. Nychka, 2006; Kaufman, Schervish, and D. W. Nychka, 2008) and multivariate (Bevilacqua, Fassò, et al., 2016) process is another likelihood-based method that offers a computationally efficient likelihood approximation for spatial statistical analysis. Considering the SRE model (2.16) and the MLE shown in Example 2, the core idea of tapering is to keep the covariances approximately unchanged at small distance lags and to reduce the covariances to zero at large distances. To implement the idea, let $c_{tap}(h)$ be an isotropic covariance function with compact support; that is, $c_{tap}(h) = 0$ if $h \geq \gamma$ for some $\gamma > 0$. Then, the tapered covariance function $\tilde{c}$ is

$$\tilde{c}(h;\vartheta) = c(h;\vartheta)c_{tap}(h), \tag{6.3}$$

and the tapered covariance matrix $\tilde{\mathbf{\Sigma}} = \mathbf{\Sigma}(h;\vartheta) \odot \mathbf{K}_{tap}(h)$, where $\odot$ denotes the Hadamard (element-wise) product and $\mathbf{K}_{tap}$ is a sparse, positive definite tapering matrix. A popular choices for taper functions are the Wendland functions. This sparsity significantly reduces the computational burden to $O(N \log N)$ or $O(N)$ for certain structures, as the inverse of $\tilde{\mathbf{\Sigma}}$ becomes sparse or can be efficiently approximated. One can use the tapered covariance function $\tilde{c}(h;\boldsymbol{\vartheta})$ for spatial interpolation and estimation as if it were the correct covariance function. Under mild conditions, tapering has been shown to yield consistent and asymptotically normal estimators for the covariance parameters.

While it introduces a bias into the likelihood, this bias is often negligible for sufficiently large sample sizes. Furrer, Genton, and D. Nychka (2006) showed that appropriate tapering does not affect the fixed-domain asymptotic mean square error of prediction for the Matérn model. Du, H. Zhang, and Mandrekar (2009) showed that, under some conditions on the taper, the tapered MLE is asymptotically as efficient as the true MLE for the $\sigma^2\theta^{2\nu}$ parameter in the Matérn model.

The sparsity of $\tilde{\mathbf{\Sigma}}$ increases as γ is decreased, but the taper function and the taper range should, of course, also be chosen such that the shape of the true covariance function is preserved, and of especial importance for asymptotic considerations is that the smoothness at the origin is preserved. However, tapering assumes that spatial dependence vanishes beyond a fixed distance, which may underestimate long-range correlations.

6.2.2 Low-rank methods

In many of the techniques for building computationally efficient models, the main assumption is that a latent zero-mean GP $z(\boldsymbol{s})$ can be expressed, or at least approximated, through some finite basis expansion:

$$z(\boldsymbol{s}) = \sum_{j=1}^{R} \phi_j(\boldsymbol{s}) w_j = \mathbf{\Phi}\boldsymbol{w}, \tag{6.4}$$

where w_j are Gaussian random variables and $\{\phi_j\}_{j=1}^{R}$ are some pre-defined basis functions. To show the computational efficiency associated with the approximation (6.4), the Example 6 considers the MLE of the SRE model (2.16). Specifically, the computational cost for evaluating the log-likelihood $\text{logL}(\mathbf{y} \mid \boldsymbol{X}; \Pi)$ reduces from $O(m^3)$ to $O(R^3)$, and by choosing $R \ll m$, the computational complexity is substantially decreased. However, the choice of basis functions involves multiple tuning parameters, such as the types of basis functions, levels of resolution, and information in the coarsest resolution, which can lead to issues with identifiability and overfitting (Song, Dai, and Genton, 2024).

A popular methods that fit into the low-rank approximations includes the Fixed-Rank Kriging (Cressie and Johannesson, 2008; Zammit-Mangion and Cressie, 2021) and LatticeKrig (D. Nychka et al.,

2015). Their recommendation is to use multi-resolution basis functions. Many other methods can be viewed as low-rank approximations, e.g. the predictive process method by Banerjee et al. (2008).

Example 6. *(Maximum-likelihood SRE estimation with low-rank approximation) We recall here the observed log-likelihood* $\text{logL}(\mathbf{y} \mid \boldsymbol{X}; \Pi)$ *associated with the SRE model* (2.16) *given by*

$$2\,\text{logL}(\mathbf{y} \mid \boldsymbol{X}; \Pi) = -n\log(2\pi) - \log|\boldsymbol{\Sigma}(\vartheta) + \sigma^2_\epsilon I_n| - (\mathbf{y} - \boldsymbol{X}'\boldsymbol{\beta})'(\boldsymbol{\Sigma}(\vartheta) + \sigma^2_\epsilon I_n)^{-1}(\mathbf{y} - \boldsymbol{X}'\boldsymbol{\beta}), \quad (6.5)$$

the parameter set $\Pi = \{\boldsymbol{\beta}, \vartheta, \sigma^2_\epsilon\}$ *can be estimated by maximising* (6.5)*. Using the low-rank approximation* (6.4)*, then* $\boldsymbol{z}(\boldsymbol{s})$ *can be written as* $\boldsymbol{z} = \boldsymbol{\Phi}\boldsymbol{w} \sim \text{N}(0, \boldsymbol{\Phi}\boldsymbol{Q}\boldsymbol{\Phi}')$ *where i-column in the matrix* $\boldsymbol{\Phi}$ *contains the basis function* $\phi_i(\boldsymbol{s})$ *evaluated at all measurement locations. Sometimes* $\boldsymbol{Q}$ *is parametrised as well, so that* $\boldsymbol{Q}(\tilde{\vartheta})$*. The log-likelihood* $\text{logL}(\mathbf{y} \mid \boldsymbol{X}; \Pi)$ *becomes*

$$2\,\text{logL}(\mathbf{y} \mid \boldsymbol{X}; \Pi) = -n\log(2\pi) - \log|M| - (\mathbf{y} - \boldsymbol{X}'\boldsymbol{\beta})'\boldsymbol{M}^{-1}(\mathbf{y} - \boldsymbol{X}'\boldsymbol{\beta}), \quad (6.6)$$

where $\boldsymbol{M} = \boldsymbol{\Phi}\boldsymbol{Q}\boldsymbol{\Phi}' + \sigma^2_\epsilon \boldsymbol{I}_n$*. Using the Sherman–Morrison–Woodbury formula,* $(\boldsymbol{M})^{-1}$ *can be written as* $\boldsymbol{M}^{-1} = (\sigma^2_\epsilon)^{-1}\boldsymbol{I}_n + (\sigma^2_\epsilon)^{-1}\boldsymbol{\Phi}\big(\boldsymbol{Q}^{-1} + (\sigma^2_\epsilon)^{-1}\boldsymbol{\Phi}'\boldsymbol{\Phi}\big)((\sigma^2_\epsilon)^{-1}\boldsymbol{\Phi})'$ *which requires* $O(R^3)$ *floating point operations.*

6.2.3 Markov approximations

A distinct and powerful class of approximations of spatial GPs is based on Gaussian Markov Random Fields (GMRFs) (Havard Rue and Held, 2005) discussed in Chapter 3, where conditional independence across a neighbour structure yields a sparse precision matrix. Whittle (1954) and Whittle (1963) showed that a Matérn GP can be characterised as the unique solution to a specific stochastic partial differential equation (SPDE) defined over the domain $\mathcal{D}$. By applying the Galerkin method (Ern and Guermond, 2004) to this SPDE, one obtains a low-rank approximation of the field η_i, which corresponds to a Gaussian Markov Random Field (GMRF) representation of the original GF. Lindgren, Håvard Rue, and Lindström (2011) gives the explicit connection between the GMRF and the Matérn GP; we refer the reader to Lindgren, Håvard Rue, and Lindström (2011), which provides the theoretical foundation for the proposed approximation. This SPDE approach enables linear-cost computation of the precision matrix (inverse of the covariance matrix) over complex domains and has recently been extended to spatio-temporal processes (Clarotto et al., 2024).

6.3 Multivariate State-Space Model framework

Building on the previous discussion, we briefly recall the multivariate SSM framework and the matrix parametrisation considered; in the next section, we introduce the multivariate Low-Rank State-Space Model (LR-SSM) designed for heterotopic observations.

Let $\boldsymbol{y}(\boldsymbol{s},t) = (y_1(\boldsymbol{s},t),\dots,y_p(\boldsymbol{s},t))' \in \mathbb{R}^p$ be a p-dimensional spatio-temporal random field defined over a spatial domain $\mathcal{D} \subset \mathbb{R}^2$ and discrete time steps $t = 1,\dots,T$, where $\boldsymbol{s} \in \mathcal{D}$ indexes the spatial location. The term $\boldsymbol{\eta}(\boldsymbol{s},t) = (\eta_1(\boldsymbol{s},t),\dots,\eta_q(\boldsymbol{s},t))'$ represents a q-dimensional, zero-mean Gaussian innovation process which is temporally independent. Moreover, we assume that the components of $\boldsymbol{\eta}(\boldsymbol{s},t)$ are mutually independent, i.e., $\eta_j(\boldsymbol{s},t) \perp \eta_{j'}(\boldsymbol{s}',t')$ for all $j \neq j'$, at any spatial locations $\boldsymbol{s},\boldsymbol{s}'$ and times t,t'. We consider the following matrix parametrisation for the continuous-space SSM (2.23)–(2.24): $\boldsymbol{F} = \text{diag}(\boldsymbol{f})$ denotes a $q \times q$ diagonal transition matrix describing the system's dynamics, and $\boldsymbol{f} = (f_1,\dots,f_q)'$ with each $f_i \in (-1,1)$ to ensure the stationarity of $\boldsymbol{z}(\boldsymbol{s},t)$. Moreover, for every $\boldsymbol{s} \in \mathcal{D}$, $\boldsymbol{z}(\boldsymbol{s},0)$ is an independent Gaussian variable with mean $\boldsymbol{\mu}_0 \in \mathbb{R}^q$ and $q \times q$ covariance matrix $\boldsymbol{\Sigma}_0$.

We also assume that each $\eta_i(\boldsymbol{s},t)$ is a zero-mean Matérn GP, with rescale parameter κ_i, unit marginal variance (i.e $\sigma_i^2 = 1$ for $i = 1\dots,q$) and smoothness parameter ν_i. κ_i is the rescale parameter which controls the spatial range of dependence by $\theta_i = \sqrt{(8\nu)}/k_i$. The smoothness parameter ν is rarely estimated directly; instead, it is typically set a priori based on the desired level of smoothness in the Matérn process or informed by physical assumptions about the underlying phenomenon. In what follows, we set $\nu_i = 1$, which corresponds to a spatial process that is continuous but not mean-square differentiable. This choice strikes a balance between model flexibility and computational efficiency, making it well-suited for applications where a certain degree of spatial roughness is expected or acceptable. While the proposed methodology allows for alternative values of ν_i, we adopt this setting to streamline both the notation and the computational workload.

We denote by $\boldsymbol{\Sigma}_\epsilon = \text{diag}(\boldsymbol{\sigma}_\epsilon^2)$ the diagonal variance matrix of $\boldsymbol{\varepsilon}(\boldsymbol{s},t)$, with $\boldsymbol{\sigma}_\epsilon^2 = (\sigma_{\epsilon,1}^2,\dots,\sigma_{\epsilon,p}^2)' \in \mathbb{R}_+^p$ being the vector of measurement error variances. Under these matrix parameterisations, the model parameter set that needs to be estimated is:

$$\Pi = \{\boldsymbol{\beta}, \boldsymbol{\sigma}_\epsilon^2, \boldsymbol{f}, \boldsymbol{W}, k_1,\dots,k_q\}. \tag{6.7}$$

As we discussed in the introduction, however, the Kalman filter requires the inversion of an $m \times m$ matrix at every time $t = 1,\dots,T$, which results in a cost of $\mathcal{O}(m^3T)$. While linear in time, it does not scale properly in terms of the number of spatial locations m. For this reason, we present in the next section a low-rank SSM based on the SPDE approach (Lindgren, Håvard Rue, and Lindström, 2011)

that aims to approximate the data-generating process (2.23)–(2.24) and that can be used to obtain inference on the latent dynamics at the lower cost of $O(mR^2T)$ with $R \ll m$. In the sequel, we also discuss the case $R = m$.

6.4 Multivariate low-rank SPDE State-Space Model

In this section, we introduce a low-rank SSM based on the GMRF approach (Lindgren, Håvard Rue, and Lindström, 2011), designed to approximate the data-generating process (2.23)–(2.24) and to enable inference on the latent dynamics at a reduced computational cost of $O(mR^2T)$ with $R \ll m$. We also examine the special case $R = m$. To this end, we discuss how to replace the GP η_i in Equation (2.24) with a low-rank approximation by expressing η_i as a linear combination of R finite basis functions, yielding a finite-dimensional representation of the state equation and, ultimately, computational gains.

Lindgren, Håvard Rue, and Lindström (2011) established a link between Matérn GPs and GMRFs by solving a Stochastic Partial Differential Equation (SPDE) using the Finite Element Method (FEM). To simplify the presentation, we present in this section the resulting low-rank representation, and defer the technical details to Appendix B.3.

Let us consider a partition of the domain $\mathcal{D} \subset \mathbb{R}^2$ into a set of non-overlapping triangles, i.e., $\mathcal{D} = \bigcup_i E_i$, where any two triangles share at most a common edge or vertex. The result of the discretisation is a mesh $\mathcal{G}_R = (\mathcal{V}_R, \mathcal{E}_R)$, with R equal to the cardinality of the vertices $\mathcal{V}_R = \{\boldsymbol{r}_1, \dots, \boldsymbol{r}_R\}$ and $\mathcal{E}_R$ being the set of the edges of the triangles. Based on the mesh $\mathcal{G}_R$, we define a set of basis functions $\boldsymbol{\psi}_R(\boldsymbol{s}) = (\psi_1(\boldsymbol{s}), \dots, \psi_R(\boldsymbol{s}))'$, where each $\psi_i : \mathbb{R}^2 \to \mathbb{R}$ is selected from the class of piecewise linear functions with compact support. These functions are constructed such that $\psi_i(\boldsymbol{r}_k) = 1$ if $i = k$, and $\psi_i(\boldsymbol{r}_k) = 0$ if $i \neq k$, for all $i, k = 1, \dots, R$. Additionally, we impose the condition that all basis functions have zero normal derivatives on the boundary of $\mathcal{D}$, which ensures that the SPDE approximation of the process $\eta_i(\boldsymbol{s}, t)$ exhibits the desired asymptotic properties, discussed in the sequel. A similar approach is presented in Lindgren, Håvard Rue, and Lindström (2011), to which we refer for further technical details.

Let $\boldsymbol{\eta}_j(\mathcal{V}_R, t)$ denote the vector:

$$\boldsymbol{\eta}_j(\mathcal{V}_R, t) = \big(\eta_j(\boldsymbol{r}_1, t), \dots, \eta_j(\boldsymbol{r}_R, t)\big)',$$

for $j = 1, \dots, q$, where $\mathcal{V}_R = \{\boldsymbol{r}_1, \dots, \boldsymbol{r}_R\}$ is a finite set of spatial locations. Then, we define the full vector $\boldsymbol{\eta}(\mathcal{V}_R, t) \in \mathbb{R}^{qR}$ as:

$$\boldsymbol{\eta}(\mathcal{V}_R, t) = \big(\boldsymbol{\eta}_1(\mathcal{V}_R, t)', \dots, \boldsymbol{\eta}_q(\mathcal{V}_R, t)'\big)'.$$

The extension to the case where each component has a different mesh, i.e. $\mathcal{G}_R^j = (\mathcal{V}_R^j, \mathcal{E}_R^j)$ for $j = 1, \ldots, q$, is straightforward. Here, for simplicity and without loss of generality, we assume that the latent process components share the same finite element mesh $\mathcal{G}_R$. In Section 6.7, we provide an example with disjoint meshes. Similarly, let $\boldsymbol{z}_j(\mathcal{V}_R, t)$ for $j = 1, \ldots, q$ be the vector $\boldsymbol{z}_j(\mathcal{V}_R, t) = (z_j(\boldsymbol{r}_1, t), \ldots, z_j(\boldsymbol{r}_R, t))'$ and $\boldsymbol{z}(\mathcal{V}_R, t)$ be the vector $\boldsymbol{z}(\mathcal{V}_R, t) = (\boldsymbol{z}_1(\mathcal{V}_R, t)', \ldots, \boldsymbol{z}_q(\mathcal{V}_R, t)')'$.

Let $\mathcal{H}^1(\mathcal{D})$ denote the first-order Sobolev space over the spatial domain $\mathcal{D}$. Consider the finite-dimensional subspace $\mathcal{H}_R^1(\mathcal{D}) \subset \mathcal{H}^1(\mathcal{D})$ spanned by basis functions $\{\psi_i\}_{i=1}^R$. We approximate each component of the innovation term in (2.24) using the Galerkin solution of the SPDE defining the Matérn GF (see the Appendix B.3, Equation (B.1)), projected onto $\mathcal{H}_R^1(\mathcal{D})$. This yields the following approximation:

$$\boldsymbol{z}(\boldsymbol{s}, t) \approx \boldsymbol{F}\boldsymbol{z}(\boldsymbol{s}, t-1) + \boldsymbol{\Psi}_R(\boldsymbol{s})\boldsymbol{\eta}(\mathcal{V}_R, t), \tag{6.8}$$

where $\boldsymbol{\Psi}_R(\boldsymbol{s}) = \boldsymbol{I}_p \otimes \boldsymbol{\psi}_R(\boldsymbol{s})'$, $\boldsymbol{\psi}_R(\boldsymbol{s}) = (\psi_1(\boldsymbol{s}), \ldots, \psi_R(\boldsymbol{s}))'$ and $\otimes$ denotes the Kronecker product.

We propose to leverage (6.8) to obtain a low-rank approximation of model (2.23)-(2.24). That is, considering (6.8) holding as equality, we replace (2.23)-(2.24) with the following LR-SSM:

$$\boldsymbol{y}(\boldsymbol{s}, t) = \boldsymbol{X}'(\boldsymbol{s}, t)\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{\Psi}_R(\boldsymbol{s})\boldsymbol{z}(\mathcal{V}_R, t)' + \boldsymbol{\varepsilon}(\boldsymbol{s}, t) \tag{6.9}$$

$$\boldsymbol{z}(\mathcal{V}_R, t) = (\boldsymbol{F} \otimes \boldsymbol{I}_R)\boldsymbol{z}(\mathcal{V}_R, t-1) + \boldsymbol{\eta}(\mathcal{V}_R, t). \tag{6.10}$$

In Equation (6.10), the continuous latent field $\boldsymbol{z}(\boldsymbol{s}, t)$ from (2.23) is replaced by a low-rank approximation of the form $\boldsymbol{\Psi}_R(\boldsymbol{s})\boldsymbol{z}(\mathcal{V}_R, t)'$, where the reduction ratio is defined as $LR = \frac{|\mathcal{V}_R|}{m} \times 100\%$. Comparing (2.23) with (6.9), the approximation introduces a model-based error. In Section 6.4.1, we show that the model-based error can be controlled by the number of vertices R. As expected, a finer finite element mesh provides increased inference accuracy.

We summarise here the advantages of using the model defined in equations (6.9)–(6.10) to make inference on model (2.23)–(2.24). First, the low-rank approximation reduces the computational complexity of the Kalman filter from $O(m^3T)$ to $O(mR^2T)$, gaining efficiency for $R \ll m$. The practitioner can therefore control the balance between computational cost and accuracy by choosing the value of R. Second, it preserves the first- and second-order moments of the original process in the limit, thereby enabling scalable inference and prediction in high-dimensional spatio-temporal settings. Third, once the random coefficients $\boldsymbol{z}(\mathcal{V}_R, t)$ are estimated, evaluating the latent process at new locations has a linear computational cost. This enables efficient construction of spatial maps of the observed processes, making the proposed model competitive even when $R = m$ ($LR = 100\%$). Details on model estimation and spatial mapping are provided in Section 6.5. Most model parameters can be estimated via closed-form expressions. For the rescaling parameters, k_i for $i = 1, \ldots, q$,

although no closed-form solution is available, the assumption of independent innovation components allows for efficient minimisation whose complexity grows linearly with the dimension of the latent component q. This makes the estimation procedure particularly efficient as compared to other SSMs such as those in Calculli et al. (2015).

6.4.1 Identifiability, Convergence, and Approximation Bounds

We begin by analysing the theoretical properties of the proposed approximation. Since the true data-generating process is unknown, we treat the LR-SSM (6.9)–(6.10) as the underlying model. In this way, we first establish point identification in a distributional sense, followed by a weak convergence result for the low-rank approximation. Finally, we quantify the approximation error by deriving a uniform bound on the difference between the full-rank SSM and its low-rank counterpart.

Assumption 1. *For a subset $\mathcal{S} \subset \mathcal{D}$, the limit $\lim_{T\to\infty} \frac{1}{T} \int_{\mathcal{S}} \boldsymbol{X}(\boldsymbol{s}, t)\, d\boldsymbol{s} = L_{\boldsymbol{X}}$ exists and is finite, and the matrix $L_{\boldsymbol{X}} L'_{\boldsymbol{X}}$ has full rank.*

Assumption 1 imposes a standard full-rank condition on the fixed-effects component of the model, ensuring identifiability of the regression coefficient vector $\boldsymbol{\beta}$.

Assumption 2. *(i) $f_1 > f_2 > \cdots > f_q$; (ii) $\boldsymbol{W}$ has full column rank, with each column scaled so that its first non-zero element is positive; (iii) $\boldsymbol{\Psi}_R(\boldsymbol{s}) = \boldsymbol{I}_p \otimes \boldsymbol{\psi}_R(\boldsymbol{s})'$ is known and has full row rank.*

Assumption 2 imposes restrictions on the parameter space, excluding alternative representations of the LR-SSM that may result from permuting the order or changing the sign of latent components. The following identifiability result is established, with the proof provided in Appendix B.2. We consider point identification in a distributional sense, where model parameters are uniquely determined by the joint distribution of the observables (Lewbel, 2019).

Theorem 1. *Let Assumptions 1–2, hold, and consider the data generated according to the LR-SSM* (6.9)–(6.10)*, with parameter set* Π*. Assume that the process* $\{\boldsymbol{y}(\boldsymbol{s}, t) : (\boldsymbol{s}, t) \in \mathcal{S} \times \mathbb{N}\}$ *is observed over a spatially dense set $\mathcal{S} \subset \mathcal{D}$ and over an arbitrarily long time horizon. Then, the parameter set* Π *is identifiable, meaning that no two distinct parameter sets* Π *and* $\Pi^\star$ *can generate the same distribution of the observed process.*

The SSM models described in (2.23)–(2.24) provide a continuous representation of the observed random process, capturing its inherent continuity. We now analyse the estimation errors that arise when the LR-SSM is employed to infer data generated by the true SSM.

First, we discuss the convergence of the low-rank approximation of the latent component. Let x^R be a sequence of $L^2(\mathcal{D})$-bounded Gaussian fields. As defined in Lindgren, Håvard Rue, and Lindström (2011), we say $x^R \xrightarrow{D\{L^2(\mathcal{D})\}} x$ if for all $f, g \in L^2(\mathcal{D})$,

$$\mathbb{E}[\langle f, x^R\rangle] \to \mathbb{E}[\langle f, x\rangle], \quad \text{and} \quad \mathrm{Cov}(\langle f, x^R\rangle, \langle g, x^R\rangle) \to \mathrm{Cov}(\langle f, x\rangle, \langle g, x\rangle).$$

As $\mathcal{H}^1_R(\mathcal{D})$ is dense in $\mathcal{H}^1(\mathcal{D})$ as $R \to \infty$, Theorem 3 of Lindgren, Håvard Rue, and Lindström (2011) ensures that for a fixed $t \in \{1, \dots, T\}$, it holds $\boldsymbol{\Psi}(\boldsymbol{s})\boldsymbol{\eta}(\mathcal{V}_R, t) \xrightarrow{D\{L^2(\mathcal{D}^q)\}} \boldsymbol{\eta}(\boldsymbol{s}, t)$. Thanks to the Markovian structure of $\boldsymbol{z}(\boldsymbol{s}, t)$, over a finite time horizon T, this result implies that the approximate latent field converges to the true latent field as well, as stated in the following result.

Theorem 2 (Weak Convergence of the Approximated State Process)**.** *For every $t = 1, \dots, T$ and $i = 1, \dots, q$, let $\eta_i(\boldsymbol{s}, t)$ satisfy the Matérn SPDE* (B.1) *with homogenous Neumann condition, and define $z_i(\boldsymbol{s}, t) = \sum_{r=0}^{t} f_i^{t-r}\eta_i(\boldsymbol{s}, r)$. Let $\boldsymbol{\eta}_i(\mathcal{V}_R, t)$ be the vector of coefficients of the Galerkin solution of the SPDE in $\mathcal{H}^1_R(\mathcal{D}, \kappa)$. Then*

$$\boldsymbol{\Psi}_R(\boldsymbol{s})\boldsymbol{z}(\mathcal{V}_R, t) \xrightarrow{D\{L^2(\mathcal{D}^{qT})\}} \boldsymbol{z}(\boldsymbol{s}, t),$$

where $\boldsymbol{z}(\mathcal{V}_R, t)$ are defined in terms of $\{\boldsymbol{\eta}_i(\boldsymbol{s}, t)\}_{i=1}^q$ in (6.10)*. In particular,*

$$\mathbb{E}[\boldsymbol{\Psi}_R(\boldsymbol{s})\boldsymbol{z}(\mathcal{V}_R, t)] \to \mathbb{E}[\boldsymbol{z}(\boldsymbol{s}, t)],$$
$$\mathrm{Cov}(\boldsymbol{\Psi}_R(\boldsymbol{s})\boldsymbol{z}(\mathcal{V}_R, t), \boldsymbol{\Psi}_R(\tilde{\boldsymbol{s}})\boldsymbol{z}(\mathcal{V}_R, \tilde{t})) \to \mathrm{Cov}(\boldsymbol{z}(\boldsymbol{s}, t), \boldsymbol{z}(\tilde{\boldsymbol{s}}, \tilde{t})).$$

The proof is provided in Appendix B.3.1, and relies on standard assumptions about the finite element mesh topology, namely that the minimal mesh angles are bounded away from zero and that the maximum edge length tends to zero. Importantly, the proposed model preserves the first- and second-order moments of the original process in the limit, thereby enabling scalable inference and prediction in high-dimensional spatio-temporal settings.

As a second result, we provide a bound on the observed process when the original model (2.23)–(2.24) is approximated by the model (6.9)–(6.10).

Theorem 3 (Bound on the Observable Process)**.** *For every $t = 1, \dots, T$ and $\boldsymbol{s} \in \mathcal{D}$, let $\boldsymbol{y}(\boldsymbol{s}, t)$ be a random process satisfying the model* (2.23)-(2.24) *with set of parameter Π. Let $\boldsymbol{y}^R(\boldsymbol{s}, t)$ the discretised version of $\boldsymbol{y}(\boldsymbol{s}, t)$ obtained by projecting each component of $\boldsymbol{\eta}_t$ into $\mathcal{H}_R(\mathcal{D})$, so that $\{\boldsymbol{y}_t^R\}_{i=1}^T$ satisfies model* (6.9)-(6.10) *with the same parameter set Π. The following bound holds:*

$$\mathbb{E}\|\boldsymbol{y}_t - \boldsymbol{y}_t^R\|_2^2 \leq C\,\frac{1-\varphi^{2(t+1)}}{1-\varphi^2}\,h^2,$$

where C is a constant that depends on $\|\boldsymbol{W}\|$ and q, $\varphi = \max_i |f_i| < 1$, h denotes the maximum edge length of the finite element mesh used in the approximation, and $\|\cdot\|_2$ stands for the $L^2(\mathcal{D})-$norm.

We prove this result in Appendix B.3.2. The result in Theorem 4 shows that the error when LR-SSM and SSM are parametrised by the same parameter set Π, is uniformly bounded in time since $\frac{1-\varphi^{2(t+1)}}{1-\varphi^2} \leq \frac{1}{1-\varphi^2}$ as $t \to \infty$. Thus, the discrepancy between the true and reduced–order outputs can never grow without bound; at worst, it is proportional to the finite element mesh width h. The linear dependence on h reflects the usual accuracy–cost trade-off: a finer mesh reduces the error while demanding greater computational effort. The constant C depends on $\|\boldsymbol{W}\|$ and q, since the matrix $\boldsymbol{W}$ modulates how the approximation error introduced by replacing $\boldsymbol{z}$ with $\boldsymbol{z}^R$ propagates to the observable output $\boldsymbol{y}$.

Lastly, we obtain the following result concerning the parameter bias.

Theorem 4. *Under Assumptions 1–2, for a fixed R, the bias in the estimation of the vector parameters* $\boldsymbol{\pi} = (\boldsymbol{\beta}', \boldsymbol{\sigma}', \mathrm{vec}(\boldsymbol{W})', \boldsymbol{f}', k_1, \ldots, k_q)'$ *using the LR-SSM* (6.9)–(6.10) *for data generated by the SSM* (2.23)–(2.24) *is of order* $O(h)$*, where h denotes the maximum edge length of the finite element mesh used in the approximation. Specifically,*

$$\left\|\boldsymbol{\pi}_R^\star - \boldsymbol{\pi}\right\| = O(h)$$

where $\boldsymbol{\pi}_R^\star$ *is the pseudo-true parameter vector minimising the Kullback-Leibler divergence (*D_{KL}*) between the true marginal law* $\mathcal{P}_{\boldsymbol{\pi}}$ *of* $\boldsymbol{y}(\boldsymbol{s}, t)$ *with parameter vector* $\boldsymbol{\pi}$ *and the misspecified LR-SSM marginal law* $\mathcal{P}_{\boldsymbol{\vartheta}}^{(R)}$ *of* $\boldsymbol{y}^R(\boldsymbol{s}, t)$ *with vector of parameter* $\boldsymbol{\vartheta}$*, i.e.* $\boldsymbol{\pi}_R^\star = \mathrm{argmin}_{\boldsymbol{\vartheta}}\, D_{KL}(\mathcal{P}_{\boldsymbol{\vartheta}}^{(R)} || \mathcal{P}_{\boldsymbol{\pi}})$*. Furthermore, the parameter* $\boldsymbol{\beta}$ *has no bias.*

The proof is reported in Appendix B.4. Together, Theorems 2–4 provide a comprehensive theoretical foundation supporting the use of the LR-SSM as a principled and efficient approximation to the full SSM, justifying its adoption in practical spatio-temporal inference and prediction tasks.

6.5 Parameter estimation

We estimate the parameter set Π using a maximum likelihood (ML) approach, implemented via the EM algorithm. This method is particularly well-suited for models with latent structures. Our objective is to estimate Π based on the observed data. Specifically, we observe each component $y_j(\boldsymbol{s}, t)$ at spatial locations $\mathcal{S}_{j,t} = \{\boldsymbol{s}_{i,t}\}_{i=1}^{m_{j,t}}$.

We assume that the mesh $\mathcal{G}_R$ —which defines the spatial domain for the latent process within the state-space model described in Equations (6.9)–(6.10)—is given. The procedure used to construct this mesh is detailed in Section 6.5.1.

The complete EM algorithmic steps, the closed-form expressions for the parameter updates, as well as the matrix formulation of the model, are presented in Appendix B.5. We denote by $\boldsymbol{Q}_{\kappa_i}^{-1}$ the covariance matrix of $\boldsymbol{\eta}_i(\mathcal{V}_R, t)$, as specified in (2.1). In Section 6.5.2, we show how the precision matrix $\boldsymbol{Q}_{\kappa_i}$ is constructed at linear cost using the FEM (Lindgren, Håvard Rue, and Lindström, 2011). This makes the maximisation procedure for the estimation of the parameter κ_i with a linear cost instead of a cubic cost.

Note that given the MLE of the parameter set Π, the plug-in predictions at new site $\boldsymbol{s}_0 \in \mathcal{D}$ and time $t = 1, \ldots, T$ are given by

$$\hat{\boldsymbol{y}}(\boldsymbol{s_0}, t) = \boldsymbol{X}'(\boldsymbol{s_0}, t)\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{W}}\boldsymbol{\Psi}_R(\boldsymbol{s}_0)\hat{\boldsymbol{z}}^T(\mathcal{V}_R, t) \tag{6.11}$$

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{y}}}(\boldsymbol{s_0}, t) = \hat{\boldsymbol{W}}\boldsymbol{\Psi}_R(\boldsymbol{s}_0)\,\mathrm{Var}[\hat{\boldsymbol{z}}^T(\mathcal{V}_R, t)]\boldsymbol{\Psi}_R(\boldsymbol{s}_0)'\hat{\boldsymbol{W}}' \tag{6.12}$$

where $\hat{\boldsymbol{z}}^T(\mathcal{V}_R, t)$ denotes the smoothed state process $\boldsymbol{z}(\mathcal{V}_R, t)$, with associated variance $\mathrm{Var}[\hat{\boldsymbol{z}}^T(\mathcal{V}_R, t)]$, estimated by the Kalman smoother variance $\boldsymbol{P}_t^T$ (see the Kalman smoother derivation in Appendix B.5). The vectors $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{W}}$ are the MLE of $\boldsymbol{\beta}$ and $\boldsymbol{W}$, respectively. Equation (6.11) results in the BLUP procedure, and it relies only on linear algebra operations with a linear computational cost of $O(R)$.

6.5.1 Choice of the mesh

Theorem 2 establishes the weak convergence $\boldsymbol{\Psi}_R(\boldsymbol{s})\boldsymbol{z}(\mathcal{V}_R, t) \rightarrow \boldsymbol{z}(\boldsymbol{s}, t)$. This result holds under standard assumptions on the finite element mesh, say $\mathcal{G}_R = (\mathcal{V}_R, \mathcal{E}_R)$, specifically that the minimal mesh angles are bounded away from zero and the maximum edge length tends to zero.

Since the choice of triangulation significantly affects both the quality of the approximation and the asymptotic convergence, it is essential to construct finite element meshes that satisfy these topological criteria. In this section, we introduce a method for generating such meshes, involving R vertices, where R is chosen to balance computational tractability with accurate and reliable parameter estimation.

Let $\mathcal{S}$ denote the set of all observed locations, that is, $\mathcal{S} = \bigcup_{i=1}^{p} \bigcup_{t=1}^{T} \mathcal{S}_{i,t}$, where $\mathcal{S}_{i,t}$ is the set of locations at which variable i is observed at time t.

Among available options, Delaunay triangulations are commonly used to construct finite element meshes because they maximise the smallest angle among all possible triangulations of a given set of points, thereby avoiding the creation of overly narrow triangles. We construct an initial triangulation $\mathcal{G}_m$ by applying the Delaunay algorithm to the full set $\mathcal{S}$, where $m = |\mathcal{S}|$. To reduce the mesh to a fixed number R of vertices, we apply an iterative procedure: at each step, the node associated with the smallest triangle area is removed, followed by local re-triangulation to preserve the Delaunay structure.

This process is repeated until exactly R vertices remain. The resulting triangulation defines the mesh $\mathcal{G}_R = (\mathcal{V}_R, \mathcal{E}_R)$. To ensure that the interior angles in the mesh are suitably bounded away from zero, we apply standard Laplacian smoothing (Field, 1988) to the triangular finite mesh $\mathcal{G}_R$. Laplacian smoothing is an effective ad hoc method for improving mesh quality, particularly in regions with varying element density, by adjusting vertex positions to produce a more regular triangulation. The Laplacian smoothing algorithm for Delaunay triangulations is presented in pseudocode in Algorithm 2. It produces a Laplacian-smoothed version of the Delaunay triangulation, denoted by $\mathcal{G}_R^* = (\mathcal{V}_R^*, \mathcal{E}_R^*)$. The resulting set of smoothed vertices $\mathcal{V}_R^*$ can then be used in place of $\mathcal{V}_R$ in the state equation of the SSM model (6.9)–(6.10). For additional details and visual illustrations demonstrating the effectiveness of this approach in improving angle quality relative to standard Delaunay triangulations, we refer the reader to Field (1988).

Algorithm 2: Laplacian smoothing algorithm for Delaunay triangulations.

Input: Triangulation $\mathcal{G}_R = (\mathcal{V}_R, \mathcal{E}_R)$; angle threshold $\theta_{\min}$

Output: Triangulation with all angles $\geq \theta_{\min}$

while *any triangle angle* $< \theta_{\min}$ **do**
  foreach *interior vertex* $\boldsymbol{r} \in \mathcal{V}_R$ **do**
    Let $\{\boldsymbol{r}_1, \dots, \boldsymbol{r}_k\}$ be the neighbours of $\boldsymbol{r}$;
    Compute $\boldsymbol{r}^* \leftarrow \frac{1}{k} \sum_{i=1}^{k} \boldsymbol{r}_i$;
    if $\boldsymbol{r}^*$ *preserves Delaunay and connectivity* **then**
      Update position: $\boldsymbol{r} \leftarrow \boldsymbol{r}^*$;
    else
      Remove $\boldsymbol{r}$; locally retriangulate;
      Insert vertex at $\boldsymbol{r}^*$;

6.5.2 Approximation of the precision matrix

In this section, we briefly recall how the approximate precision matrix $\boldsymbol{Q}$ is obtained within the SPDE framework of Lindgren, Håvard Rue, and Lindström (2011); further details are given in Appendix B.3.

Let $\mathcal{G}_R = (\mathcal{V}_R, \mathcal{E}_R)$ be a triangulation of the spatial domain $\mathcal{D}$ with $R = |\mathcal{V}_R|$ vertices associated with piecewise–linear basis functions $\{\psi_i\}_{i=1}^R$ of the finite element space. Let $\boldsymbol{C}$ and $\boldsymbol{G}$ denote the mass matrix and stiffness matrix, respectively, defined as

$$\boldsymbol{C}_{ij} = \langle \psi_i, \psi_j \rangle, \qquad \boldsymbol{G}_{ij} = \langle \nabla\psi_i, \nabla\psi_j \rangle.$$

We then define the matrix

$$\boldsymbol{K}_{\kappa_i^2} = \kappa_i^2 \boldsymbol{C} + \boldsymbol{G}.$$

Using the Neumann boundary conditions, the finite-dimensional representations of the solutions to Equation (B.25) have precision

$$\boldsymbol{Q}_{k_i} = \boldsymbol{K}_{\kappa_i^2} \boldsymbol{C}^{-1} \boldsymbol{K}_{\kappa^2}. \tag{6.13}$$

The matrix $\boldsymbol{C}^{-1}$ is dense, which leads to a dense precision matrix in the resulting model. To address this issue, Lindgren, Håvard Rue, and Lindström (2011, Appendix C.5) show that $\boldsymbol{C}$ can be approximated by a diagonal matrix $\widetilde{\boldsymbol{C}}$, where $\widetilde{C}_{ii} = \langle \psi_i, 1 \rangle$. This approximation yields a sparse precision matrix, resulting in a GMRF representation of the original Matérn GF with rescale k_i. Because each basis function overlaps only a fixed, small number of neighbours, $\widetilde{\boldsymbol{C}}$, $\boldsymbol{G}$, $\boldsymbol{K}_{\kappa_i^2}$ contain $O(R)$ non-zero entries. Thus, all matrix products can be assembled with linear complexity $O(R)$. Due to the independence assumption of the q components of the random vector $\boldsymbol{\eta}(\mathcal{V}_R, t)$, the resulting precision matrix is $\boldsymbol{Q} = \text{blockdiag}(\boldsymbol{Q}_{k_1}, \ldots, \boldsymbol{Q}_{k_q})$

The result in (6.13) is derived using Neumann boundary conditions to solve the SPDE (B.25). We summarise the main considerations about the boundary effect and the boundary-mitigation strategy in the next paragraph and refer the reader to Appendix B.6 for further details.

Boundary effects

Imposing Neumann boundary conditions on the field $\boldsymbol{\eta}(\boldsymbol{s}, t)$ induces a folding effect in the corresponding covariance function. This behaviour was analysed in detail by Lindgren, Håvard Rue, and Lindström (2011); Appendix B.6 revisits the main arguments and demonstrates that, at distances greater than twice the Matérn range parameter from the domain boundary, the folded covariance is practically indistinguishable from the stationary Matérn covariance. We leverage this result by enlarging the vertex set through the addition of auxiliary boundary vertices; that is, we work on the augmented finite element mesh $\mathcal{V}_R^* = \mathcal{V}_R \cup \mathcal{V}_R^{\text{aux}}$ where $\mathcal{V}_R^{\text{aux}}$ is chosen so that every vertex in $\mathcal{V}_R$ lies at least two Matérn ranges away from the domain edge. On this extended finite element mesh, we build the SPDE-based precision matrix $\boldsymbol{Q}_{k_i}^*$ as in (6.13). Finally, the target precision matrix is recovered as the principal submatrix

$$\boldsymbol{Q}_{k_i} = \boldsymbol{Q}_{k_i}^*[\, \mathcal{I}, \mathcal{I} \,],$$

where $\mathcal{I}$ indexes the original vertices $\mathcal{V}_R$.

6.5.3 Asymptotic distribution and Wald confidence intervals

For inference on the estimated parameters, we assume that the true data-generating process follows model (2.23)–(2.24), while the proposed low-rank formulation (6.9)–(6.10) serves as a computationally efficient approximation. Specifically, we assume that the number of points R for the construction of the mesh is sufficiently large so that the approximation error is negligible, and small enough to guarantee the estimation of the model with available resources in a short time. In this sense, the low-rank model can be treated as the underlying model for the observed data.

Standard error estimates for $\hat{\Pi}$ can be computed in various ways. For example, it is sometimes possible to evaluate the Hessian matrix after convergence (Shumway, Stoffer, and Stoffer, 2000). Alternatively, one may obtain estimates of the standard error by perturbing the likelihood function and using numerical differentiation (e.g. Shumway, Stoffer, and Stoffer (2000)). However, although the likelihood-based parameter estimates are consistent and asymptotically normal, the asymptotics are often not applicable for the relatively small sample sizes. In that case, a bootstrap procedure is appropriate and is very simple to implement. For example, Stoffer and Wall (1991) describes a simple bootstrap sampling algorithm for parameter estimates in general state-space models that is appropriate for the spatio-temporal setting discussed here. In addition, Wall and Stoffer (2002) describes how bootstrap resampling in this context can also give appropriate estimates of conditional forecast accuracy.

Here we consider inference on $\hat{\Pi}$ by evaluating the Hessian matrix after convergence (Shumway, Stoffer, and Stoffer, 2000). Indeed, under standard regularity conditions, including identifiability of the parameter vector Π, finite fourth moments of the variables involved, and differentiability of the likelihood function with respect to Π, the maximum likelihood estimator $\hat{\Pi}$ is asymptotically normal. For precise assumptions in the context of state-space models and EM-based estimation, see Jensen and Petersen (1999) and McLachlan and Krishnan (2008), respectively. Then:

$$\sqrt{T}\,(\hat{\Pi} - \Pi_0) \;\xrightarrow{d}\; \mathcal{N}\Big(\mathbf{0},\; \boldsymbol{\mathcal{I}}^{-1}(\Pi_0)\Big), \tag{6.14}$$

where $\boldsymbol{\mathcal{I}}(\Pi_0)$ is the observed information matrix, i.e., the negative Hessian of the marginal log-likelihood $\ell(\Pi) = \log p(\boldsymbol{y}_{1:T} \mid \boldsymbol{X}_{1:T}; \Pi)$ evaluated at the true value Π_0. Since $\ell(\Pi)$ is obtained through the Kalman filter and smoother, its second derivatives can be computed analytically. Let

$$\widehat{\boldsymbol{\mathcal{I}}} \;\equiv\; -\frac{\partial^2 \ell(\Pi)}{\partial \Pi\, \partial \Pi'}\bigg|_{\Pi=\hat{\Pi}},$$

denote the observed information evaluated at $\hat{\Pi}$. Then the estimated asymptotic covariance matrix is given by

$$\widehat{\boldsymbol{V}} \;=\; \widehat{\boldsymbol{\mathcal{I}}}^{-1}. \tag{6.15}$$

For each component Π_j of Π, a Wald-type $100(1-\alpha)\%$ confidence interval is given by:

$$\Pi_j \in \left[\hat{\Pi}_j \pm z_{1-\alpha/2}\sqrt{\widehat{V}_{jj}}\right], \tag{6.16}$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the standard normal distribution.

For parameters constrained to the positive real line, such as the variances σ_i^2, we compute intervals on the logarithmic scale and back-transform the bounds to enforce positivity.

6.6 Simulations

In this section, we evaluate the finite-sample performance of the proposed LR-SSM. To this end, we simulate model (2.23)–(2.24) using the parameter vector denoted by Π^0. The inference is carried out using the proposed LR-SSM (6.9)–(6.10). Indeed, Theorem 3 shows that the approximation error when using model (6.9)-(6.10) for inference on a data-generating process of model (2.23)-(2.24) remains uniformly bounded in time. We consider a realistic setup, where the observed process is of dimension $p = 3$ and a latent space of dimension $q = 2$, allowing for a fully heterotopic case.

6.6.1 Experimental design

To study the identifiability of the model defined by (6.9)-(6.10), parametrised by Π^0, synthetic data are generated using the SSM defined by equations (2.23)–(2.24). Let $\mathcal{L}_\delta = \{(i\delta, j\delta) \in \mathbb{R}^2 : i, j \in \{0, 1, \ldots, 24\}\}$ denote a 25×25 regular lattice on the domain $\mathcal{D} = [0,1]^2$, with spacing $\delta = 1/25$ between adjacent points. For each variable $i \in \{1, \ldots, p\}$, the training set $\mathcal{S}_i \subset \mathcal{L}_\delta$ is constructed by selecting a random subset of m locations. The remaining points, $\mathcal{S}_i^* = \mathcal{L}_\delta \setminus \mathcal{S}_i$, are used to evaluate the model's predictive performance.

We consider three different cases, corresponding to $m = 100$, 225, and 400 spatial locations while LR is set to 100%, 75%, 50%, 25% and 15%. As an example, Figure 35 shows the generated finite element mesh for the simulations when $m = 100$. Except for the case $LR = 100\%$, all the other cases consider the Laplacian smoothing of the Delaunay triangulation. Details on Laplacian smoothing and boundary effects are in Section 6.5.1 and Section 6.5.2. Moreover, we consider the effect of time series length by considering $T = 50$ and 100.

Summing up, for each rank R, the simulation campaign is based on six different simulation setups. For each simulation setup, $M = 100$ Monte Carlo replications are performed.

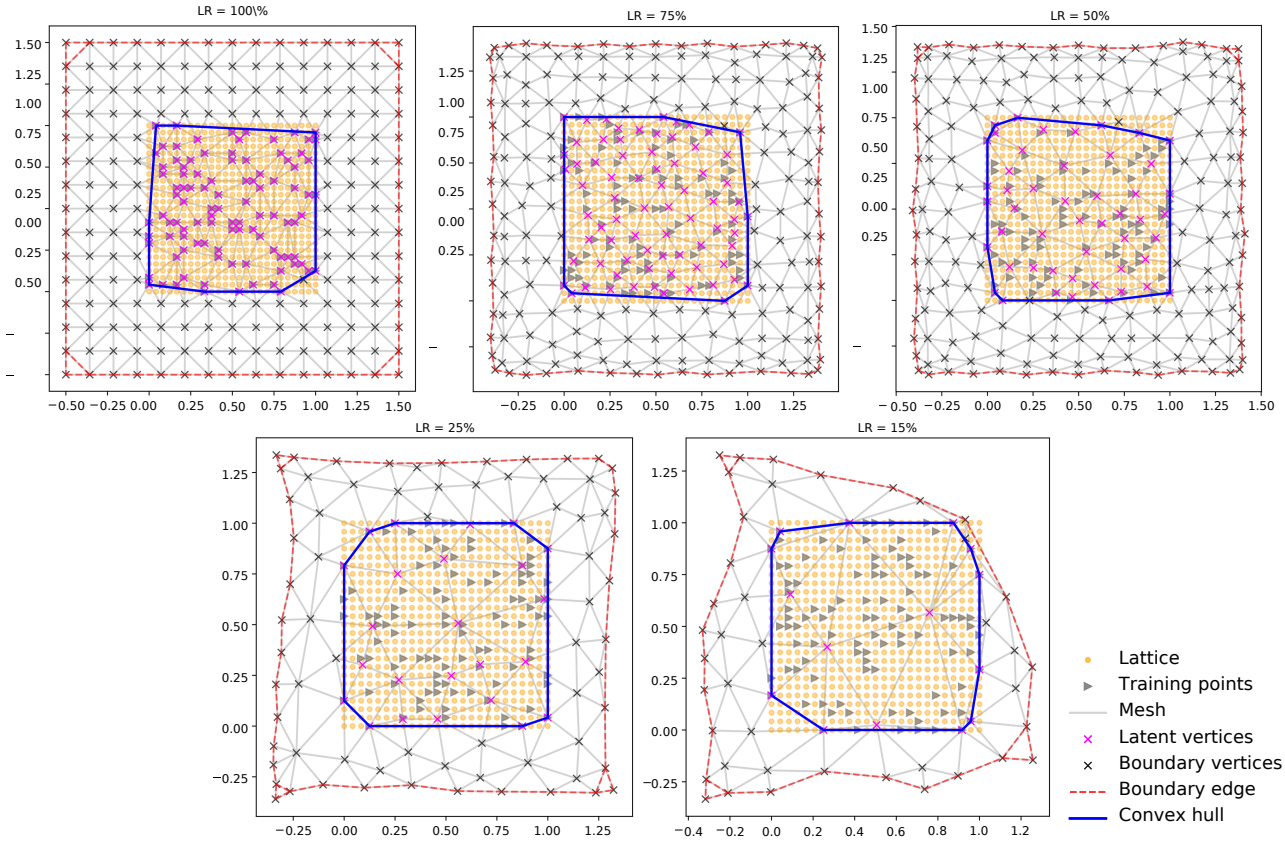


Figure 35: Simulated $m = 100$ spatial points location on the domain $\mathcal{D} = [0,1]^2$, along with the resulting finite element mesh for $LR = 100\%, 75\%, 50\%, 25\%$ and 15%(from left to right, top to bottom). Laplacian smoothing is performed, providing a refined vertices (magenta crosses) of the original Delaunay triangulation of the observed points (light grey triangles). The predictive performances are computed on a regular 25×25 lattice $\mathcal{L}_\delta$ (light orange points). The observed points are shown in light grey. The boundary vertices are shown in grey.

The data-generating process is specified by $\Pi^0 = \{\boldsymbol{\beta}^0, \sigma^{2,0}, \boldsymbol{f}^0, \boldsymbol{W}^0, k_1^0, \dots, k_q^0\}$ with the following parameter values: $\boldsymbol{\beta}^0 = (1, 2, -1)'$, and

$$\boldsymbol{X}'(\boldsymbol{s}, t) = \begin{bmatrix} x_1(\boldsymbol{s}, t) & 0 & 0 \\ 0 & x_2(\boldsymbol{s}, t) & 0 \\ 0 & 0 & x_3(\boldsymbol{s}, t) \end{bmatrix},$$

where each $x_i(\boldsymbol{s}, t)$ is independently drawn from a standard normal distribution $\mathcal{N}(0, 1)$, for all spatial locations $\boldsymbol{s} \in \mathcal{L}_\delta$, time steps $t = 1, \dots, T$, and for $i = 1, 2, 3$. The variance parameters are set to $\boldsymbol{\sigma}^{2,0} = (0.5, 1.5, 1)'$. The autoregressive coefficients are defined as $\boldsymbol{f} = (0.85, -0.5)'$, and the $p \times q$ interaction matrix is given by

$$\boldsymbol{W}^0 = \begin{bmatrix} 0.5 & 1 \\ 0.5 & 0.25 \\ 0.2 & 0.8 \end{bmatrix}.$$

The spatial range parameters are set to $\boldsymbol{\kappa}^0 = (\kappa_1^0, \kappa_2^0)' = (7\sqrt{8}, 2\sqrt{8})'$. The latent variables $\boldsymbol{z}(\boldsymbol{s}, t)$ are simulated on the same spatial grid with initial state drawn from $\boldsymbol{z}(\boldsymbol{s}, 0) \sim \mathcal{N}(1, 1)$.

6.6.2 EM estimation

Once a single dataset j, for $j = 1, \ldots, M$ is simulated, the EM algorithm is executed with particular starting values. Specifically, as starting values for the estimation procedure, we use a random point uniformly drawn from $[0, 1]^2$ for $\boldsymbol{\beta}$, a random point in $(0.2, 0.8) \times (-0.2, -0.8)$ for $\boldsymbol{f}$, each element of $\boldsymbol{W}$ is sampled independently from $\mathcal{U}[0.2, 2]$, each κ_i is drawn from $\mathcal{U}(\sqrt{8}, 7\sqrt{8})$, and each σ_i^2 from $\mathcal{U}[0.1, 2]$.

The maximum iteration number of the EM algorithm is set to 300, while the exit condition is based on the convergence criteria $|\text{logL}_{k+1} - \text{logL}_k|/|\text{logL}_k| < 0.0001$ where logL_k denote the log-likelihood at iteration k.

6.6.3 Simulation results

We evaluate the models by comparing their parameter estimates using the bias and root mean square error (RMSE) where for each $\boldsymbol{\theta} \in \Pi$ the RMSE is computed as $\sqrt{N^{-1} \sum_{i=1}^{M} \|\hat{\boldsymbol{\theta}}^i - \boldsymbol{\theta}_0\|^2}$ where $\hat{\boldsymbol{\theta}}^i$ corresponds to the estimation of the parameter $\boldsymbol{\theta}$ in the ith Monte Carlo simulation, $\|\cdot\|$ is the euclidean norm and $\boldsymbol{\theta}_0$ is the true parameter value. In addition, we assess predictive performance on the validation set $\mathcal{S}_i^*$ for $i = 1, \ldots, p$, by comparing the observed process $y_i(\boldsymbol{s}, t)$ with its estimate $\hat{y}_i(\boldsymbol{s}, t)$. Prediction errors are defined as $e_i(\boldsymbol{s}, t) = y_i(\boldsymbol{s}, t) - \hat{y}_i(\boldsymbol{s}, t)$, for $\boldsymbol{s} \in \mathcal{S}_i^*$. The different LR-SSM versions are then compared in terms of RMSE computed over these validation errors.

Tables 22, 23 presents the bias and RMSE of the estimated parameters under each combination of m and T for $LR = 100\%, 75\%$ and $LR = 15\%$ respectively (the simulation results for the cases with $LR = 75\%$ and $LR = 25\%$ are showed in the Supplementary material). The results indicate that estimation accuracy improves as the number of spatial locations and time steps increases, as evidenced by decreasing RMSE and absolute bias values across all parameters. The ML estimator of $\boldsymbol{\beta}$ has an RMSE close to zero, improving with increasing observed data both in T and m. The noise variances $\boldsymbol{\sigma}^2$ show more pronounced improvement with increasing m. Spatial range parameters $\boldsymbol{\kappa}$ are more challenging to estimate accurately; although RMSE decreases with finer grids, a noticeable bias persists. This is a known issue in spatial statistics, as range parameters often exhibit weak identifiability, particularly when data are limited in spatial resolution or extent (H. Zhang, 2004). Moreover, the RMSE do not decrease fast to zero and maybe do not converge for $T \to \infty$. Note

that this is not surprising as it is consistent with asymptotic theory, which requires consistent initial estimates. The ML estimator of $\boldsymbol{W}$ has small RMSE, and the accuracy improves with increasing time length T. Similar behaviour is shown by the estimator of the temporal autocorrelation parameters $\boldsymbol{f}$.

	T = 50						T = 100					
	m = 100		m = 225		m = 400		m = 100		m = 225		m = 400	
Parameter	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
β_1	0.002		0.000		0.001		0.000		0.001		-0.001	
β_2	0.001	0.017	0.001	0.011	0.001	0.008	0.003	0.012	0.001	0.007	0.000	0.006
β_3	-0.008		-0.001		-0.001		0.001		0.000		0.000	
σ_1^2	0.160		0.067		0.013		0.150		0.060		0.014	
σ_2^2	0.278	0.193	0.165	0.105	0.079	0.048	0.266	0.180	0.153	0.096	0.079	0.046
σ_3^2	0.121		0.055		0.019		0.114		0.052		0.013	
k_1	-3.526	0.207	-3.842	0.200	-4.767	0.262	-3.097	0.174	-3.084	0.159	-4.386	0.236
k_2	0.390		-0.368		-0.414		0.619		-0.185		-0.355	
W_1	0.018		-0.022		0.005		0.051		-0.002		0.041	
W_2	0.143		0.226		0.191		0.154		0.215		0.197	
W_3	-0.026	0.177	-0.032	0.223	-0.025	0.259	0.014	0.172	-0.009	0.206	0.022	0.280
W_4	0.007		0.046		0.046		0.014		0.043		0.035	
W_5	0.018		0.004		0.020		0.036		0.009		0.030	
W_6	0.121		0.180		0.149		0.127		0.177		0.165	
f_1	-0.067	0.133	-0.052	0.094	-0.083	0.169	-0.082	0.121	-0.056	0.077	-0.074	0.141
f_2	-0.021		-0.007		-0.015		-0.008		-0.006		-0.017	
$\text{RMSE}_{\text{train}}$	1.022		0.989		0.964		1.015		0.983		0.961	
$\text{RMSE}_{\text{test}}$	1.233		1.178		1.143		1.217		1.158		1.124	
Runtime (hours)	0.11		0.49		1.65		0.15		0.74		3.02	

Table 22: $LR = 100\%$. Bias and RMSE of the estimated parameters are reported for different numbers of time steps (T) and spatial locations (m). $RMSE_{\text{train}}$ and $RMSE_{\text{test}}$ denote the average prediction RMSEs computed on the training set $\mathcal{S}_i$ and the test set $\mathcal{S}_i^*$, respectively, across $M = 100$ Monte Carlo replications. Time indicates the average computation time over replications.

Overall, as expected for the ML estimation, these results demonstrate that the proposed model yields consistent and increasingly accurate parameter estimates as more spatial and temporal information becomes available, confirming the robustness of the estimation procedure under increasing low-rank approximation. This means that, in general, the low-rank approximation can lead to a speed-up in the computation without losing a precise estimate. Furthermore, the estimate shows that the model performs well in the heterotopic case, meaning that it is able to estimate the cross-correlation among the observed processes. By evaluating the test RMSE under increasing levels of low-rank approximation R, while keeping m fixed, we observe a slight increase in RMSE that is not statistically significant compared to the case $LR = 100\%$. This demonstrates that the proposed approximation procedure substantially reduces computation time while maintaining consistent parameter estimates and accurate spatial mapping performance. The computation time refers to a machine equipped with an Intel Xeon Platinum 8460Y+ CPU and 1007 GB RAM using Python 3.10.15.

	T = 50						T = 100					
	m = 100		m = 225		m = 400		m = 100		m = 225		m = 400	
Parameter	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
β_1	-0.002		0.000		0.001		-0.004		-0.001		0.000	
β_2	-0.004	0.018	0.002	0.011	-0.001	0.008	0.004	0.013	-0.001	0.007	-0.001	0.006
β_2	-0.004		0.001		0.001		-0.001		-0.001		0.001	
σ_1^2	0.435		0.274		0.199		0.419		0.256		0.180	
σ_2^2	0.324	0.304	0.207	0.190	0.149	0.137	0.332	0.298	0.214	0.183	0.141	0.125
σ_3^2	0.153		0.083		0.053		0.142		0.073		0.043	
k_1	-0.527	0.316	-1.407	0.145	-2.224	0.137	-0.295	0.220	-0.788	0.096	-1.426	0.088
k_2	-0.228		-0.463		-0.396		0.002		-0.412		-0.280	
W_1	0.029		-0.018		-0.016		0.081		0.003		-0.001	
W_2	0.170		0.214		0.197		0.153		0.227		0.222	
W_3	-0.006	0.196	-0.033	0.216	-0.025	0.203	0.036	0.195	-0.006	0.220	-0.004	0.218
W_4	0.047		0.067		0.064		0.043		0.074		0.066	
W_5	0.014		-0.002		0.000		0.038		0.005		0.004	
W_6	0.139		0.177		0.166		0.136		0.190		0.188	
f_1	-0.036	0.097	-0.027	0.083	-0.036	0.072	-0.056	0.131	-0.031	0.058	-0.039	0.062
f_2	0.002		-0.004		-0.013		-0.001		-0.012		-0.005	
$\text{RMSE}_{\text{train}}$	1.088		1.042		1.022		1.084		1.038		1.015	
$\text{RMSE}_{\text{test}}$	1.230		1.169		1.132		1.223		1.158		1.121	
Runtime (hours)	0.05		0.17		0.64		0.05		0.23		0.88	

Table 23: $LR = 50\%$. Bias and RMSE of the estimated parameters are reported for different numbers of time steps (T) and spatial locations (m). $RMSE_{\text{train}}$ and $RMSE_{\text{test}}$ denote the average prediction RMSEs computed on the training set $\mathcal{S}_i$ and the test set $\mathcal{S}_i^*$, respectively, across $M = 100$ Monte Carlo replications. Time indicates the average computation time over replications.

6.7 Analysis of a multivariate spatio-temporal air quality dataset

We now illustrate how the model is used to analyse a large spatio-temporal dataset with multiple pollutants, heterogeneous spatial coverage, and substantial missingness.

We focus on two pollutants: coarse particulate matter with aerodynamic diameter less than 10 μm (PM_{10}) and fine particulate matter with diameter less than 2.5 μm ($PM_{2.5}$) in Italy. These pollutants are typically observed at partially different monitoring stations (heterotopic) and are known to be highly correlated. Fassò, Finazzi, and Ndongo (2016) shows that the estimated correlation between co-located PM_{10} and $PM_{2.5}$ at the European level is close to 0.93 after controlling for meteorology and orography. Hence, leveraging this correlation, the aim is to obtain maps of daily concentrations at the national level. This requires adopting a statistical model capable of handling the complexity of the bivariate pollutant phenomenon, while controlling for meteorology and orography, as well as accounting for an unbalanced monitoring network, heterotopic observations, and extensive missing data.

	T = 50						T = 100					
	m = 100		m = 225		m = 400		m = 100		m = 225		m = 400	
Parameter	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
β_1	-0.001		-0.002		-0.001		0.000		0.000		-0.001	
β_2	-0.005	0.018	-0.001	0.011	0.000	0.008	0.002	0.012	0.001	0.008	-0.001	0.006
β_3	-0.001		0.001		0.000		0.001		-0.002		0.000	
σ_1^2	0.813		0.556		0.416		0.804		0.541		0.394	
σ_2^2	0.582	0.562	0.403	0.381	0.301	0.283	0.577	0.553	0.412	0.375	0.303	0.273
σ_3^2	0.300		0.179		0.125		0.289		0.169		0.113	
k_1	-8.255	0.502	-0.062	0.311	7.295	0.572	-6.818	0.453	-0.250	0.270	8.843	0.675
k_2	3.349		1.015		0.471		3.607		0.948		0.840	
W_1	0.711		0.360		0.201		0.691		0.406		0.242	
W_2	0.215		0.140		0.206		0.278		0.173		0.197	
W_3	0.144	0.750	0.204	0.384	0.169	0.322	0.175	0.755	0.239	0.424	0.217	0.350
W_4	0.093		0.058		0.093		0.121		0.073		0.092	
W_5	0.643		0.241		0.115		0.599		0.269		0.124	
W_6	0.179		0.111		0.160		0.225		0.138		0.150	
f_1	-0.780	0.869	-0.227	0.299	-0.076	0.116	-0.741	0.839	-0.227	0.269	-0.085	0.115
f_2	0.073		0.047		0.039		0.086		0.052		0.053	
$\text{RMSE}_{\text{train}}$	1.222		1.147		1.106		1.217		1.144		1.100	
$\text{RMSE}_{\text{test}}$	1.285		1.210		1.165		1.283		1.205		1.158	
Runtime (hours)	0.03		0.05		0.14		0.03		0.06		0.24	

Table 24: $LR = 15\%$. Bias and RMSE of the estimated parameters are reported for different numbers of time steps (T) and spatial locations (m). $RMSE_{\text{train}}$ and $RMSE_{\text{test}}$ denote the average prediction RMSEs computed on the training set $\mathcal{S}_i$ and the test set $\mathcal{S}_i^*$, respectively, across $M = 100$ Monte Carlo replications. Time indicates the average computation time over replications.

6.7.1 Data description

We use the AQCLIM-GRINS dataset provided by Fusta Moro and Fassò (2025), which contains daily air quality and meteorological measurements across Italy, covering the five-and-a-half-year period from July 1, 2018, to December 30, 2023. The AQCLIM-GRINS dataset extends the AGRIMONIA dataset (Fassò, Rodeschini, et al., 2023a) to the entire Italian territory and has been harmonised by the GRINS project (`https://www.grins.it/`, accessed 26/07/2025).

The particulate matter data consists of daily average concentrations (in $\mu g/m^3$) collected at 661 ground-level monitoring stations irregularly distributed across Italy. Also, we include the following covariates: Altitude (in metres above sea level), Temperature at 2 metres (in °C), Relative Humidity (RH; in %), and Wind Speed (in m/s), along with an intercept term. Table 25 summarises the variables used in this study and the associated main statistics.

It can be noticed that the resulting measurement network in Figure 36 is partially heterotopic, meaning the variables are observed at intersecting but non-identical sets of locations. The considered period spans $T = 2009$ days, and the dataset includes approximately 1.15 million observations of PM_{10} from 574 monitoring stations, with 28.7% missing values, and about 0.6 million observations

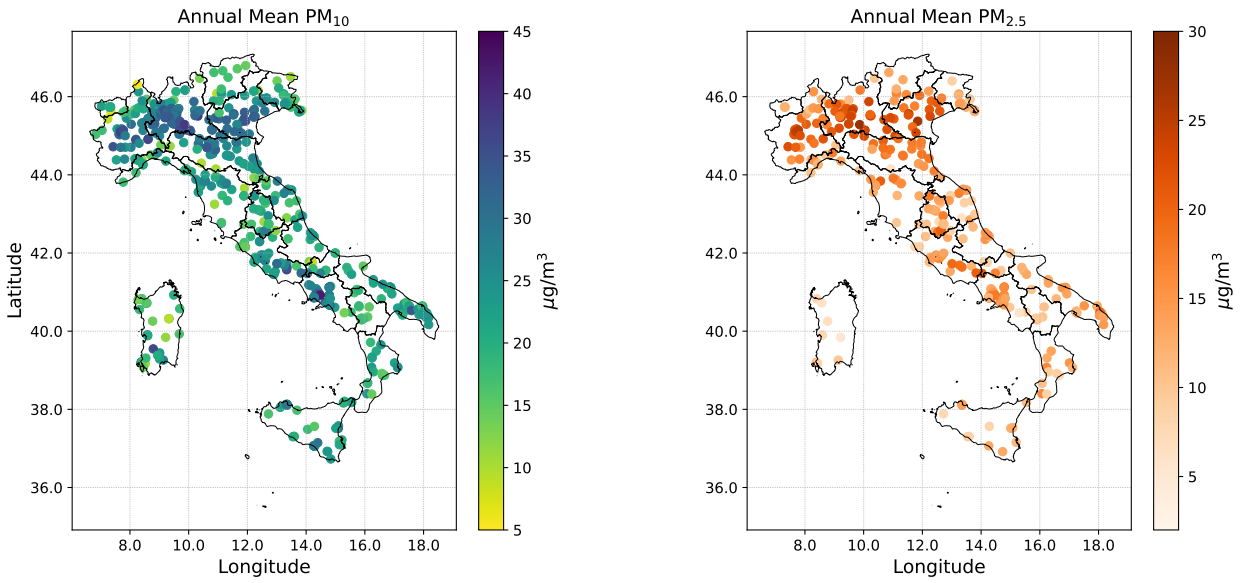


Figure 36: Monitoring network and overall averages of PM_{10} (left) and $PM_{2.5}$(right) concentration over Italy. A total of $S = 661$ stations are available: $S_1 = 574$ stations measuring PM_{10}, and $S_2 = 313$ measuring $PM_{2.5}$, with 304 stations providing co-located measurements.

of $PM_{2.5}$ from 313 stations, with 34.1% missing values. In the sequel, concentrations above the 99.9 percentile are considered outliers and replaced by a missing value.

6.7.2 Preliminary analysis

Figure 36 highlights spatial patterns in pollutant average concentrations: nearby stations tend to record similar values, suggesting the presence of spatial autocorrelation and cross-correlation between the two particulate matter variables. This is confirmed by Figure 37 reporting the weighted least squares Matérn variograms and covariogram of station averages for PM_{10} and $PM_{2.5}$ (Wackernagel, 2003; Cressie, 1985). The variograms reveal a nugget effect, with estimated measurement error std $\hat{\sigma}_i$ of 2.88 $\mu g/m^3$ for PM_{10} and 2.24 $\mu g/m^3$ for $PM_{2.5}$. Accordingly, the model of the next section includes a measurement error component as specified in Equation (6.10). The estimated marginal variances of the spatial process are approximately 15.4 and 6.8 $(\mu g/m^3)^2$. As expected, the autocovariances of the two detrended processes decrease with distance. The estimated spatial ranges are about 0.37° (41 km) and 0.30° (33 km), respectively for PM_{10} and $PM_{2.5}$. Similarly, the correlation between PM_{10} and $PM_{2.5}$ is particularly high, see Figure 37 (right panel), and has a spatial range of 0.37° (41 km), which is close to the PM_{10} behaviour since networks are almost overlapping.

The temporal variation of the observed PM_{10} and $PM_{2.5}$ concentrations is shown as daily averages among the sites in Figure 38 (left panel). Not surprisingly, there is a clear seasonality with higher concentrations in winter. Moreover, the empirical autocorrelation function (ACF) is shown in Figure

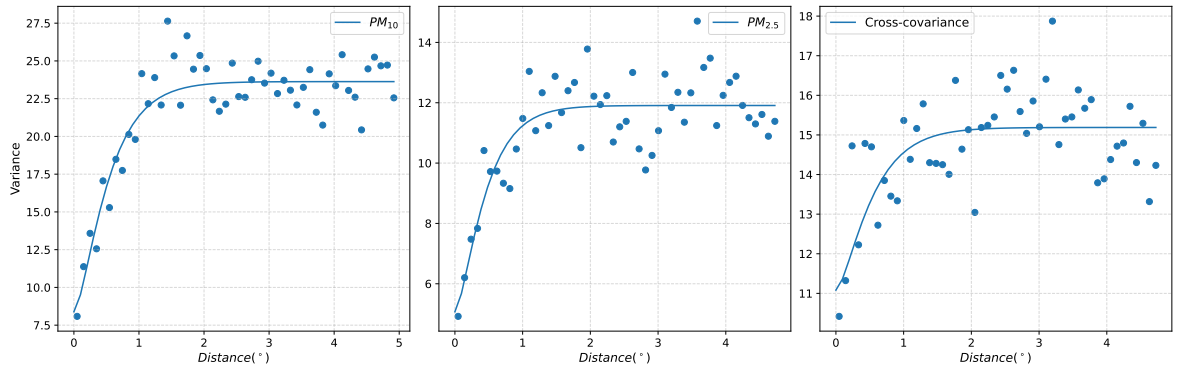


Figure 37: Spatial variograms and covariogram computed from the average concentration at each spatial location in $(\mu g/m^3)^2$. The plots show the variogram of PM_{10} (left), the variogram of $PM_{2.5}$ (center), and the cross-covariogram between PM_{10} and $PM_{2.5}$ (right). The fitted Matérn models are estimated with weighted least squares (Cressie, 1985).

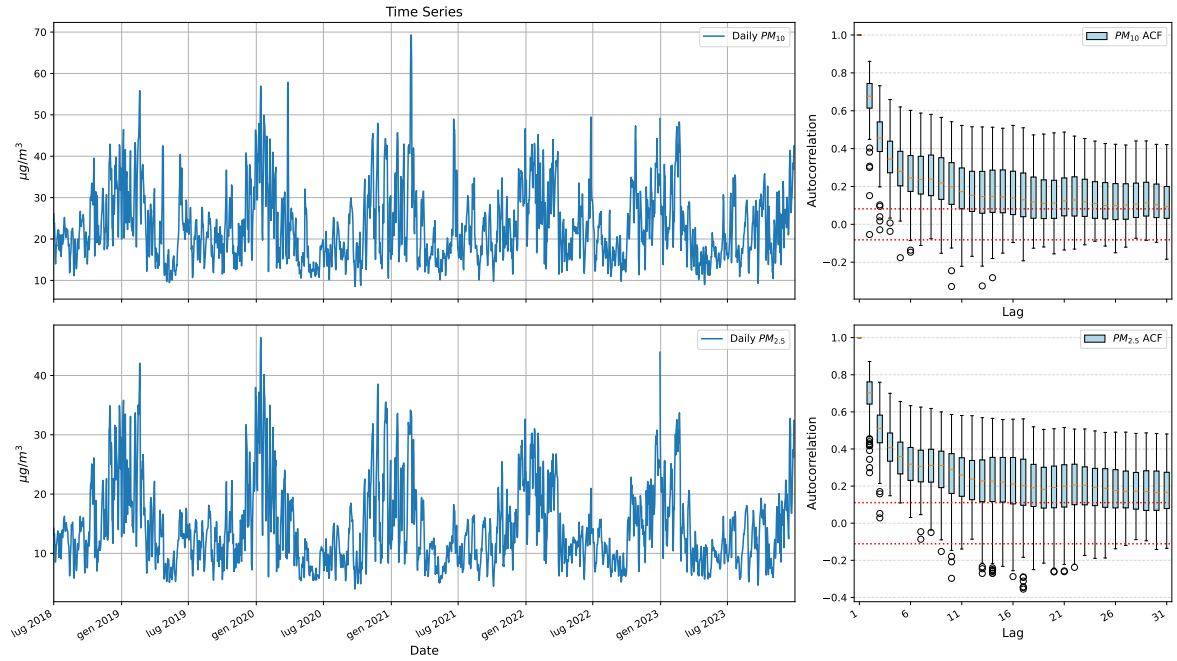


Figure 38: Temporal structure of particulate matter concentrations. The left column shows daily averages of PM_{10} (top) and $PM_{2.5}$ (bottom); the right column displays corresponding autocorrelation functions (ACFs) over the first 30 lags, summarised across stations using boxplots.

38. The first lag is close to 0.7 for both pollutants, indicating the presence of a temporal dependence. Time series and ACF grouped by regions are summarised in the Supplementary material.

6.7.3 Application of the LR-SSM

To apply the proposed LR-SSM (6.9)–(6.10) defined in Section 6.4, we discretise the considered domain using a Delaunay triangulation to form the spatial meshes $\mathcal{V}_R^j$ for $j = 1, 2$. This triangulation maximises the smallest angle among all triangles, promoting mesh regularity. To further ensure good mesh quality, we apply the Laplacian smoothing with a minimum internal angle threshold of 0.15°.

Table 25: Summary statistics of variables from the AQCLIM-GRINS dataset across all locations.

Statistics Unit of measurement	PM_{10} $\mu g/m^3$	$PM_{2.5}$ $\mu g/m^3$	Altitude m	Temperature °C	RH% 	Wind Speed m/s
Percentiles						
Minimum	0.00	0.00	-2.00	-24.68	15.27	0.11
5th	7.00	4.00	4.00	1.63	52.95	0.71
25th	13.76	8.00	26.00	8.31	66.59	1.18
50th	20.00	12.00	114.00	13.84	75.70	1.64
75th	29.55	19.00	271.00	20.29	83.63	2.45
95th	56.80	41.40	736.00	26.32	92.01	4.64
Maximum	137.40	102.60	1770.00	35.75	99.98	16.54
Descriptive statistics						
Mean	24.14	15.66	209.31	14.03	74.51	2.02
Standard deviation	16.20	12.49	274.34	7.84	11.99	1.29
Skewness	1.98	2.28	2.43	-0.14	-0.49	2.02
Kurtosis	5.69	6.87	7.61	-0.47	-0.12	5.82
Number of stations m	574	313	661	661	661	661
Missing proportion	0.29	0.34	0.00	0.00	0.00	0.00

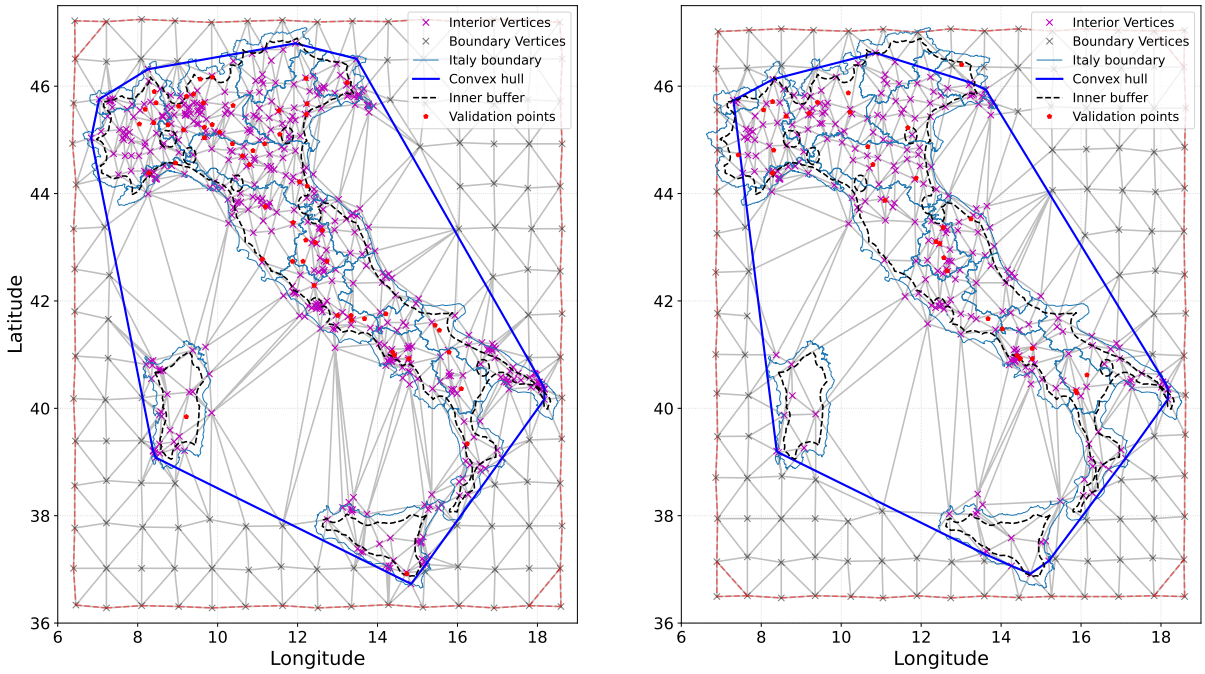


Figure 39: Spatial meshes generated via Laplacian smoothing. The left panel corresponds to the mesh constructed for the PM_{10} dataset, and the right panel to the $PM_{2.5}$ dataset. Interior vertices (magenta), boundary vertices (green), convex hulls (blue), inner buffer (black), and validation locations (red) are shown.

Additional details on the triangulation and smoothing procedures are given in Section 6.5.1. The resulting discretised domain is shown in Figure 39.

We validate the model using a data split validation strategy tailored to the spatio-temporal setting (see Otto, Fassò, and Maranzano, 2024, for a review). Specifically, we adopt a leave-many-stations-out validation scheme. Basically, the entire time series of a subset of stations (the validation set) is withheld during training and later used to assess the model's predictive performance. We construct the validation set by randomly selecting 10% of the total stations of each pollutant, while the remaining 90% are used for training. To avoid edge effects, we exclude stations close to the domain edge by

Table 26: Performance and computational statistics of the estimation procedure across varying levels of low-rank approximation LR. The table reports the number of vertices (R) used in the approximation for each observed pollutant PM_{10} and $PM_{2.5}$, RMSE and R^2 scores for both training and validation sets, as well as estimation time, and number of EM iterations. The computation time refers to a machine equipped with an Intel Xeon Platinum 8460Y+ CPU and 1007 GB RAM using Python 3.10.15.

LR	R	**Variable**	**Train**		**Validation**		**Runtime**	**EM Iterations**
			RMSE	R^2	RMSE	R^2		
100%	516	PM_{10}	3.52	0.95	6.88	0.81	3h 37m	46
	281	$PM_{2.5}$	2.24	0.96	5.44	0.80		
75%	387	PM_{10}	4.26	0.92	6.87	0.81	1h 51m	38
	211	$PM_{2.5}$	2.94	0.93	5.47	0.80		
50%	258	PM_{10}	5.00	0.89	6.44	0.83	1h 00m	35
	141	$PM_{2.5}$	3.51	0.90	5.62	0.79		
25%	129	PM_{10}	6.07	0.84	7.31	0.78	16m	33
	71	$PM_{2.5}$	4.43	0.84	6.26	0.74		

selecting only stations located at least 0.15° away from the boundary (see the inner buffer in Figure 39). The resulting training set includes 516 and 281 spatial points, while the validation set includes 58 and 32 spatial points for PM_{10} and $PM_{2.5}$, respectively.

To validate the low-rank approach and given a set of candidate R values, the mean square error optimality is of interest. Hence, considering $\hat{y}_i(\boldsymbol{s}, t)$ as in (6.11), errors $e_i(\boldsymbol{s}, t) = y_i(\boldsymbol{s}, t) - \hat{y}_i(\boldsymbol{s}, t)$ with $\boldsymbol{s} \in \mathcal{D}$ are computed and the LR-SSM candidates are compared in terms of RMSE and R^2 statistic, where, for each pollutant i, $R_i^2 = 1 - \sum_{t=1}^{T} \sum_{\boldsymbol{s}\in\mathcal{S}_{i,t}} (y_i(\boldsymbol{s}, t) - \hat{y}_i(\boldsymbol{s}, t))^2 / \sum_{t=1}^{T} \sum_{\boldsymbol{s}\in\mathcal{S}_{i,t}} (y_i(\boldsymbol{s}, t) - \bar{y}_i)^2$, where $\bar{y}_i = T^{-1} \sum_{t=1}^{T} |\mathcal{S}_{i,t}|^{-1} \sum_{\boldsymbol{s}\in\mathcal{S}_{i,t}} y_i(\boldsymbol{s}, t)$.

We consider low-rank approximations where LR is set to 100%, 75%, 50%, and 25%. Table 26 summarises the validation results across models. As expected, for both pollutants, the training RMSE increases monotonically as the degree of approximation increases, while the test RMSE shows only a slight increase, at most by 6% in the worst case. These results confirm that the low-rank model preserves predictive accuracy while significantly reducing computational time. Remarkably, even at the roughest approximation level (25% of the rank), the validation RMSE increases by only 6% for PM_{10} and 15% for $PM_{2.5}$ compared to the full-rank model, while computational time is reduced from 3h 37m to just 16m. This demonstrates that substantial efficiency gains can be achieved without compromising accuracy in a practically meaningful way. Notably, across all low-rank configurations, the test RMSEs remain comparable to those reported in Rodeschini, Fassò, et al. (2024) and Fassò, Finazzi, and Ndongo (2016).

Table 27 reports the ML estimates and the associated uncertainty of the parameter vector $\hat{\Pi}_{MLE}$ for different low-rank approximation levels. For both PM_{10} and $PM_{2.5}$, altitude exhibits a small negative association, indicating that higher elevation is associated with lower pollution levels. Temperature shows a negative effect, especially for $PM_{2.5}$. Relative Humidity has a positive correlation with

Table 27: ML estimates of regression coefficients and random effect parameters for the bivariate model of PM_{10} and $PM_{2.5}$, for different low-rank approximation levels (LR = 100%, 75%, 50% and 25%). Covariates include Altitude, Temperature at 2 metres, RH, and Wind Speed. Values are followed by their standard deviations in parentheses..

	$LR = 100\%$		$LR = 75\%$		$LR = 50\%$		$LR = 25\%$	
Parameter	PM_{10}	$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}	$PM_{2.5}$
Intercept	15.762	12.154	16.809	12.694	16.814	13.145	15.220	12.673
	(0.029)	(0.026)	(0.034)	(0.032)	(0.038)	(0.036)	(0.044)	(0.044)
Altitude	-0.012	-0.007	-0.014	-0.008	-0.012	-0.007	-0.012	-0.009
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Temperature	-0.055	-0.236	-0.060	-0.239	-0.056	-0.257	-0.053	-0.269
	(<0.001)	(<0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
RH	0.088	0.062	0.082	0.057	0.074	0.054	0.096	0.069
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Wind Speed	-0.760	-0.874	-0.854	-0.863	-0.788	-0.830	-0.837	-0.926
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
W_1	6.998	2.439	6.515	2.575	6.200	2.752	6.008	3.375
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
W_2	3.467	3.688	3.351	3.356	3.425	3.198	3.584	3.184
	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)
σ_1^2	18.535	7.825	24.642	11.633	30.808	15.107	41.300	21.828
	(0.008)	(0.004)	(0.012)	(0.008)	(0.016)	(0.011)	(0.025)	(0.020)
f	0.823	0.869	0.827	0.889	0.844	0.903	0.895	0.946
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
k	3.349	3.232	2.676	2.197	2.118	1.844	1.578	1.904
	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.004)

both pollutants, more pronounced for PM_{10}. Wind Speed has a substantial negative effect on both processes, with higher winds contributing to the dispersion of pollutants. The results are coherent with the analysis of Rodeschini, Fassò, et al. (2024).

Finally, Figures 40, 41, 42, and 43 show the estimated average random effect component of the PM_{10} across the study domain. Each map depicts the mean over all available years, corresponding to the average over time of $\hat{\boldsymbol{W}}\boldsymbol{H}(\boldsymbol{s})\hat{\boldsymbol{z}}_t$ and its standard deviation, as defined in the predictive equations (6.11) and (6.12). As expected, prediction performance decreases as the approximation level increases. Nevertheless, the average standard deviation remains low across all cases. The magnitude of the random effects suggests some level of model misspecification, specifically, that in certain areas, the available covariates are insufficient to fully capture the variability of the process. In Figure 40, the Po Valley in Northern Italy, one of the most polluted regions in Europe, is clearly distinguishable due to its large positive random effects. This result aligns with findings from other studies focusing on Northern Italy, such as Rodeschini, Fassò, et al. (2024). As the approximation level increases (LR decreases), the latent process becomes smoother, which is also reflected in the reduced values of the estimated rescaling parameters k_i. In the approximation case $LR = 25\%$ in Figure 43, some local variability present in the data is lost. Indeed, the difference from the full-rank case (i.e., $LR = 100\%$) is especially noticeable in regions with a high density of stations but where the mesh is relatively sparse. Regarding the estimated uncertainty, it remains consistent with the test RMSE, which is approximately 6 $\mu g/m^3$.

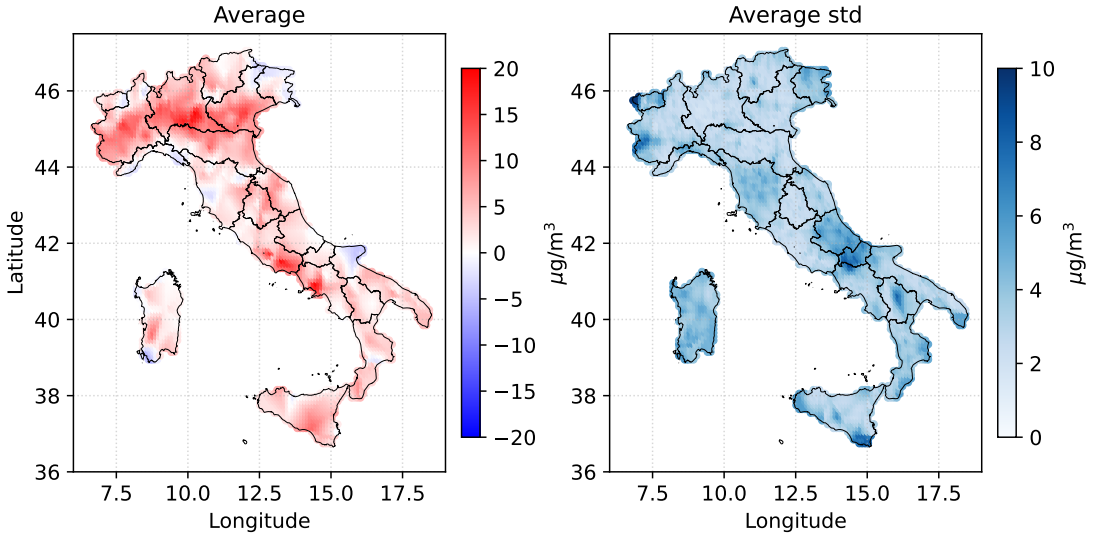


Figure 40: Estimated average latent effect component on PM_{10} (left) and associated with the average standard deviation (right).

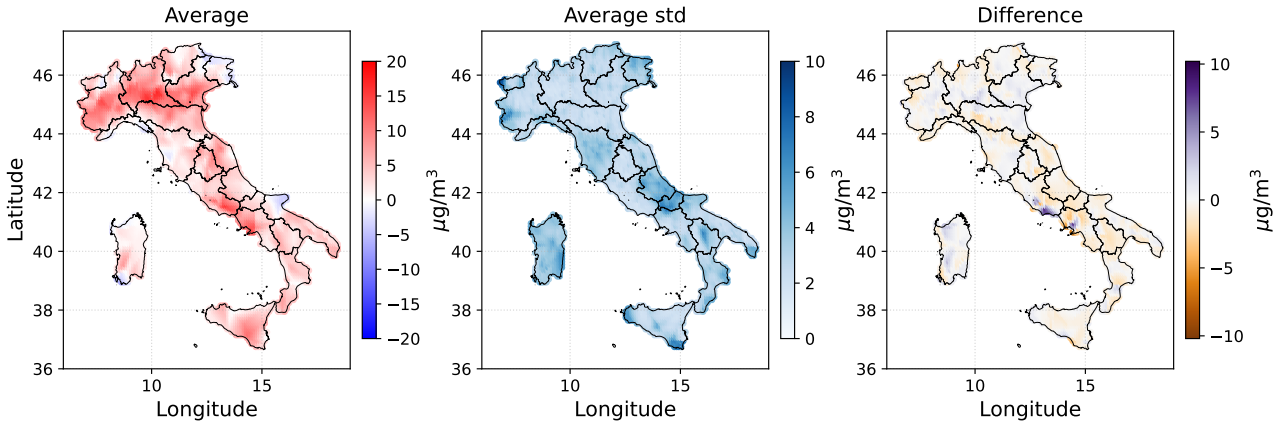


Figure 41: $LR = 75\%$. Estimated average latent effect component on PM_{10} (left), associated with the average standard deviation (right) and average difference between $LR = 100\%$ and $LR = 75\%$ estimations (right).

Notably, no boundary effects are observed, and the uncertainty remains stable even near the edges of the domain.

6.8 Software implementation

The increasing complexity and size of modern spatio-temporal datasets make software implementation and computational efficiency central to statistical practice. Beyond theoretical models, scalable algorithms and practical tools are essential to handle high-dimensional data and extract meaningful insights. This necessity is particularly evident in large-scale applications, such as satellite remote sensing or national monitoring programs, where traditional methods quickly become computationally infeasible.

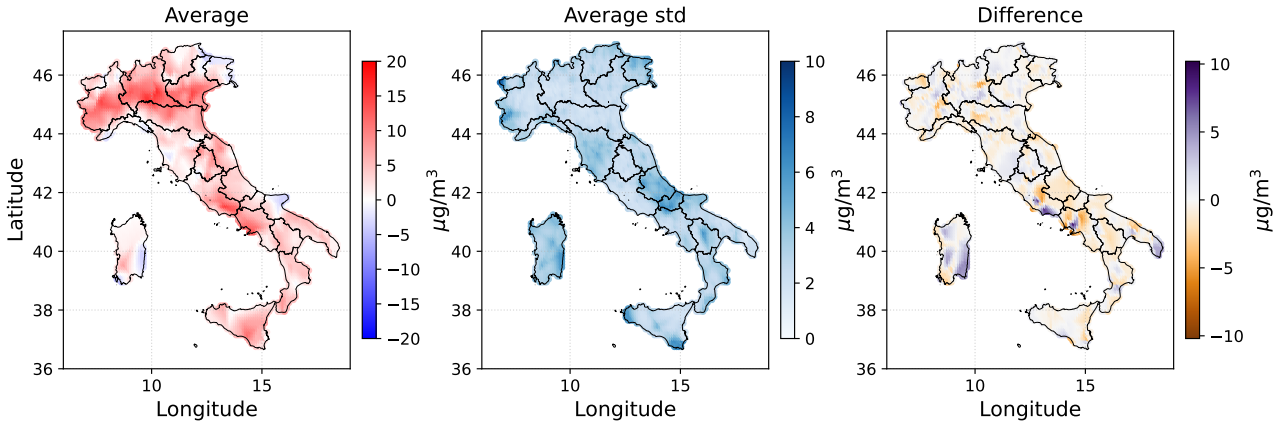


Figure 42: $LR = 50\%$. Estimated average latent effect component on PM_{10} (left), associated with the average standard deviation (right) and average difference between $LR = 100\%$ and $LR = 50\%$ estimations (right).

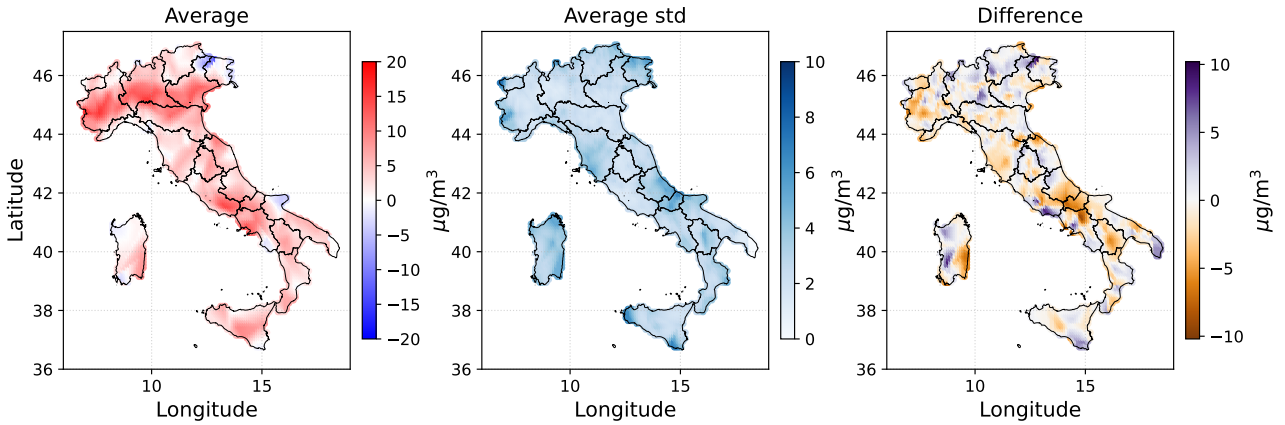


Figure 43: $LR = 25\%$. Estimated average latent effect component on PM_{10} (left), associated with the average standard deviation (right) and average difference between $LR = 100\%$ and $LR = 25\%$ estimations (right).

Accounting for multivariate spatio-temporal dependencies is particularly critical in the "big data" era. For example, the GOES-16 geostationary satellite captures high-resolution spatial and temporal details, recording radiance across 16 wavelength bands with spatial resolutions as fine as 1 km and temporal intervals as short as one minute. Such data have been utilised, for instance, by Guinness (2022) to estimate stationary multivariate spatial and spatio-temporal spectra. Similarly, Datta et al. (2016) analysed over 114,000 georeferenced forest biomass plots from the USDA FIA program, encountering memory constraints due to the large distance matrices required for spatial modelling. These challenges have spurred the development of scalable methodologies to efficiently process and analyse large spatial datasets. Moreover, Heaton et al. (2019) directly addresses the high-dimensional data issue in the spatial context, evaluating the limitations of traditional models and exploring novel approaches to improve tractability and efficiency.

Scalability plays a critical role in this context (see Jordan (2013)). Broadly, one can distinguish between strong and weak scalability. The former is relevant when working with fixed-size datasets,

where increasing computational resources should reduce runtime. For example, matrix operations such as transposition or exponentiation on a fixed-size matrix should ideally scale well across multiple processors, with each processor handling a subset of the data in parallel. Weak scalability, by contrast, becomes important when the problem size grows proportionally with the number of processors. In such cases, typical of larger spatial datasets or finer temporal resolutions, the computational workload per processor remains constant. Operations like matrix creation and deformation, which involve reshaping or partitioning data for independent processing, are well-suited to weak scalability.

These scalability considerations are particularly relevant because spatio-temporal models often rely on a recurring set of computationally intensive operations. Matrix creation, transposition, deformation, and exponentiation facilitate reshaping spatial data, modelling temporal or spatial relationships, and analysing higher-order dependencies. This approach is exemplified in Li, Yuping Wang, and Fang (2024), where higher-order spatial autoregressive models are effectively analysed. Sorting large arrays is another frequent pre-processing step, necessary for tasks such as spatial indexing (essential in spatial databases), nearest-neighbour searches in geostatistics (Altman, 1992), optimising sparse matrix operations for large covariance structures (Hsieh et al., 2011), and organising results for further analysis or visualisation. Matrix operations such as cross-products, inversion (e.g., for OLS estimation of regression models), eigenvalue decomposition, and Cholesky decomposition form the backbone of many statistical methods. Additionally, Fast Fourier Transforms (FFT) are vital for decomposing time series or spatial data into frequency components, aiding in the analysis of periodic or cyclic behaviours (Guinness, 2019). Recursive algorithms also play a significant role, especially within MCMC and EM estimation procedures: the former being widely used in Bayesian spatial statistics (e.g., Otto, Doğan, and Taşpınar (2023)), and the latter in frequentist approaches. Finally, robust input/output operations (e.g., reading and writing large CSV files) are crucial for efficiently handling high-dimensional datasets, ensuring seamless integration between data storage and analysis workflows.

In Tedesco, Rodeschini, and Otto (2025), a comprehensive evaluation of the computational performance of `MATLAB`, `Julia`, `R`, and `Python` for spatial and spatio-temporal modelling was carried out, focusing on high-dimensional datasets typical in geospatial analysis. The study benchmarked each language across key tasks, including matrix operations, mathematical functions, iterative routines, and input/output processes. Results indicate that while `MATLAB` excels in matrix-based computations and `Julia` delivers robust performance across a wide range of tasks, `Python` offers competitive performance in several numerical operations while also serving as a highly versatile general-purpose language. Meanwhile, `R` can achieve competitive speed when linked to optimised libraries (e.g., via `Rcpp`).

In light of these results, we chose to implement our software in a `Python` environment. Moreover, Python extends differential calculus through the JAX package. JAX is a library designed for high-performance numerical computing. It combines the familiar NumPy API with automatic differentiation, just-in-time compilation (via XLA), and GPU/TPU acceleration. JAX is recommended for developing scientific packages because it allows you to write code that looks like standard NumPy but runs efficiently on modern hardware, while also being differentiable. This is particularly powerful in optimisation and machine learning tasks, where differential calculus is required, for example, to compute gradients and Hessian matrices. With JAX, you can automatically compute the Hessian of a scalar-valued function using its built-in differentiation tools (`grad`, `jacfwd`, `jacrev`), avoiding manual and numerical differentiations and ensuring efficiency. This choice ensures that our methodology can be readily adopted, extended, and integrated with modern computational tools.

The Python code (.py scripts) and instructions for reproducing the results in this Chapter are available at our GitHub repository: `https://github.com/jacopoRodeschini/Low_Rank_State_Space_Model`.

Chapter 7

Conclusions and future work

7.1 Main contributions of the theses

This thesis develops methodological contributions within the SSM framework suitable for environmental datasets. We propose several parameterisations for SSM by combining valid spatial statistical structures with simple physical assumptions. As noted in Shumway, Stoffer, and Stoffer (2000), the spatio-temporal context often requires additional restrictions on parameter matrices, and their algorithm includes modifications to accommodate such fully restricted settings. These assumptions lead to partial restrictions on the transition and/or covariance matrices, ensuring both interpretability and feasibility. In particular, in Chapter 5, we develop a multivariate SSM suitable for handling heteroskedastic spatio-temporal processes by allowing for a time-varying error variance. We apply the proposed method to the Agrimonia dataset, which can be regarded as a medium-sized dataset, for which exact Gaussian inference, and thus the traditional continuous-space SSM, remains feasible. However, as the spatial dimension increases, inference with continuous-space SSMs becomes infeasible. In Chapter 6, we develop a multivariate SSM, called a low-rank SSM, that leverages the GMRF representation of continuous-space GPs to handle large spatio-temporal datasets.

The main contributions of this thesis are the following:

1. The first contribution of this thesis is a procedure for combining heterogeneous data sources, such as satellite observations, model outputs, and in situ measurements, which have differing spatial and temporal resolutions, making them suitable for studying air quality variability. Specifically, Chapter 4 develops the Agrimonia dataset and details the harmonisation process applied to combine different sources at a daily temporal resolution and at the spatial scale of air

quality monitoring stations. The Agrimonia dataset, which integrates livestock, meteorological, and air quality data for the Lombardy region of Italy. Agrimonia enables advanced modelling of livestock emission impacts and evaluation of management strategies, representing a foundational contribution to environmental and agricultural science. Since pollutant concentrations are observed at monitoring sites, while explanatory variables are typically available on administrative units or gridded domains, we develop a gridded representation of the Agrimonia dataset, called the Agrimonia Gridded Dataset (AGC), which aligns air pollution and explanatory variables, thereby facilitating spatial modelling and analysis. In the Agrimonia dataset, numerous variables, including air quality pollutants, are available in both concentration and emission forms, which enables data fusion approaches. Combining these complementary representations can improve the consistency and accuracy of spatial analyses. In particular, integrating emission inventories with observed concentrations allows for a more robust characterisation of pollution dynamics and supports advanced modelling strategies. The data fusion perspective can be considered in future developments.

2. The second main contribution of this thesis is comparing different classes of models in explaining $PM_{2.5}$ daily variability. Indeed, statistical and machine learning models can provide valuable insights into air quality behaviour, including the identification of pollution sources, the factors influencing its variability, and the forecasting of future pollution levels under different scenarios, such as changes in emissions or weather patterns. We applied HDGM, GAMM and RFSTK models on the Agrimonia dataset, and we compared the performance in cross-validation (CV). All three models used are designed to handle spatio-temporal data, although each employs different methods to model external factors and spatio-temporal dependence. The results suggest that geostatistical approaches, such as the HDGM, yield the best predictive performance, that is, the validation RMSE is 6.31 $\mu g/m^3$, while being computationally efficient. However, more complex algorithms like RFSTK can enhance the identification of nonlinear and interaction effects. The comparison of models in the field of air quality highlighted that the spatio-temporal correlation is a crucial aspect that requires careful consideration. However, this correlation is also very sensitive. If not handled properly, it can lead to overfitting the model to the specific data used. This limitation is illustrated by the discrepancy in the in-sample performances versus the CV performances. In summary, this study underscores the efficacy of conventional techniques in modelling correlated spatio-temporal data, concurrently highlighting the complementary potential of machine learning and classical statistical approaches.

3. The third main contribution of this thesis is the development of a multivariate heteroskedastic spatio-temporal model. In particular, we propose to model heteroskedasticity in two main frameworks: *(i)*: time-varying unstructured flexible error variance, *(ii)* stochastic heteroskedastic functions characterised by a lower number of degrees of freedom. Building on the model comparison, we implement the two frameworks within a geostatistical approach by extending the HDGM to account for heteroskedasticity in the error terms. In both cases, this results in a high-dimensional parameter dimension efficiently handled by the expectation-maximisation (EM) algorithm. Using the Agrimonia dataset, we applied the proposed heteroskedastic HDGM to assess the $PM_{2.5}$ sensitivity to ammonia emissions through a "what-if" scenario analysis. Our findings suggest that considering a reduction of 26% in ammonia emission, the $PM_{2.5}$ overall reduction is close to 3.5%, while considering the 50% ammonia reduction, the $PM_{2.5}$ overall reduction is close to 7%.

   One key limitation of the current development is the assumption of stationarity, which fails to capture non-stationary behaviour often observed in real-world data. Moreover, handling large spatial datasets poses computational challenges. Indeed, the heteroskedasticity HDGM model has a complexity $O(m^3T)$, which becomes infeasible even for moderate m.

4. The fourth main contribution of this thesis is the development of a multivariate Low-Rank State-Space Model (LR-SSM) designed for heterotopic observations to tackle the "Big n problem". Our approach incorporates the SPDE–GMRF framework on a finite–element mesh to significantly reduce the rank of the spatial covariance matrix, from m to $R \ll m$. By embedding this sparse representation inside the Kalman filter, this reduces the computational cost of the Kalman filter to $O(mR^2T)$, while preserving the physical interpretation of the original model. The practitioner can therefore control the balance between computational cost and accuracy by choosing the value of R. We prove theoretical results, including asymptotic properties and error bounds for the proposed low-rank model. The maximum likelihood estimation (MLE) of the model parameters is carried out via EM, which results in closed-form updates for most parameters. Extensive simulations validate both the accuracy and efficiency of the method. The air-quality application showed the method's ability to handle large, multivariate heterotopic datasets with substantial missingness. In this application, we reduce computation time by about 93%, with only a 15% increase in the validation error. Although our discussion focuses on domains $\Omega \subset \mathbb{R}^2$, the approach extends naturally to higher-dimensional manifolds, and future studies may further adapt it to networks and graphs, highlighting the model's flexibility and broad applicability. A further refinement could allow the mesh to evolve in time, thereby allo-

cating resolution where the signal is most variable without inflating the overall computational budget. A comparison with the current implementation of the SSM for spatio-temporal data will be pursued in future research.

This dissertation demonstrates the flexibility of SSMs within the EM framework. It is reasonable to ask what the advantages and disadvantages of the EM approach presented here are compared to a fully Bayesian approach. This work does not perform a comparison between a spatio-temporal dynamic approach and a joint modelling approach, but focuses mainly on the SSM framework. Benchmark case studies comparing the performance of the various models proposed in the literature are also needed (see, for instance, Heaton et al. (2019) for comparison of spatial models). Our model comparison shows that such studies allow a close look at the advantages and disadvantages of each model, and would enable validation of critical decisions involved in the modelling process. It would also be useful to test the performance of the various algorithms against a single or a few known spatio-temporal fields. Such reference spatio-temporal fields should not be mere realisations of a Gaussian random field with a convenient separable space-time covariance; instead, they should correspond to an outcome of a complex transfer function based on a realistic input parameter field. For example, a pressure field derived from simulating flow in a highly structured, non-Gaussian field could be used as a reference, against which the predictive ability of different space–time stochastic models could be evaluated. It would also be interesting to compare in practice batch inference algorithms (using all available data) within the Kalman–EM machinery.

Appendices

Appendix A

Appendix for Chapter 4

A.1 Validation of AQ variables

The last column of Table 2 shows that the AQ network, intended as a multi-pollutant monitoring network, is unbalanced as each sensor is settled according to a pollutant-specific risk and exposure assessment, following EU regulations. As a result, the various stations often have different sensors. In addition, most sensors are subject to maintenance and other issues, resulting in quasi-random missing values. Unfortunately, NH_3 sensors are very few. NH_3 sensors are not required by EU regulations and are somewhat experimental, as explained above. This results in long non-operating periods. Providing a complete dataset with all (systematic and quasi-random) missing data imputed by some estimation method is an interesting objective. It may be pursued using three approaches:

- A suitable approach is based on multivariate statistical models applied to the unbalanced network and some additional auxiliary variables, see e.g. Fassò, Finazzi, and Ndongo (2016).
- Mathematical and numerical techniques may be used to simulate the physical and chemical processes that affect air pollutants as they disperse and react in the atmosphere. For example, see the Models-3/Community Multiscale Air Quality (CMAQ; http://www.epa.gov/asmdnerl/CMAQ) model.
- A mixture of numerical and statistical models may be used to leverage the previous two.

Indeed, the AgrImOnIA project is involved in this task for PM_{10} and $PM_{2.5}$, but this is out of the scope of building the current dataset.

The missing values imputation for the hourly and bi-hourly time series is performed using the State-Space Model (SSM) Durbin and Koopman (2012) and the relative Kalman smoother Harvey

(1990). For any hour $t \in \{1, \cdots, 24T\}$, where T is the number of days as before, let x_t be the scalar state describing the dynamics of the underlying AQ "true" concentrations and let y_t be the scalar observation of the observed hourly AQ series. Moreover, let u_t and ϵ_t be Gaussian white noises with unit variance representing the innovation and measurement error, respectively, with u_t and ϵ_t uncorrelated. The SSM here used for missing imputations is defined by:

$$\begin{cases} x_t = Ax_{t-1} + Bu_t \\ y_t = x_t + \epsilon_t \end{cases} \tag{A.1}$$

where the A and B parameters describe the dynamics and the additive error structure on the state x_t, respectively. Both A and B are estimated for each hourly time series using numerical optimisation of the likelihood function with initial values set to one. Assuming the hourly state errors $x_t - \hat{x_t}$ uncorrelated, we propagate the imputed uncertainty given by the smoother, through the mean of the generic day (d) as $\sigma_d = \sqrt{1/24^2 \sum_{t \in d} Var(x_t \mid y_1, ..., y_{24T})}$ where $Var(x_t \mid y_1, ..., y_{24T})$ is the variance of the smoothed states x_t for the hour $t \in \{1, \cdots, 24T\}$.

A validation experiment concerns NH_3 at the station called 'Bergamo Via Meucci', where ARPA Lombardy provides both hourly and daily time series. The daily time series is validated by ARPA but has a shorter coverage than the hourly time series. Indeed, the latter is not validated by ARPA as discussed in the Method Section and has several missing values; see Figure 44. The idea is to use the longer hourly time series to reconstruct the daily time series by our proposed method. So, we verify the performance of the missing imputation process by comparing the daily data obtained through the Kalman smoother and the daily data provided by the agency. Figure 44 shows the two daily time series: the blue line depicts our method, while the orange diamonds are the ARPA Lombardy daily data. From Figure 44, we see that the ARPA daily data and the daily mean of the hourly data may not be exactly the same in some cases, also for those days without missing values. This is due to the validation process performed by the ARPA on the daily data. Figure 45 shows the imputed uncertainty of daily average concentrations computed from imputed hourly time series. It can be observed that the time series obtained by our missing imputation method is very close to the daily one, with the Root Mean Square Error (RMSE) equal to 0.1710.

Another important experiment concerns the stations that sample $PM_{2.5}$ bi-hourly. In this case, if the sampling frequency is regular during the day, we do not expect bias problems in the associated daily time series, although we still have 12 missing values for each day spread every other hour. This situation occurs, for example, in the station called 'Parona Via della Miseria'. In fact, the data on $PM_{2.5}$ concentrations are available from May 2021, are validated and have a bi-hourly frequency.

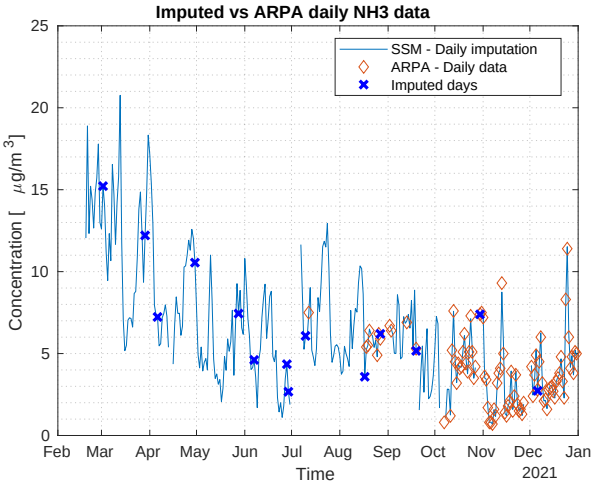


Figure 44: Hourly NH_3 data. Impact of Kalman smoother on daily data for the monitoring station named 'Bergamo Via Meucci'. Daily time series obtained with our method (blue line) with highlighted imputed days (blue crosses) and daily raw data (orange diamonds), RMSE = 0.1710.

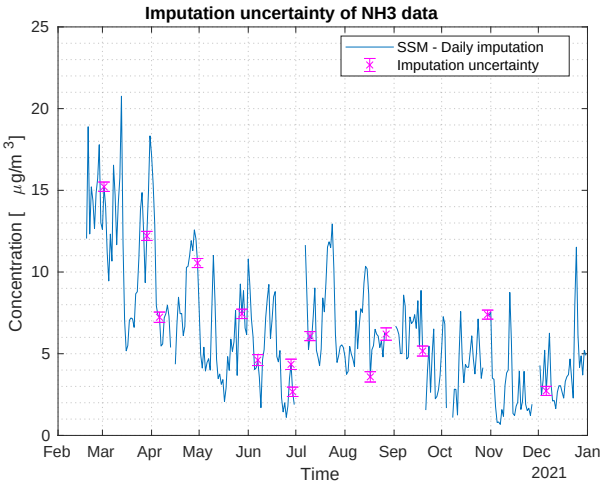


Figure 45: Hourly NH_3 data. Impact of Kalman smoother on daily data for the monitoring station named 'Bergamo Via Meucci'. Imputed daily data (blue line) with associated imputed uncertainty ($\pm 2\sigma_d$, magenta bar).

Figure 46 shows the daily mean computed with our approach (blue line) and the imputed days (blue crosses) compared to the daily mean (orange diamonds) computed without considering missing values. Instead, Figure 47 draws the imputed uncertainty of daily average concentrations computed from imputed hourly time series and shows that it is small.

We examined the dataset for the presence of anomalous values that are clearly outliers after the daily time series construction. Table 28 lists the three instances of anomalous values found. These extreme values have been replaced with the 'NaN' value. It is to be noted that this process is not to be considered a process of searching and removing outliers which is outside the context of this work.

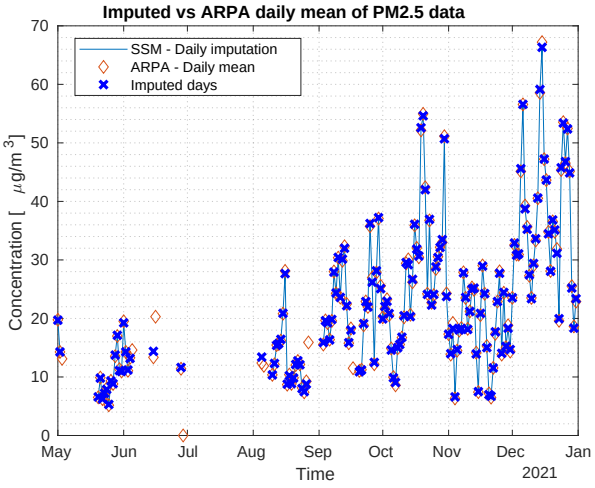


Figure 46: Bi-hourly $PM_{2.5}$ data. Impact of Kalman smoother on daily data for the monitoring station named 'Parona Via della Miseria'. Daily time series obtained with our method (blue line) with highlighted imputed days (blue crosses) and daily mean (orange diamonds) computed without considering missing values, RMSE = 0.3078.

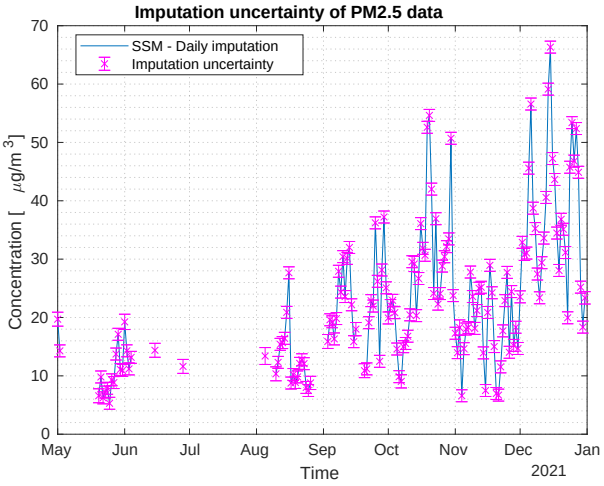


Figure 47: Bi-hourly $PM_{2.5}$ data. Impact of Kalman smoother on daily data for the monitoring station named 'Parona Via della Miseria'. Imputed daily data (blue line) with associated imputed uncertainty ($\pm 2\sigma_d$, magenta bar).

IDStations	Latitude	Longitude	Time	Pollutant	Value
STA.IT1582A	45.1361	8.4452	2021-04-27	AQ_pm_10	2399
STA.IT2121A	45.6894	8.4584	2021-04-04	AQ_pm_25	2794
STA.IT1751A	44.5740	8.9510	2016-10-10	AQ_so2	152.03

Table 28: Extremely large values for the AQ variables [$\mu g/m^3$] identified in the dataset. The values are replaced with the 'NaN' value.

A.1.1 Validation of WE variables

ARPA meteorological sensors effectively understand local meteorological conditions at the station site. This data has been used in various studies to adjust air quality estimates for meteorology (see e.g. Fassò, Maranzano, and Otto (2022).) The AgrImOnIA project aims at producing maps of the

livestock impact on AQ, that is, estimates of air quality and relations where no stations are available. For this reason, we used the ERA5 dataset, which provides average meteorological conditions over each pixel $0.25° \times 0.25°$. This means that ERA5 is not aimed at reproducing local conditions at the station. Nonetheless, the issue of which meteorological data provides a better predictor for air quality in Lombardy is an open issue and deserves attention for future research.

A.2 Validation of EM and LI variables

Livestock emission variables may be retrieved either from the emission inventory of the Lombardy region known as INEMARARPA Lombardia Settore Monitoraggi Ambientali, INEMAR (2022) or from the above-mentioned Copernicus product here named CAMS. We prefer CAMS as it fulfils the aims of the AgrImOnIA project mentioned in the Introduction section, namely generalisability and mappability. In particular, CAMS provides spatially resolved information monthly, whilst INEMAR provides municipally aggregated data yearly. The former is also readily available for the extended Lombardy, which overlaps with other Italian regions and Switzerland. Hence it is easier for inter-region and/or inter-country comparisons.

The original time resolution for EM and LI variables is lower than daily. In particular, EM variables are provided with a monthly temporal resolution, while LI variables are available every six months. Interpolation techniques are required for daily estimates.

We should consider that EM variables depend on manure management and spreading calendar, which are not continuos over time but concentrated on certain days for each farm. Similarly, the basic quantities for LI are livestock counts changing according to population dynamics driven by animal births, deaths, sales and purchases. Especially the latter two are often concentrated in certain time moments. Hence, the following interpolation splines must be interpreted as smooth approximations of an underlying process with irregular steps.

To avoid oscillations, overshoots, edge effects and negative values, we use PCHIP Fritsch and Carlson (1980) and Kahaner, Moler, and Nash (1989) interpolation. This method interpolates the data using a piecewise cubic polynomial while retaining the shape and monotonicity of the original data. Since the points to be interpolated are all positive (e.g. the number of bovines), negative values that could occur with classic splines are avoided. For example, Figure 48 shows the classic piecewise cubic spline, PCHIP and modified Akima piecewise cubic Hermite interpolation (Makima) Akima (1970), fitted on data from the ‘Corte de Cortesi’ station.

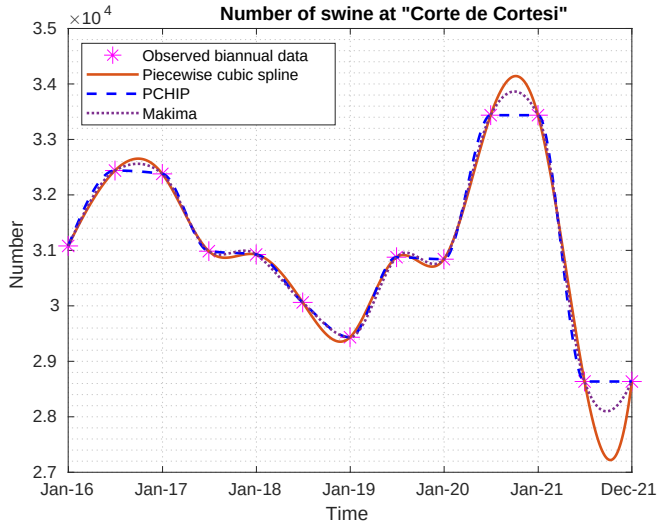


Figure 48: Piecewise cubic spline, PCHIP and Makima interpolation methods applied to swine time series for the monitoring station named 'Corte de Cortesi'.

A.2.1 Validation of LA variables

Data related to land cover and land and soil use are taken from ERA5-Land, CLC and SIARL datasets, respectively. The main considerations concern land use and soil use. Land use classifies the territory from the urbanisation and/or nature point of view (urban, industrial, road, agricultural, forest, marine and other). Since CLC data are available only for 2018 in the study period (2016-2021), we assume the land use to be constant in this work. We notice that CLC is defined on a fine grid, which is needed to accomplish the spatially resolved mapping objective of the AgrImOnIA project. Also, the former is better for extended Lombardy and other comparisons such as inter-region and/or inter-country comparisons.

Soil use classifies the land from the agricultural production point of view (cultivation type). It is expected to change over time as the cultivation type is often changed and/or rotated for greater yield. As an example, Figure 49 shows the soil use change in 2019 provided by the SIARL dataset for the station named 'Corte dei Cortesi'.

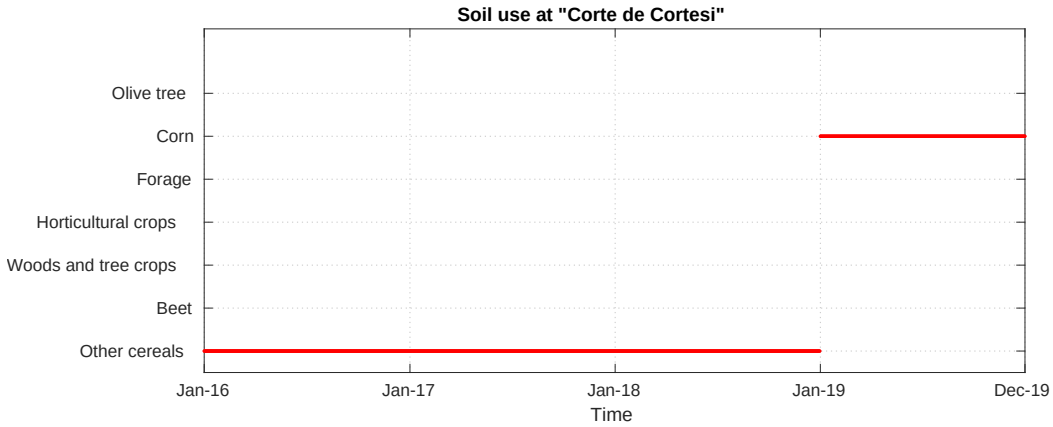


Figure 49: Piecewise constant function for soil use provided by SIARL dataset for the station named ‘Corte de Cortesi’. Note that the SIARL dataset covers data up to 2019 only.

Appendix B

Appendix for Chapter 6

Appendix

This appendix collects supporting material for the main text. Section B.1 defines the notation used throughout the appendix. Section B.2 provides the detailed proof of the identifiability theorem. Section B.3 discusses the weak convergence properties of the low-rank approximation to the latent process, with full theoretical justifications. Section B.5 presents the derivation and implementation details of the EM algorithm used for inference. Finally, Section B.6 explores the role of Neumann boundary conditions in the SPDE approximation and their impact on the induced covariance structure.

B.1 Notation

We introduce the following notation. For two matrices A and B of dimensions $r_A \times c$ and $r_B \times c$, respectively, we denote by $[A; B]$ the $(r_A + r_B) \times c$ matrix obtained by stacking A on top of B. We use $A \otimes B$ to denote the Kronecker product. Analogously, for two matrices A and B with dimensions $r \times c_A$ and $r \times c_B$, respectively, we define the column-wise juxtaposition of A and B, denoted by $[A, B]$, as the $r \times (c_A + c_B)$ matrix obtained by placing A to the left of B. Moreover, for matrices A and B of dimensions $r_A \times c_A$ and $r_B \times c_B$, respectively, we denote by $A \oplus B$ the block diagonal matrix of dimension $(r_A + r_B) \times (c_A + c_B)$, constructed with A and B placed as diagonal blocks and zeros elsewhere. Lastly, denote by $|A|$ the determinant of a square matrix A and by $\boldsymbol{I}_n$ the identity matrix of dimension n.

B.2 Proof of Theorem 1

Proof. *Proof* Assume two parameter sets

$$\Pi = (\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{f}, \boldsymbol{W}, \{k_i\}_{i=1}^q), \qquad \Pi^\star = (\boldsymbol{\beta}^\star, \boldsymbol{\sigma}^{2\star}, \boldsymbol{f}^\star, \boldsymbol{W}^\star, \{k_i^\star\}_{i=1}^q)$$

generate the same joint distribution for the observable field $\{\boldsymbol{y}(\boldsymbol{s}, t) : (\boldsymbol{s}, t) \in \mathcal{S} \times \mathbb{N}\}$ under the LR-SSM (6.9)–(6.10) for a fixed valueof R. As the latent state sequence $\{\boldsymbol{z}(\mathcal{V}_R, t)\}_{t\in\mathbb{N}}$ is strictly stationary, ergodic and has mean zero, the Strong Law of Large Numbers yields

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{z}(\mathcal{V}_R, t) \xrightarrow{\text{a.s.}} \mathbf{0} \quad (T \to \infty).$$

Define $L_{\boldsymbol{X}} = \lim_{T\to\infty} \frac{1}{T} \int \boldsymbol{X}(\boldsymbol{s}, t) d\boldsymbol{s}$ which exists finite by hypothesis. Then,

$$\boldsymbol{\beta} = (L_{\boldsymbol{X}} L_{\boldsymbol{X}}')^{-1} L_{\boldsymbol{X}}' \lim_{T\to\infty} \frac{1}{T} \int y(\boldsymbol{s}, t) d\boldsymbol{s} = \boldsymbol{\beta}^\star.$$

Centre the observations with $\boldsymbol{r}(\boldsymbol{s}, t) = \boldsymbol{y}(\boldsymbol{s}, t) - \boldsymbol{X}'(\boldsymbol{s}, t)\boldsymbol{\beta} = \boldsymbol{W}\boldsymbol{\Psi}_R(\boldsymbol{s})\boldsymbol{z}_t + \boldsymbol{\varepsilon}(\boldsymbol{s}, t)$, where $\boldsymbol{z}_t = \boldsymbol{z}(\mathcal{V}_R, t)$ obeys $\boldsymbol{z}_t = (\operatorname{diag}(\boldsymbol{f}) \otimes \boldsymbol{I}_R)\boldsymbol{z}_{t-1} + \boldsymbol{\eta}_t$.

Let $\boldsymbol{\Gamma}_1(\boldsymbol{s}, \boldsymbol{s}') = \operatorname{Cov}\big(\boldsymbol{r}(\boldsymbol{s}, t), \boldsymbol{r}(\boldsymbol{s}', t-1)\big)$. Using stationarity of $\boldsymbol{z}_t$ with covariance $\boldsymbol{V}_{\boldsymbol{z}} = \oplus_{i=1}^q (1 - f_i^2)^{-1} \boldsymbol{Q}_{\kappa_i}^{-1}$,

$$\boldsymbol{\Gamma}_1(\boldsymbol{s}, \boldsymbol{s}') = \boldsymbol{W}\boldsymbol{\Psi}_R(\boldsymbol{s})\big[\operatorname{diag}(\boldsymbol{f}) \otimes \boldsymbol{I}_R\big]\boldsymbol{V}_{\boldsymbol{z}}\boldsymbol{\Psi}_R(\boldsymbol{s}')'\boldsymbol{W}'.$$

Define

$$\boldsymbol{H} = \Big(\int_{\mathcal{S}^2} \boldsymbol{\Psi}_R'(\boldsymbol{s})\boldsymbol{W}^\dagger \boldsymbol{\Gamma}_0(\boldsymbol{s}, \boldsymbol{s}')\boldsymbol{W}^{\dagger\prime}\boldsymbol{\Psi}_R(\boldsymbol{s}')\, d\boldsymbol{s}\, d\boldsymbol{s}'\Big)^{-1} \int_{\mathcal{S}^2} \boldsymbol{\Psi}_R'(\boldsymbol{s})\boldsymbol{W}^\dagger \boldsymbol{\Gamma}_1(\boldsymbol{s}, \boldsymbol{s}')\boldsymbol{W}^{\dagger\prime}\boldsymbol{\Psi}_R(\boldsymbol{s}')\, d\boldsymbol{s}\, d\boldsymbol{s}',$$

with $\boldsymbol{W}^\dagger$ any right inverse of $\boldsymbol{W}$. The integrals average over a dense spatial grid, so $\boldsymbol{H} \to \operatorname{diag}(\boldsymbol{f})$. Equality of distributions implies the same limit for $\Pi^\star$, and the strict ordering $f_1 > \cdots > f_q$ (Assumption 2 (ii)) forces $\boldsymbol{f} = \boldsymbol{f}^\star$.

Stacking a dense set of locations gives

$$\operatorname{Cov}(\boldsymbol{r}_t) = (\boldsymbol{I}_m \otimes \boldsymbol{W})(\boldsymbol{P}\boldsymbol{V}_{\boldsymbol{z}}\boldsymbol{P}')(\boldsymbol{I}_m \otimes \boldsymbol{W})' + \boldsymbol{I}_m \otimes \boldsymbol{\Sigma},$$

where $\boldsymbol{P}$ stacks the basis evaluations. Because $\boldsymbol{P}$ and $\boldsymbol{W}$ have full column rank (Assumption 2 (iii)–(iv)), standard factor-analysis arguments (Bai and P. Wang, 2015), together with the column-scaling convention, yield $\boldsymbol{W} = \boldsymbol{W}^\star$ and $\boldsymbol{V}_{\boldsymbol{z}} = \boldsymbol{V}_{\boldsymbol{z}}^\star$. The SPDE construction gives a one-to-one map $k_i \mapsto \boldsymbol{Q}_{k_i}$, so $k_i = k_i^\star$ for all i.

From the diagonal of $\operatorname{Var}\big(\boldsymbol{r}(\boldsymbol{s}, t)\big)$, every element of the diagonal matrix $\boldsymbol{\Sigma}$ is determined; hence $\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^{2\star}$. All components coincide, so $\Pi = \Pi^\star$. □ □

B.3 Weak Convergence of the Low-Rank Approximation

In this section, we discuss the weak convergence of the process $\boldsymbol{\Psi}_R(\boldsymbol{s})\boldsymbol{z}(\mathcal{V}_R, t)$ to the continuous process $\boldsymbol{z}(\boldsymbol{s}, t)$. To this end, we first introduce the concept of Gaussian Markov Random Fields (GMRFs).

Let $G = (\mathcal{V}, \mathcal{E})$ be a finite undirected graph, where $\mathcal{V}$ is a set of vertices with cardinality $R < \infty$, and $\mathcal{E}$ is the set of edges. Let $\mathcal{X} = \{\mathcal{X}_v\}_{v \in \mathcal{V}}$ denote a collection of random variables indexed by $\mathcal{V}$. We say that $\mathcal{X}$ forms a zero-mean Markov Random Field defined by G if it satisfies the local Markov property 3.3. The explicit link between Gaussian Field (GF) parameters and GMRF parameters is established in Lindgren, Håvard Rue, and Lindström (2011). The goal is to approximate a GF with covariance matrix Σ using a GMRF with precision matrix Q such that Q^{-1} approximates Σ in a suitable norm. A GMRF representation of a Matérn GF can be efficiently constructed using the stochastic partial differential equation (SPDE) approach, in combination with the Finite Element Method (FEM) over a triangulated domain.

Let $\mathcal{D} \subset \mathbb{R}^d$ be a bounded domain with a smooth boundary $\partial\mathcal{D} \in C^2(\mathbb{R}^d)$. As shown in Guttorp and Gneiting (2006), a one-dimensional GF $x(\boldsymbol{s})$ with Matérn covariance characterised by parameters (σ^2, ν, κ) is a stationary solution of:

$$(\kappa^2 - \Delta)^{\alpha/2} x(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}, \quad \kappa > 0, \tag{B.1}$$

where $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ is the Laplacian and $\alpha = \nu + d/2$. The operator $(\kappa^2 - \Delta)^{\alpha/2}$ is defined via its spectral properties, as described in Whittle (1954) and Whittle (1963). Here, $\mathcal{W}$ denotes spatial Gaussian white noise with unit variance, and $x(\boldsymbol{s})$ has marginal variance:

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^{2\nu}}.$$

For simplicity, we consider the common spatial case with $d = 2$ and $\alpha = 2$, implying $\sigma^2 = (4\pi\kappa^2)^{-1}$. The spectral density of the stationary solution to (B.1) is given by:

$$R(\boldsymbol{k}) = (2\pi)^{-2}(\kappa^2 + \|\boldsymbol{k}\|^2)^{-2}.$$

Using the Fourier transform, the fractional Laplacian is defined as:

$$\mathcal{F}\{(\kappa^2 - \Delta)\varphi\}(\boldsymbol{k}) = (\kappa^2 + \|\boldsymbol{k}\|^2)\mathcal{F}\varphi(\boldsymbol{k}), \tag{B.2}$$

where φ is any function on $\mathbb{R}^d$ with an invertible right-hand side under the Fourier transform. To avoid null-space solutions, Neumann boundary conditions (zero normal derivatives on $\partial\mathcal{D}$) are imposed.

To numerically solve (B.1), we apply FEM (Ciarlet, 2002; Brenner, 2008; Quarteroni and Valli, 2008). Let $\mathcal{D}$ be partitioned into non-overlapping triangles via Delaunay triangulation, resulting in a mesh graph $G = (\mathcal{V}, \mathcal{E})$ with $R = |\mathcal{V}|$.

The stochastic weak formulation of the SPDE requires defining the inner product:

$$\langle f, g \rangle = \int_{\mathcal{D}} f(\boldsymbol{s}) g(\boldsymbol{s}) \, d\boldsymbol{s}.$$

The weak solution satisfies:

$$\langle \phi_j, (\kappa^2 - \Delta) x \rangle \overset{w}{=} \langle \phi_j, \mathcal{W} \rangle, \quad \forall j, \tag{B.3}$$

for suitable test functions $\{\phi_j\}_{j=1}^R$. We represent the FEM approximation as:

$$x^R(\boldsymbol{s}) = \sum_{k=1}^{R} \psi_k(\boldsymbol{s}) w_k = \boldsymbol{\psi}_R' \boldsymbol{w}, \tag{B.4}$$

where $\boldsymbol{\psi}_R(\boldsymbol{s}) = (\psi_1(\boldsymbol{s}), \ldots, \psi_R(\boldsymbol{s}))'$ are piecewise linear basis functions, and $\boldsymbol{w}$ are Gaussian weights. Boundary conditions are enforced by choosing basis functions with zero normal derivatives on $\partial\mathcal{D}$, ensuring solutions reside in the Hilbert space $\mathcal{H}_R(\mathcal{D})$.

Choosing $\phi_k = \psi_k$ yields the Galerkin approximation. Define matrices:

$$\begin{aligned} \boldsymbol{C}_{ij} &= \langle \psi_i, \psi_j \rangle, \\ \boldsymbol{G}_{ij} &= \langle \nabla\psi_i, \nabla\psi_j \rangle, \\ \boldsymbol{K}_{\kappa^2} &= \kappa^2 \boldsymbol{C} + \boldsymbol{G}. \end{aligned}$$

Under Neumann conditions, the precision matrix for $\boldsymbol{w}$ becomes:

$$\boldsymbol{Q}_{\kappa^2} = \boldsymbol{K}_{\kappa^2} \boldsymbol{C}^{-1} \boldsymbol{K}_{\kappa^2}. \tag{B.5}$$

While $\boldsymbol{C}^{-1}$ is dense, one can replace $\boldsymbol{C}$ with a diagonal matrix $\tilde{\boldsymbol{C}}$, where $\tilde{\boldsymbol{C}}_{ii} = \langle \psi_i, 1 \rangle$, to obtain a sparse $\boldsymbol{Q}$ (Lindgren, Håvard Rue, and Lindström, 2011).

As shown in equation (11) of Lindgren, Håvard Rue, and Lindström (2011), the approximation error is bounded as follows:

$$\sup_{f \in \mathcal{H}^1, \|f\|_{\mathcal{H}^1} \leq 1} \mathbb{E}\left[\langle f, x^R - x \rangle_{\mathcal{H}^1}^2 \right] \leq c h^2, \tag{B.6}$$

where $\mathcal{H}^1(\mathcal{D}, \kappa)$ is the Hilbert space:

$$\langle \phi, \psi \rangle_{\mathcal{H}^1(\mathcal{D},\kappa)} = \kappa^2 \langle \phi, \psi \rangle + \langle \nabla\phi, \nabla\psi \rangle,$$

and h denotes the mesh resolution.

We consider the following convergence definition,as formulated in Lindgren, Håvard Rue, and Lindström (2011). Let x^R be a sequence of $L^2(\mathcal{D})$-bounded Gaussian fields. We say $x^R \xrightarrow{D\{L^2(\mathcal{D})\}} x$ if for all $f, g \in L^2(\mathcal{D})$,

$$\mathbb{E}[\langle f, x^R\rangle] \to \mathbb{E}[\langle f, x\rangle], \quad \text{and} \quad \text{Cov}(\langle f, x^R\rangle, \langle g, x^R\rangle) \to \text{Cov}(\langle f, x\rangle, \langle g, x\rangle).$$

Theorem 5. *Let x be a weak solution to the SPDE $\mathcal{L}x = \mathcal{W}$ with $\mathcal{L} = (\kappa^2 - \Delta)$ and Neumann boundary conditions. Let x^R be its FEM approximation in $\mathcal{H}^1_R(\mathcal{D}, \kappa)$ using Gaussian white noise $\mathcal{W}$. Then,*

$$x^R \xrightarrow{D\{L^2(\mathcal{D})\}} x,$$
$$\mathcal{L}x^R \xrightarrow{D\{L^2(\mathcal{D})\}} \mathcal{L}x,$$

for $R \to \infty$ provided that $\bigcup_R \mathcal{H}^1_R(\mathcal{D}, \kappa)$ is dense in $\mathcal{H}^1(\mathcal{D}, \kappa)$ and x^R is the Galerkin solution.

The proof is available in Lindgren, Håvard Rue, and Lindström (2011).

B.3.1 Proof of Theorem 2

Proof. By Theorem 5, and fixing the component i of the process $\boldsymbol{\eta}(\cdot, t)$ model, we have $\eta_i(\cdot, t)^R \xrightarrow{D\{L^2(\mathcal{D})\}} \eta_i(\cdot, t)$ where $\eta_i^R(\cdot, t) = \boldsymbol{\psi}_R(s)'\boldsymbol{\eta}_i(\mathcal{V}_R, t)$ for some Gaussian weights $\boldsymbol{\eta}_i(\mathcal{V}_R, t)$, for each $t = 0, \ldots, T$. Due to independence, and the fact that $t = 1, \ldots, T$ varies over a finite set, we have that $\eta_i^R(\cdot, \cdot) \xrightarrow{D\{L^2(\mathcal{D}^T)\}} \eta_i(\cdot, \cdot)$. Now, due to SSM (6.9)-(6.10), we have that $z_i(\cdot, t) = \sum_{j=0}^t f_i^j \eta_i(\cdot, j)$. Thus, by the linearity of the approximation $\eta_i^R(\cdot, t)$, we have an immediate approximation of $z_i(\cdot, \cdot)$ as $z_i^R(\boldsymbol{s}, t) = \boldsymbol{\psi}_R(\boldsymbol{s})\boldsymbol{z}_i(\mathcal{V}_R, t)$ where the Gaussian weights are $\boldsymbol{z}_i(\mathcal{V}_R, t) = \sum_{j=0}^T f_i^{t-j}\boldsymbol{\eta}_i(\mathcal{V}_R, t)$. Therefore, $\eta_i^R(\cdot, \cdot) \xrightarrow{D\{L^2(\mathcal{D}^T)\}} \eta_i(\cdot, \cdot)$ implies $z_i^R(\cdot, \cdot) \xrightarrow{D\{L^2(\mathcal{D}^T)\}} z_i(\cdot, \cdot)$. As the number of components is finite $i = 1, \ldots, q$, we obtain the aimed result:

$$\boldsymbol{z}^R(\boldsymbol{s}, t) = \boldsymbol{\Psi}_R(\boldsymbol{s})\boldsymbol{z}(\mathcal{V}_R, t) \xrightarrow{D\{L^2(\mathcal{D}^{qT})\}} \boldsymbol{z}(\boldsymbol{s}, t).$$

This convergence result also implies that

$$\mathbb{E}[\boldsymbol{\Psi}_R(\boldsymbol{s})\boldsymbol{z}(\mathcal{V}_R, t)] \to \mathbb{E}[\boldsymbol{z}(\boldsymbol{s}, t)],$$
$$\text{Cov}(\boldsymbol{\Psi}_R(\boldsymbol{s})\boldsymbol{z}(\mathcal{V}_R, t), \boldsymbol{\Psi}_R(\tilde{\boldsymbol{s}})\boldsymbol{z}(\mathcal{V}_R, \tilde{t})) \to \text{Cov}(\boldsymbol{z}(\boldsymbol{s}, t), \boldsymbol{z}(\tilde{\boldsymbol{s}}, \tilde{t})).$$

□ □

B.3.2 Proof of Theorem 3

Proof. The two observation equations (2.23) and (6.9) differ only in the latent state, therefore

$$\begin{aligned}\boldsymbol{y}(\boldsymbol{s},t) - \boldsymbol{y}^R(\boldsymbol{s},t) &= \boldsymbol{X}_t'\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{z}(\boldsymbol{s},t) + \boldsymbol{\varepsilon}(\boldsymbol{s},t) - \boldsymbol{X}_t'\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{z}^R(\boldsymbol{s},t) + \boldsymbol{\varepsilon}(\boldsymbol{s},t) \\ &= \boldsymbol{W}(\boldsymbol{z}(\boldsymbol{s},t) - \boldsymbol{z}^R(\boldsymbol{s},t)),\end{aligned}$$

so the mean–square bias $\mathbb{E}\big[\|\boldsymbol{y}_t - \boldsymbol{y}_t^R\|^2\big]$ is controlled by the mean–square error of the state approximation:

$$\mathbb{E}\big[\|\boldsymbol{y}_t - \boldsymbol{y}_t^R\|_2^2\big] \le C \cdot \mathbb{E}\big[\|\boldsymbol{z}_t - \boldsymbol{z}_t^R\|_2^2\big], \tag{B.7}$$

where C is a constant that depends on $\|W\|$. The inequality (B.6) implies

$$\mathbb{E}\big[\|\eta_i^R - \eta_i\|_{\mathcal{H}^{-1}}^2\big] \;\le\; c\,h^2,$$

for each component $i = 1, \dots, q$ of $\boldsymbol{\eta}$ where $\|\cdot\|_{\mathcal{H}^{-1}}$ is the dual norm of $\mathcal{H}^1$. On the bounded domain $\mathcal{D}$, the inclusion $L^2(\mathcal{D}) \subset \mathcal{H}^{-1}(\mathcal{D})$ leads to

$$\|g\|_{L^2} \;\le\; C_{\text{emb}}\,\|g\|_{H^{-1}} \quad \text{(for every } g \in H^{-1} \text{ with zero mean)},$$

that applied to $g = \boldsymbol{\eta}_t^R - \boldsymbol{\eta}_t$ and taking expectation leads to

$$\mathbb{E}\big[\|\boldsymbol{\eta}^R - \boldsymbol{\eta}\|_{L^2}^2\big] \;\le\; C_{\text{emb}}^2 \sum_{i=1}^{q} \mathbb{E}\big[\|\eta_i^R - \eta_i\|_{H^{-1}}^2\big] \;\le\; q\,C_{\text{emb}}^2\,c\,h^2.$$

As we are interested in a general time $t = 1, \dots, T$, the AR(1) equations (2.24) and (6.10) yield

$$\begin{aligned}\|\boldsymbol{z}_t - \boldsymbol{z}_t^R\|_2 &\le \varphi\|\boldsymbol{z}_{t-1} - \boldsymbol{z}_{t-1}^R\|_2 + \|\boldsymbol{\eta}_t - \boldsymbol{\eta}_t^R\|_2, \\ &\le \varphi\|\boldsymbol{z}_{t-1} - \boldsymbol{z}_{t-1}^R\|_2 + q\,C_{\text{emb}}^2\,c\,h^2\end{aligned}$$

with C being a constant $\varphi = \max_i |f_i| < 1$. Iterating backwards in time, we obtain

$$\mathbb{E}\big[\|\boldsymbol{z}_t^R - \boldsymbol{z}_t\|_{L^2}^2\big] \le \sum_{r=0}^{t} \varphi^{2r}\,q\,C_{\text{emb}}^2\,c\,h^2 = \frac{1-\varphi^{2(t+1)}}{1-\varphi^2}\,q\,C_{\text{emb}}^2\,c\,h^2 \tag{B.8}$$

Equations (B.7) and (B.8) yield the result. □ □

B.4 Proof of Theorem 4

Let $\boldsymbol{y}(\boldsymbol{s},t)$ distributed as specified in SSM (2.23)–(2.24) with set of parameter Π. We are interested in studying the bias on the estimation of parameters when, for a fixed R we use the LR-SSM in the estimation. Let $\boldsymbol{\pi}$ the vector $(\boldsymbol{\beta}', \boldsymbol{\sigma}', \boldsymbol{f}', k_1, \dots, k_q)'$. Denote by

$$\mathcal{P}_{\boldsymbol{\pi}} \equiv \mathcal{N}\big(\boldsymbol{X}'(\boldsymbol{s},t)\boldsymbol{\beta}, \boldsymbol{\Sigma}_y(\boldsymbol{\pi})\big), \qquad \boldsymbol{\Sigma}_y(\boldsymbol{\pi}) = \boldsymbol{W}\boldsymbol{\Sigma}_z(\boldsymbol{\pi})\boldsymbol{W}' + \boldsymbol{\Sigma},$$

the law of $\boldsymbol{y}(\boldsymbol{s},t)$, with $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\sigma}^2)$ and latent covariance $\boldsymbol{\Sigma}_z(\boldsymbol{\pi})$. Replacing $\boldsymbol{\eta}(\boldsymbol{s},t)$ by its FEM-approximation $\boldsymbol{\Psi}_R(\boldsymbol{s})\boldsymbol{\eta}(\mathcal{V}_R,t)$ yields a low-rank covariance

$$\boldsymbol{\Sigma}_y^{(R)}(\boldsymbol{\pi}) = \boldsymbol{W}\boldsymbol{\Sigma}_{z,R}(\boldsymbol{\pi})\boldsymbol{W}' + \boldsymbol{\Sigma}.$$

Define the covariance perturbation

$$\boldsymbol{\Delta}_R(\boldsymbol{\pi}) := \boldsymbol{\Sigma}_z(\boldsymbol{\pi}) - \boldsymbol{\Sigma}_{z,R}(\boldsymbol{\pi}).$$

By the FEM approximation, see equation (11) of Lindgren, Håvard Rue, and Lindström (2011), the Galerkin error is bounded as

$$\left\|\Delta_R(\boldsymbol{\pi})\right\| \leq C\, h, \tag{B.9}$$

were h is the diameter of the largest circle that can be inscribed in a triangle in the triangulation, and C is a constant. Because the LR-SSM is misspecified, the likelihood based on (6.9)–(6.10) concentrates around the minimiser of the Kullback–Leibler divergence

$$\boldsymbol{\pi}_R^{\star} := \arg\min_{\boldsymbol{\vartheta}\in\Theta_R} \operatorname{KL}\big(\mathcal{P}_{\boldsymbol{\pi}} \,\|\, \mathcal{P}_{\boldsymbol{\vartheta}}^{(R)}\big) \tag{B.10}$$

where $\mathcal{P}_{\boldsymbol{\vartheta}}^{(R)} \equiv \mathcal{N}(\boldsymbol{X}(\boldsymbol{s},t)'\boldsymbol{\beta}, \boldsymbol{\Sigma}_y^{(R)}(\boldsymbol{\vartheta}))$. For Gaussian measures with the same mean function, the Kullback–Leibler divergence takes the form

$$\operatorname{KL}(\boldsymbol{\pi},\boldsymbol{\vartheta}) = \tfrac{1}{2}\left\{\operatorname{tr}\big[\boldsymbol{\Sigma}_y^{(R)}(\boldsymbol{\vartheta})^{-1}\boldsymbol{\Sigma}_y(\boldsymbol{\pi})\big] - \log\det\big[\boldsymbol{\Sigma}_y^{(R)}(\boldsymbol{\vartheta})^{-1}\boldsymbol{\Sigma}_y(\boldsymbol{\pi})\big] - p\right\}. \tag{B.11}$$

Setting the gradient to zero yields, after vectorising (vec) the matrices,

$$0 = \boldsymbol{J}'\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\pi}) \otimes \boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\pi})\ \operatorname{vec}\boldsymbol{\Delta}_R(\boldsymbol{\pi}) + \ \boldsymbol{H}\,(\boldsymbol{\pi}_R^{\star} - \boldsymbol{\pi}) + o(\|\Delta_R(\boldsymbol{\pi})\|)\,,$$

where $\boldsymbol{J} = \nabla_{\boldsymbol{\vartheta}} \operatorname{vec}\boldsymbol{\Sigma}_y^{(R)}(\boldsymbol{\pi})\big|_{\boldsymbol{\vartheta}=\boldsymbol{\pi}}$ and $\boldsymbol{H}$ is the Fisher-information matrix of the correct model, which is positive definite. Solving the equation, we obtain the exact first-order bias

$$\boldsymbol{\pi}_R^{\star} - \boldsymbol{\pi} = -\boldsymbol{H}^{-1}\boldsymbol{J}'\big(\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\pi}) \otimes \boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\pi})\big)\operatorname{vec}\boldsymbol{\Delta}_R(\boldsymbol{\pi})\ +\ o\big(\|\boldsymbol{\Delta}_R(\boldsymbol{\pi})\|\big) \tag{B.12}$$

Combining (B.12) with the covariance bound (B.9) delivers the rate

$$\left\|\boldsymbol{\pi}_R^{\star} - \boldsymbol{\pi}\right\| = O(h).$$

Note that, the term (B.11) does not involve the mean of the processes, which implies the parameter $\boldsymbol{\beta}$ has no bias. □

B.5 EM Algorithm

We are interested in estimating the set parameter Π based on the observed data. We recall here the notation. Denote by $\boldsymbol{y}_t$ the $(m_t = \sum_{j=1}^{p} m_{j,t})$-dimensional vector

$$\boldsymbol{y}_t = \left(\boldsymbol{y}_1(\mathcal{S}_{1,t}, t)', \boldsymbol{y}_2(\mathcal{S}_{2,t}, t)', \dots, \boldsymbol{y}_p(\mathcal{S}_{p,t}, t)'\right)',$$

where $\mathcal{S}_{j,t} = \{\boldsymbol{s}_{i,t}\}_{i=1}^{m_{j,t}}$ is the set of locations where the process $y_j(\cdot, t)$ has been observed, and

$$\boldsymbol{y}_{j,t} = \boldsymbol{y}_j(\mathcal{S}_{j,t}, t) = \Big(y_j(\boldsymbol{s}_{1,t}, t), \dots, y_j(\boldsymbol{s}_{m_{j,t},t}, t)\Big)'.$$

Let $\boldsymbol{y}_{1:T}$ denote the vector $(\boldsymbol{y}_1', \dots, \boldsymbol{y}_T')'$.

An analogous notation is used for $\boldsymbol{z}_{1:T} = (\boldsymbol{z}_1', \dots, \boldsymbol{z}_T')'$, where

$$\boldsymbol{z}_t = \left(\boldsymbol{z}_1(\mathcal{V}_R, t)', \dots, \boldsymbol{z}_q(\mathcal{V}_R, t)'\right)', \quad \text{with} \quad \boldsymbol{z}_{j,t} = \boldsymbol{z}_j(\mathcal{V}_R, t) = \left(z_j(\boldsymbol{r}_1, t), \dots, z_j(\boldsymbol{r}_R, t)\right)'.$$

Moreover, let $\boldsymbol{X}_{1:T} = [\boldsymbol{X}_1; \dots; \boldsymbol{X}_T]$, where each $\boldsymbol{X}_t$ is a $b \times m_t$ matrix, with $b = \sum_{j=1}^{p} b_j$, defined as

$$\boldsymbol{X}_t = \boldsymbol{X}_{1,t} \oplus \boldsymbol{X}_{2,t} \oplus \cdots \oplus \boldsymbol{X}_{p,t}.$$

Each component matrix $\boldsymbol{X}_{j,t}$ is constructed as

$$\boldsymbol{X}_{j,t} = [\boldsymbol{X}_j(\boldsymbol{s}_{1,t}, t), \dots, \boldsymbol{X}_j(\boldsymbol{s}_{m_{j,t},t}, t)],$$

with $\boldsymbol{X}_{j,t}$ being a $b_j \times m_{j,t}$ matrix, and each column vector $\boldsymbol{X}_j(\boldsymbol{s}, t) \in \mathbb{R}^{b_j}$ representing the covariate values associated with the j-th component of $\boldsymbol{y}(\boldsymbol{s}, t)$ at location $\boldsymbol{s}$ and time t.

We assume that the mesh $\mathcal{G}_R = (\mathcal{V}_R, \mathcal{E}_R)$, on which the latent process in the state-space model (6.9)–(6.10) is defined, is given. The procedure for selecting this mesh is discussed in Section 6.5.1.

Let $L(\Pi; \boldsymbol{y}_{1:T}, \boldsymbol{X}_{1:T}, \boldsymbol{z}_{1:T}) = p(\boldsymbol{y}_{1:T}, \boldsymbol{z}_{1:T} | \boldsymbol{X}_{1:T}; \Pi)$ be the full likelihood function. The MLE of Π is obtained by maximising the marginal likelihood of the observed data likelihood

$$p(\boldsymbol{y}_{1:T} | \boldsymbol{X}_{1:T}; \Pi) = \int p(\boldsymbol{y}_{1:T} | \boldsymbol{X}_{1:T}, \boldsymbol{z}_{1:T}, \Pi) p(\boldsymbol{z}_{1:T} | \Pi) d\boldsymbol{z}_{1:T}. \tag{B.13}$$

However, this quantity is intractable since $\boldsymbol{z}_{1:T}$ is unobserved and its distribution is unknown before attaining Π. The EM algorithm seeks to find the ML estimate of the marginal likelihood by iteratively applying these two steps:

E-step Define $Q(\Pi \mid \Pi^{(k)})$ as the expected value of the log-likelihood function of Π, with respect to the current conditional distribution of $\boldsymbol{z}_{1:T}$ given $(\boldsymbol{y}_{1:T}, \boldsymbol{X}_{1:T})$ and the current estimates of the parameters $\Pi^{(k)}$:

$$Q(\Pi \mid \Pi^{(k)}) = \mathbb{E}_{\boldsymbol{z}_{1:T} \sim p(\cdot | \boldsymbol{y}_{1:T}, \boldsymbol{X}_{1:T}; \Pi^{(k)})} [\log p(\boldsymbol{y}_{1:T} \mid \boldsymbol{X}_{1:T}, \boldsymbol{z}_{1:T}; \Pi)].$$

M-step Obtain the next value of Π, namely $\Pi^{(k+1)}$ as the maximiser of $Q(\Pi|\Pi^{(k)})$:

$$\Pi^{(k+1)} = \arg\max_{\Pi\in\mathcal{P}} Q(\Pi \mid \Pi^{(k)}), \tag{B.14}$$

where the maximisation is done on a suitable compact set $\mathcal{P}$ which contains Π as interior point.

The computation of the conditional distribution of the latent states $\boldsymbol{z}_{1:T}$ given the observed data and the current value of the parameter $\Pi^{(i)}$ can be efficiently obtained using the Kalman filter and smoother, and the formulas are given in the Appendix B.5 together with the formulas to obtain arg max in (B.14). Several stopping criteria can be considered, including a maximum number of iterations, a tolerance on the norm of the difference $\Pi^{(k)} - \Pi^{(k+1)}$, or a tolerance on the percentage change in the likelihood. Suppose that the algorithm stops after K iterations, then the value $\hat{\Pi} = \Pi^{(K)}$ corresponds to the proposed estimator of Π.

Close formulas We now obtain close formulas of the parameters $\{\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{f}, \boldsymbol{w}\}$ and express the objective function that we need to numerically minimise to obtain an estimate of the value of κ_i for $i = 1, \ldots, q$. First, we write explicitly the full log-likelihood $\ell(\Pi; \boldsymbol{y}_{1:T}, \boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}) = \log p(\boldsymbol{y}_{1:T}, \boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T}; \Pi)$. Denote by w_{ij} the element at row i and column j of the matrix $\boldsymbol{W}$ in (6.9) and define the matrix $\boldsymbol{\Psi}_t^{\boldsymbol{w}}$ as

$$\boldsymbol{\Psi}_t^{\boldsymbol{w}} = \begin{bmatrix} w_{11}\boldsymbol{\Psi}_{11,t} & \ldots & w_{1q}\boldsymbol{\Psi}_{1q,t} \\ \vdots & \ddots & \vdots \\ w_{p1}\boldsymbol{\Psi}_{p1,t} & \ldots & w_{pq}\boldsymbol{\Psi}_{pq,t} \end{bmatrix}, \tag{B.15}$$

where $\boldsymbol{\Psi}_{ij,t}$ is a known basis matrix that characterises the effect of the vector $\boldsymbol{z}_j(\mathcal{V}_j, t)$ on $\boldsymbol{y}_i(\mathcal{S}_{i,t}, t)$, that is $\boldsymbol{\Psi}_{ij,t} = [\psi_R(\boldsymbol{s}_{1,t})'; \ldots; \psi_R(\boldsymbol{s}_{m_{i,t},t})']$. Using the notation introduced in Section 6.5, we can write the SSM (6.9)-(6.10) in matrix form as

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{X}_t'\boldsymbol{\beta} + \boldsymbol{\Psi}_t^{\boldsymbol{w}}\boldsymbol{z}_t + \boldsymbol{\varepsilon}_t \\ \boldsymbol{z}_t &= \boldsymbol{F}\boldsymbol{z}_{t-1} + \boldsymbol{\eta}_t, \end{aligned}$$

where $\boldsymbol{\varepsilon}_t = (\varepsilon_1(\mathcal{S}_{1,t}, t)', \ldots, \varepsilon_p(\mathcal{S}_{p,t}, t))'$ with $\varepsilon_i(\mathcal{S}_{i,t}, t) = (\varepsilon_i(\boldsymbol{s}_{1,t}, t), \ldots, \varepsilon_i(\boldsymbol{s}_{m_{i,t},t}, t))'$. Thus, the full log-likelihood $\ell(\Pi; \boldsymbol{y}_{1:T}, \boldsymbol{X}_{1:T}, \boldsymbol{z}_{1:T})$ satisfies

$$\begin{aligned} -2\ell(\Pi; \boldsymbol{y}_{1:T}, \boldsymbol{X}_{1:T}, \boldsymbol{z}_{1:T}) &= \log|\boldsymbol{\Sigma}_0| + (\boldsymbol{z}_0 - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{z}_0 - \boldsymbol{\mu}_0) \\ &+ \sum_{t=1}^{T}\left(\log|\boldsymbol{\Sigma}_t| + \boldsymbol{e}_t'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{e}_t\right) - T\log|\boldsymbol{Q}_{\boldsymbol{\kappa}}| + \sum_{t=1}^{T}(\boldsymbol{z}_t - \boldsymbol{F}\boldsymbol{z}_{t-1})'\,\boldsymbol{Q}_{\boldsymbol{\kappa}}\,(\boldsymbol{z}_t - \boldsymbol{F}\boldsymbol{z}_{t-1}), \end{aligned}$$

where $\boldsymbol{e}_t = \left(\boldsymbol{y}_t - \boldsymbol{X}_t'\boldsymbol{\beta} - \boldsymbol{\Psi}_t^{\boldsymbol{w}}\boldsymbol{z}_t\right)$, and $\boldsymbol{\Sigma}_t = (\sigma_1 \boldsymbol{I}_{m_1,t}) \oplus \cdots \oplus (\sigma_p \boldsymbol{I}_{m_p,t})$ is the covariance matrix of $\boldsymbol{\varepsilon}_t$, and $\boldsymbol{Q}_{\boldsymbol{\kappa}} = \boldsymbol{Q}_{\kappa_1} \oplus \cdots \oplus \boldsymbol{Q}_{\kappa_q}$ is the precision matrix of the innovation term $\boldsymbol{z}_t - \boldsymbol{F}\boldsymbol{z}_{t-1}$, the matrix $\boldsymbol{F} = (f_1 \boldsymbol{I}_{R_1}) \oplus \cdots \oplus (f_q \boldsymbol{I}_{R_q})$ corresponds to the transition matrix and $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are the vector mean and covariance matrix of $\boldsymbol{z}_0$, respectively. Thus, the conditional expectation of the complete-data log-likelihood given the current estimate of Π at pass (k) of the EM algorithm takes the form

$$Q(\Pi|\Pi^{(k)}) = \mathbb{E}_{\Pi^{(k)}}\left[-2\ell(\Pi; \boldsymbol{y}_{1:T}, \boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T})|\boldsymbol{y}_{1:T}, \boldsymbol{X}_{1:T}\right] \tag{B.16}$$

where $\mathbb{E}_{\Pi^{(k)}}$ stands for the expected value considering the posterior distribution of the latent variables $\boldsymbol{z}_{1:T}$ given the observed data $(\boldsymbol{y}_{1:T}, \boldsymbol{X}_{1:T})$ and the current parameter estimate $\Pi^{(k)}$.

Applying the conditional expectation, it follows that

$$\begin{aligned}
Q(\Pi|\Pi^{(k)}) &= \log|\boldsymbol{\Sigma}_0| + \operatorname{Tr}\left[\boldsymbol{\Sigma}_0^{-1}\left((\boldsymbol{z}_0^T - \boldsymbol{\mu}_0)(\boldsymbol{z}_0^T - \boldsymbol{\mu}_0)' + \boldsymbol{P}_0^T\right)\right] \\
&+ \sum_{t=1}^{T}\log|\boldsymbol{\Sigma}_t| + \sum_{t=1}^{T}\operatorname{Tr}\left[\boldsymbol{\Sigma}_t^{-1}\left(\left(\boldsymbol{y}_t - \boldsymbol{X}_t'\boldsymbol{\beta} - \boldsymbol{\Psi}_t^{\boldsymbol{w}}\boldsymbol{z}_t^T\right)\left(\boldsymbol{y}_t - \boldsymbol{X}_t'\boldsymbol{\beta} - \boldsymbol{\Psi}_t^{\boldsymbol{w}}\boldsymbol{z}_t^T\right)' + \boldsymbol{\Psi}_t^{\boldsymbol{w}}\boldsymbol{P}_t^T\boldsymbol{\Psi}^{\boldsymbol{w}}\right)\right] \\
&- T\log|\boldsymbol{Q}_{\boldsymbol{\kappa}}| + \operatorname{Tr}\left[\boldsymbol{Q}_k\sum_{t=1}^{T}(\boldsymbol{z}_t^T - \boldsymbol{F}\boldsymbol{z}_{t-1}^T)(\boldsymbol{z}_t^T - \boldsymbol{F}\boldsymbol{z}_{t-1}^T)'\right] \\
&+ \operatorname{Tr}\left[\boldsymbol{Q}_{\boldsymbol{\kappa}}\sum_{t=1}^{T}\left(\boldsymbol{P}_t^T + \boldsymbol{F}\boldsymbol{P}_{t-1}^T\boldsymbol{F}' - \boldsymbol{F}\boldsymbol{P}_{t,t-1}^T - (\boldsymbol{P}_{t,t-1}^T)'\boldsymbol{F}'\right)\right]
\end{aligned} \tag{B.17}$$

where the quantities $\boldsymbol{z}_t^t$ and $\boldsymbol{P}_t^t$ for $t = 0, \ldots, T$ are obtained sequentially using the Kalman Filter equations, whose forward pass for $t =, 1, \ldots, T$ consists of:

$$\begin{aligned}
\textbf{Prediction:}\quad \boldsymbol{z}_t^{t-1} &= \boldsymbol{F}^{(k)}\boldsymbol{z}_{t-1}^{t-1}, \\
\boldsymbol{P}_t^{t-1} &= \boldsymbol{F}^{(k)}\boldsymbol{P}_{t-1}^{t-1}\boldsymbol{F}^{(k)'} + \boldsymbol{Q}_{\boldsymbol{\kappa}^{(k)}}^{-1}, \\
\text{with}\quad \boldsymbol{z}_0^0 &= \boldsymbol{\mu}_0^{(k)}, \quad \boldsymbol{P}_0^0 = \boldsymbol{\Sigma}_0^{(k)}. \\
\textbf{Update (Filter):}\quad \boldsymbol{\epsilon}_t &= \boldsymbol{y}_t - \boldsymbol{X}_t'\boldsymbol{\beta}^{(k)} - \boldsymbol{\Psi}_t^{\boldsymbol{w}^{(k)}}\boldsymbol{z}_t^{t-1}, \\
\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_t} &= \boldsymbol{\Psi}_t^{\boldsymbol{w}^{(k)}}\boldsymbol{P}_t^{t-1}(\boldsymbol{\Psi}_t^{\boldsymbol{w}})' + \boldsymbol{\Sigma}_t, \\
\boldsymbol{K}_t &= \boldsymbol{P}_t^{t-1}(\boldsymbol{\Psi}_t^{\boldsymbol{w}^{(k)}})'(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_t})^{-1}, \\
\boldsymbol{z}_t^t &= \boldsymbol{z}_t^{t-1} + \boldsymbol{K}_t\boldsymbol{\epsilon}_t, \\
\boldsymbol{P}_t^t &= \boldsymbol{P}_t^{t-1} - \boldsymbol{K}_t\boldsymbol{\Psi}_t^{\boldsymbol{w}^{(k)}}\boldsymbol{P}_t^{t-1}.
\end{aligned}$$

and where the quantities z_t^T and P_t^T for $t = 0, \ldots, T$ are obtained sequentially using the Kalman Smoothing equations, whose backwards pass for $t = T, \ldots, 1$ consists of:

$$\begin{aligned}
\textbf{Smoothing:} \quad z_T^T &= z_T^T, \quad P_T^T = P_T^T, \quad \text{(for } t = T \text{ only)} \\
J_{t-1} &= P_{t-1}^{t-1} F^{(k)'} (P_t^{t-1})^{-1}, \\
z_{t-1}^T &= z_{t-1}^{t-1} + J_{t-1}(z_t^T - z_t^{t-1}), \\
P_{t-1}^T &= P_{t-1}^{t-1} + J_{t-1}(P_t^T - P_t^{t-1}) J_t'.
\end{aligned}$$

The Smoothed lag-one covariance $P_{t,t-1}^T$ for $t = 1, \ldots, T$ is also obtain sequentially for $t = T, \ldots, 2$ by

$$\begin{aligned}
\textbf{Smoothed lag-one covariance:} \quad P_{T,T-1}^T &= (I - K_T \Psi_{w^{(k)}}^T) F P_{T-1}^{T-1}, \quad \text{(for } t = T \text{ only)} \\
P_{t,t-1}^T &= P_{t,t-1} J_{t-2}' + J_{t-2}(P_{t-1,t-1}^T - F P_{t-1}^{t-1}) J_{t-2}'.
\end{aligned}$$

Now, we differentiate $Q(\Pi|\Pi^{(k)})$ from equation (B.17) concerning β, σ, w, f to derive closed-form expressions for the values that minimise the function. Meanwhile, the objective function, which must be minimised numerically, provides an estimate for κ_i for $i = 1, \ldots, q$.

Denote by e_i the $q-$dimensional vector with all entries equal to zero except for the $i-$th that is equal to 1. Moreover, for $i = 1, \ldots, p$, define $\Psi_{i,t} = \Psi_{i1,t} \oplus \cdots \oplus \Psi_{iq,t}$ that is a $(qm_{i,t}) \times (qR)$ matrix.

We obtain:

$$\beta = \left(\sum_{t=1}^{T} X_t' (\Sigma_t^{(k)})^{-1} X_t \right)^{-1} \left(\sum_{t=1}^{T} X_t' (\Sigma_t^{(k)})^{-1} \left(y_t - \Psi_t^{w^{(k)}} z_t^T \right) \right) \tag{B.18}$$

For $i = 1, \ldots, p$, update $(\sigma_i^2)^{(k+1)}$ as:

$$\begin{aligned}
(\sigma_i^2)^{(k+1)} = \left(\sum_{t=1}^{T} m_{i,t} \right)^{-1} \sum_{t=1}^{T} \mathrm{Tr} \Bigg\{ & \left(y_{i,t} - X_{i,t}' \beta^{(k)} - \left[w_{i1}^{(k)} \Psi_{i1}^t, \ldots, w_{iq}^{(k)} \Psi_{iq}^t \right] z_t^T \right) \\
& \times \left(y_{i,t} - X_{i,t}' \beta^{(k)} - \left[w_{i1}^{(k)} \Psi_{i1}^t, \ldots, w_{iq}^{(k)} \Psi_{iq}^t \right] z_t^T \right)' \\
& + \left[w_{i1}^{(k)} \Psi_{i1}^t, \ldots, w_{iq}^{(k)} \Psi_{iq}^t \right] P_t^T \left[w_{i1}^{(k)} \Psi_{i1}^t, \ldots, w_{iq}^{(k)} \Psi_{iq}^t \right]' \Bigg\}.
\end{aligned} \tag{B.19}$$

Continuing, we have

$$f_i = \frac{\mathrm{Tr}\left[Q_{\kappa_i^{(k)}} \sum_{t=1}^{T} (e_i' \otimes I_R)(P_{t,t-1}^T + z_{t-1}^T z_t^{T'})(e_i \otimes I_R) \right]}{\mathrm{Tr}\left[Q_{\kappa_i^{(k)}} \sum_{t=1}^{T} (e_i' \otimes I_R)(P_{t-1}^T + z_{t-1}^T z_{t-1}^{T'})(e_i \otimes I_R) \right]}. \tag{B.20}$$

Moreover,

$$\mu_0 = z_0^T, \quad \text{and} \quad \Sigma_0 = P_0^T.$$

Finally, to derive the closed-form expression for $\boldsymbol{w}$, proceed as follows.

Let us consider the quantity in the likelihood involving the only $\boldsymbol{w}_i = (w_{i1}, \dots, w_{iq})'$, that is:

$$\sum_{t=1}^{T} \mathbb{E}\left[\left\| \boldsymbol{y}(\mathcal{S}_{i,t}, t) - \boldsymbol{X}^i(\mathcal{S}_{i,t}), t - \sum_{j=1}^{q} w_{ij} (\boldsymbol{e}_j' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t \right\|^2 \Big| \boldsymbol{y}_{1:T}, \boldsymbol{X}_{1:T} \right],$$

where $\| \cdot \|$ denotes the euclidean norm. Observe that

$$\sum_{j=1}^{q} w_{ij} (\boldsymbol{e}_j' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t = \left[(\boldsymbol{e}_1' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t, \dots, (\boldsymbol{e}_j' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t \right] \boldsymbol{w}_i.$$

Therefore, we have to minimise in $\boldsymbol{w}_i$ the quantity

$$\begin{aligned} &- 2 \sum_{t=1}^{T} \left(\boldsymbol{y}(\mathcal{S}_{i,t}, t) - \boldsymbol{X}^i(\mathcal{S}_{i,t}, t) \right)' \mathbb{E}\left[(\boldsymbol{e}_1' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t, \dots, (\boldsymbol{e}_j' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t \Big| \boldsymbol{y}_{1:T}, \boldsymbol{X}_{1:T} \right] \boldsymbol{w}_i \\ &+ \boldsymbol{w}_i' \mathbb{E}\Bigg[\left[(\boldsymbol{e}_1' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t, \dots, (\boldsymbol{e}_j' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t \right]' \\ &\qquad \times \left[(\boldsymbol{e}_1' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t, \dots, (\boldsymbol{e}_j' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t \right] \Bigg| \boldsymbol{y}_{1:T}, \boldsymbol{X}_{1:T} \Bigg] \boldsymbol{w}_i. \end{aligned} \tag{B.21}$$

The first term in (B.21) is equivalent to $-2\boldsymbol{g}_i' \boldsymbol{w}_i$, where

$$\boldsymbol{g}_i' = \sum_{t=1}^{T} \left(\boldsymbol{y}(\mathcal{S}_{i,t}, t) - \boldsymbol{X}^i(\mathcal{S}_{i,t}, t) \boldsymbol{\beta}^{(k)} \right)' \left[(\boldsymbol{e}_1' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t^T, \dots, (\boldsymbol{e}_q' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t^T \right]. \tag{B.22}$$

The second term in (B.21) is equivalent to $\boldsymbol{w}_i' \boldsymbol{R}_i \boldsymbol{w}_i$, where $\boldsymbol{R}_i = [r_{i,jk}]_{j,k=1}^{q}$ is a $q \times q$ matrix with

$$\begin{aligned} r_{i,jk} &= \sum_{t=1}^{T} \mathbb{E}\left[\boldsymbol{z}_t' \boldsymbol{\Psi}_{i,t}' (\boldsymbol{e}_j' \otimes \boldsymbol{I}_{m_i,t})' (\boldsymbol{e}_k' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t | \boldsymbol{y}_{1:T}, \boldsymbol{X}_{1:T} \right] \\ &= \sum_{t=1}^{T} \boldsymbol{z}_t^{T'} \boldsymbol{\Psi}_{i,t}' (\boldsymbol{e}_j' \otimes \boldsymbol{I}_{m_i,t})' (\boldsymbol{e}_k' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{z}_t^T + \operatorname{Tr}\left(\boldsymbol{\Psi}_{i,t}' (\boldsymbol{e}_j' \otimes \boldsymbol{I}_{m_i,t})' (\boldsymbol{e}_k' \otimes \boldsymbol{I}_{m_i,t}) \boldsymbol{\Psi}_{i,t} \boldsymbol{P}_t^T \right) \\ &= \sum_{t=1}^{T} \operatorname{Tr}\left(\boldsymbol{\Psi}_{i,t}' \left((\boldsymbol{e}_j \boldsymbol{e}_k') \otimes \boldsymbol{I}_{m_i,t} \right) \boldsymbol{\Psi}_{i,t} \left(\boldsymbol{z}_t^T \boldsymbol{z}_t^{T'} + \boldsymbol{P}_t^T \right) \right). \end{aligned}$$

Therefore, we have to minimise

$$-2\boldsymbol{g}_i' \boldsymbol{w}_i + \boldsymbol{w}_i' \boldsymbol{R}_i \boldsymbol{w}_i, \tag{B.23}$$

and therefore

$$\boldsymbol{w}_i^{(k+1)} = \boldsymbol{R}_i^{-1} \boldsymbol{g}_i \tag{B.24}$$

For the updating the value of κ_i we numerically minimise the the function $g_i^{(k)} : \mathbb{R}_+ \to \mathbb{R}_+$ defined as

$$g_i(\kappa_i) = -T \log |\boldsymbol{Q}_{\boldsymbol{\kappa}_i}| + \text{Tr}\Big[\boldsymbol{Q}_{\boldsymbol{\kappa}_i} \sum_{t=1}^{T} (\boldsymbol{e}_i' \otimes \boldsymbol{I}_R)(\boldsymbol{z}_t^T - \boldsymbol{F}^{(k)} \boldsymbol{z}_{t-1}^T)(\boldsymbol{z}_t^T - \boldsymbol{F}^{(k)} \boldsymbol{z}_{t-1}^T)'(\boldsymbol{e}_i \otimes \boldsymbol{I}_R)\Big]$$
$$+ \text{Tr}\Big[\boldsymbol{Q}_{\boldsymbol{\kappa}_i} \sum_{t=1}^{T} (\boldsymbol{e}_i' \otimes \boldsymbol{I}_R)\big(\boldsymbol{P}_t^T + \boldsymbol{F}^{(k)} \boldsymbol{P}_{t-1}^T \boldsymbol{F}^{(k)'} - \boldsymbol{F}^{(k)} \boldsymbol{P}_{t,t-1}^T - (\boldsymbol{P}_{t,t-1}^T)' \boldsymbol{F}^{(k)'}\big)(\boldsymbol{e}_i \otimes \boldsymbol{I}_R)\Big],$$

therefore

$$\kappa_i^{(k+1)} = \arg\min_{\kappa_i} g_i^{(k)}(\kappa_i).$$

B.6 Boundary Effects

In this section, we recall the study of the impact of the Neumann boundary conditions on the approximation of the error term $\boldsymbol{\eta}(\boldsymbol{s}, t)$ as in Lindgren, Håvard Rue, and Lindström (2011). In particular, consider the SPDE:

$$\begin{aligned} &(\kappa^2 - \Delta)^{\alpha/2} x(\boldsymbol{u}) = \mathcal{W}(\boldsymbol{u}), \quad \boldsymbol{u} \in \mathcal{D}, \\ &\partial_n (\kappa^2 - \Delta)^j x(\boldsymbol{u}) = 0, \quad \boldsymbol{u} \in \partial\mathcal{D}, \quad j = 0, 1, \ldots, \left\lfloor \frac{\alpha - 1}{2} \right\rfloor, \end{aligned} \tag{B.25}$$

where (B.25) uses the same notation as in Appendix B.3. In the one-dimensional case, the following result characterises the covariance structure induced by these boundary conditions. The result naturally extends to higher-dimensional domains that are generalised rectangles in $\mathbb{R}^d$.

Theorem 6. *Let x be the solution to* (B.25) *on the domain $\mathcal{D} = [0, L]$, with integer order $\alpha > 0$. Then the covariance function of x is given by*

$$cov\{x(u), x(v)\} = \sum_{k=-\infty}^{\infty} \{r_M(u, v - 2kL) + r_M(u, 2kL - v)\},$$

where r_M denotes the Matérn covariance function defined on the real line $\mathbb{R}$.

This result reveals that the covariance can be interpreted as a sum over reflections (or foldings) of the Matérn covariance across the domain boundaries. When the effective range of the Matérn covariance is small compared to the domain size L, an accurate approximation is obtained by retaining only the three dominant terms:

$$\begin{aligned} \text{cov}\{x(u), x(v)\} &\approx r_M(u, v) + r_M(u, -v) + r_M(u, 2L - v), \\ &= r_M(0, v - u) + r_M(0, v + u) + r_M(0, 2L - (v + u)). \end{aligned}$$

Furthermore, away from the boundaries—specifically, at distances greater than twice the range parameter from the domain edges—the folded covariance becomes nearly indistinguishable from the stationary Matérn covariance on $\mathbb{R}$. This argument justifies the proposed construction of the precision matrix $\boldsymbol{Q}_{k_i}$ in Section 6.5.2.

B.7 Parameters Estimation

	T = 50						T = 100					
	m = 100		m = 225		m = 400		m = 100		m = 225		m = 400	
Parameter	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
β_1	-0.001	0.016	-0.001	0.011	-2.59E-05	0.007	0.003	0.012	0.000	0.008	0.000	0.005
β_2	-0.002		-0.001		0.002		0.000		0.000		0.001	
β_3	-0.002		0.001		0.000		0.004		0.001		-0.001	
σ_1^2	0.299	0.234	0.175	0.138	0.117	0.086	0.287	0.223	0.166	0.130	0.105	0.079
σ_2^2	0.282		0.169		0.100		0.273		0.165		0.096	
σ_3^2	0.123		0.065		0.031		0.114		0.053		0.026	
k_1	-0.828	0.215	-2.101	0.147	-2.789	0.154	-1.474	0.135	-1.522	0.107	-2.078	0.123
k_2	-0.235		-0.454		-0.415		-0.131		-0.315		-0.285	
W_1	0.025	0.185	-0.008	0.211	-0.009	0.222	0.044	0.187	0.011	0.225	0.009	0.240
W_2	0.159		0.210		0.221		0.179		0.232		0.242	
W_3	-0.022		-0.030		-0.016		0.010		-0.006		0.004	
W_4	0.037		0.053		0.065		0.048		0.065		0.072	
W_5	0.005		4.99E-05		0.002		0.011		0.004		0.007	
W_6	0.138		0.176		0.186		0.154		0.198		0.198	
f_1	-0.034	0.100	-0.034	0.088	-0.042	0.080	-0.042	0.083	-0.038	0.068	-0.055	0.116
f_2	-0.018		-0.023		-0.013		-0.016		-0.006		-0.011	
$\text{rmse}_{\text{train}}$	1.049	0.014	1.012	0.009	0.992	0.006	1.042	0.010	1.006	0.006	0.987	0.004
$\text{rmse}_{\text{test}}$	1.230	0.010	1.170	0.008	1.133	0.009	1.217	0.007	1.154	0.005	1.117	0.006
time	282.247	71.186	1,221.860	294.052	4,202.809	1,017.998	305.851	82.131	1,646.330	487.763	6,242.530	2,337.307

Table 29: $R = 75\%m$. Bias and RMSE of estimated parameters across different time lengths (T) and number of locations (m).

	T = 50						T = 100					
	m = 100		m = 225		m = 400		m = 100		m = 225		m = 400	
Parameter	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
β_1	-0.005		0.000		-0.001		0.000		-0.001		-0.001	
β_2	0.005	0.018	-0.001	0.011	0.003	0.008	0.000	0.013	0.002	0.008	-0.001	0.006
β_3	-0.002		0.001		0.000		0.000		0.000		0.001	
σ_1^2	0.652		0.432		0.321		0.636		0.410		0.296	
σ_2^2	0.443	0.441	0.314	0.295	0.229	0.217	0.464	0.438	0.309	0.282	0.222	0.203
σ_3^2	0.222		0.129		0.091		0.218		0.119		0.078	
k_1	-2.437	0.304	2.890	0.345	0.396	0.225	-2.091	0.287	3.489	0.366	0.073	0.087
k_2	1.097		0.152		-0.312		1.112		0.147		-0.151	
W_1	0.278		0.092		0.006		0.327		0.119		0.017	
W_2	0.071		0.162		0.249		0.104		0.200		0.246	
W_3	0.142	0.331	0.071	0.204	0.004	0.259	0.162	0.363	0.100	0.243	0.016	0.249
W_4	0.014		0.063		0.092		0.030		0.078		0.093	
W_5	0.204		0.054		0.010		0.226		0.060		0.014	
W_6	0.070		0.132		0.202		0.095		0.162		0.200	
f_1	-0.224	0.329	-0.054	0.099	-0.036	0.087	-0.214	0.299	-0.055	0.080	-0.041	0.065
f_2	0.009		-0.001		0.013		0.014		0.019		0.017	
`rmse_train`	1.160	0.015	1.101	0.008	1.069	0.006	1.160	0.012	1.094	0.005	1.061	0.004
`rmse_test`	1.255	0.010	1.186	0.008	1.144	0.008	1.250	0.007	1.180	0.006	1.138	0.006
`time`	130.022	48.105	269.991	81.193	857.997	311.061	118.222	39.473	415.764	149.401	1,413.564	536.759

Table 30: $R = 25\%m$. Bias and RMSE of estimated parameters across different time lengths (T) and number of locations (m).

List of Figures

List of Tables

References

Akima, Hiroshi (1970). "A New Method of Interpolation and Smooth Curve Fitting Based on Local Procedures". In: *J. ACM* 17(4), pp. 589–602. DOI: https://doi.org/10.1145/321607.321609.

Alduchov, Oleg A and Robert E Eskridge (1996). "Improved Magnus Form Approximation of Saturation Vapor Pressure". In: *J. Appl. Meteorol. Climatol.* 35(4), pp. 601–609. DOI: https://doi.org/10.1175/1520-0450(1996)035<0601:IMFAOS>2.0.CO;2.

Altman, Naomi S (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". In: *The American Statistician* 46.3, pp. 175–185.

Amisigo, BA and NC Van De Giesen (2005). "Using a spatio-temporal dynamic state-space model with the EM algorithm to patch gaps in daily riverflow series". In: *Hydrology and Earth System Sciences* 9.3, pp. 209–224.

ARPA Lombardia Settore Monitoraggi Ambientali (2019). *Progetto ammoniaca: Relazione finale triennio 2017-2019. ARPA Lombardy portal* https://www.arpalombardia.it/Pages/Aria/Aria-Progetti/Progetto-Ammoniaca.aspx.

ARPA Lombardia Settore Monitoraggi Ambientali, INEMAR (2022). *INEMAR, inventario emissioni in atmosfera: Emissioni in regione Lombardia nell'anno 2019, version in public review. ARPA Lombardy portal* https://www.inemar.eu/xwiki/bin/view/InemarDatiWeb/Risultati+Regionali.

Baayen, Harald et al. (2017). "The cave of shadows: Addressing the human factor with generalized additive mixed models". In: *Journal of Memory and Language* 94, pp. 206–234.

Bai, Jushan and Peng Wang (2015). "Identification and Bayesian estimation of dynamic factor models". In: *Journal of Business & Economic Statistics* 33.2, pp. 221–240.

Banerjee, Sudipto et al. (2008). "Gaussian predictive process models for large spatial data sets". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70.4, pp. 825–848.

Belgiu, Mariana and Lucian Drăguţ (2016). "Random forest in remote sensing: A review of applications and future directions". In: *ISPRS journal of photogrammetry and remote sensing* 114, pp. 24–31.

Benth, Jūratė Saltytė, Fred Espen Benth, and Paulius Jalinskas (2007). "A Spatial-temporal Model for Temperature with Seasonal Variance". In: *Journal of Applied Statistics* 34.7, pp. 823–841. DOI: 10.1080/02664760701511398. eprint: https://doi.org/10.1080/02664760701511398. URL: https://doi.org/10.1080/02664760701511398.

Berrocal, Veronica J et al. (2020). "A comparison of statistical and machine learning methods for creating national daily maps of ambient $PM_{2.5}$ concentration". In: *Atmospheric Environment* 222, p. 117130.

Bertaccini, Pancrazio, Vanja Dukic, and Rosaria Ignaccolo (2012). "Modeling the short-term effect of traffic and meteorology on air pollution in Turin with generalized additive models". In: *Advances in Meteorology* 2012.

Besag, Julian (1974). "Spatial interaction and the statistical analysis of lattice systems". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2, pp. 192–225.

Besag, Julian and Charles Kooperberg (1995). "On conditional and intrinsic autoregressions". In: *Biometrika* 82.4, pp. 733–746.

Bevilacqua, Moreno, Alessandro Fassò, et al. (2016). "Covariance tapering for multivariate Gaussian random fields estimation". In: *Statistical Methods & Applications* 25.1, pp. 21–37.

Bevilacqua, Moreno, Jorge Mateu, et al. (2010). "Weighted composite likelihood-based tests for space-time separability of covariance functions". In: *Statistics and Computing* 20, pp. 283–293.

Boldo, Elena et al. (2014). "Air quality modeling and mortality impact of fine particles reduction policies in Spain". In: *Environmental research* 128, pp. 15–26.

Bono, Roser, Rafael Alarcón, and María J Blanca (2021). "Report quality of generalized linear mixed models in psychology: A systematic review". In: *Frontiers in psychology* 12, p. 666182.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45, pp. 5–32.

Brenner, Susanne C (2008). *The mathematical theory of finite element methods*. Springer.

Büttner, G. et al. (2021). *Corine Land Cover user manual. Copernicus Land Monitoring Service, European Environment Agency,* https://land.copernicus.eu/user-corner/technical-library/clc-product-user-manual.

Calculli, Crescenza et al. (2015). "Maximum likelihood estimation of the multivariate Hidden Dynamic Geostatistical Model with application to air quality in Apulia, Italy". In: *Environmetrics* 26.6, pp. 406–417.

Cameletti, Michela, Rosaria Ignaccolo, and Stefano Bande (2011). "Comparing spatio-temporal models for particulate matter in Piemonte". In: *Environmetrics* 22.8, pp. 985–996.

Cameletti, Michela, Finn Lindgren, et al. (2013). “Spatio-temporal modeling of particulate matter concentration through the SPDE approach”. In: *AStA Advances in Statistical Analysis* 97, pp. 109–131.

Chang, Le and Tao Zou (2022). “Spatio-temporal analysis of air pollution in North China Plain”. In: *Environmental and Ecological Statistics* 29.2, pp. 271–293.

Chen, Lin et al. (2019). “Assessment of multi-wavelength SAR and multispectral instrument data for forest aboveground biomass mapping using random forest kriging”. In: *Forest ecology and management* 447, pp. 12–25.

Chiles, Jean-Paul and Pierre Delfiner (2012). *Geostatistics: modeling spatial uncertainty*. John Wiley & Sons.

Ciarlet, Philippe G (2002). *The finite element method for elliptic problems*. SIAM.

Clappier, Alain et al. (2021). “Impact of SO_x, NO_x and NH_3 emission reductions on $PM_{2.5}$ concentrations across Europe: Hints for future measure development”. In: *Environment International* 156, p. 106699.

Clarotto, Lucia et al. (2024). “The SPDE approach for spatio-temporal datasets with advection and diffusion”. In: *Spatial Statistics*, p. 100847.

Colombi, C et al. (2024). “Monitoring ammonia concentrations in more than 10 stations in the Po Valley for the period 2007–2022 in relation to the evolution of different sources”. In: *Frontiers in Environmental Health* 3, p. 1249457.

Cressie, Noel (1985). “Fitting variogram models by weighted least squares”. In: *Journal of the international Association for mathematical Geology* 17, pp. 563–586.

Cressie, Noel and Hsin-Cheng Huang (1999). “Classes of nonseparable, spatio-temporal stationary covariance functions”. In: *Journal of the American Statistical association* 94.448, pp. 1330–1339.

Cressie, Noel and Gardar Johannesson (2008). “Fixed rank kriging for very large spatial data sets”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70.1, pp. 209–226.

Cressie, Noel and John Kornak (2003). “Spatial statistics in the presence of location error with an application to remote sensing of the environment”. In: *Statistical science*, pp. 436–456.

Cressie, Noel, Tao Shi, and Emily L Kang (2010). “Fixed rank filtering for spatio-temporal data”. In: *Journal of Computational and Graphical Statistics* 19.3, pp. 724–745.

Cressie, Noel and Christopher K Wikle (1998). “The variance-based cross-variogram: you can add apples and oranges”. In: *Mathematical Geology* 30.7, pp. 789–799.

— (2002). “Space-time Kalman filter”. In: *Encyclopedia of environmetrics* 4, pp. 2045–2049.

Cressie, Noel and Christopher K Wikle (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Datta, Abhirup et al. (2016). "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets". In: *Journal of the American Statistical Association* 111.514, pp. 800–812.

Daw, Ranadeep and Christopher K Wikle (2023). "REDS: Random ensemble deep spatial prediction". In: *Environmetrics* 34.1, e2780.

De Boor, Carl and Carl De Boor (1978). *A practical guide to splines*. Vol. 27. springer New York.

De Meij, A et al. (2009). "The sensitivity of the CHIMERE model to emissions reduction scenarios on air quality in Northern Italy". In: *Atmospheric environment* 43.11, pp. 1897–1907.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society: series B (methodological)* 39.1, pp. 1–22.

Diggle, Peter J, Jonathan A Tawn, and Rana A Moyeed (1998). "Model-based geostatistics". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 47.3, pp. 299–350.

Dominici, Francesca et al. (2006). "Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases". In: *Jama* 295.10, pp. 1127–1134.

Du, Juan, Hao Zhang, and VS2549562 Mandrekar (2009). "Fixed-domain asymptotic properties of tapered maximum likelihood estimators". In.

Durbin, James and Siem Jan Koopman (2012). *Time Series Analysis by State Space Methods*. Vol. 38. OUP Oxford.

Engle, Robert F. (1982). "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation". In: *Econometrica* 50.4, pp. 987–1007. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912773 (visited on 01/27/2024).

Ern, Alexandre and Jean-Luc Guermond (2004). *Theory and practice of finite elements*. Vol. 159. Springer.

European Environment Agency (EEA) (2022). *Premature deaths due to exposure to fine particulate matter in Europe (8th EAP)*. URL: https://www.eea.europa.eu/ims/health-impacts-of-exposure-to.

European Environmental Agency (2018). *European Union emission inventory report 1990–2016 under the UNECE Convention on Long-range Transboundary Air Pollution (LRTAP)*. Publications Office. DOI: doi/10.2800/571876.

— (Nov. 2022). *Air quality in Europe*. Last Access: 18-03-2023.

European Environmental Agency (EEA) - Air pollution section (2021). *Data Air Quality e-Reporting (AQ e-Reporting). European Environmental Agency portal* https://www.eea.europa.eu/ds_resolveuid/d4b3817c7a1640459e783f9342aac786.

European Union (2018). *Corine Land Cover (CLC), version 2020_20u1. Copernicus Land Monitoring Service, European Environment Agency* https://land.copernicus.eu/pan-european/corine-land-cover/clc2018.

Fahrmeir, Ludwig, Thomas Kneib, and Stefan Lang (2004). "Penalized structured additive regression for space-time data: a Bayesian perspective". In: *Statistica Sinica*, pp. 731–761.

Fahrmeir, Ludwig and Stefan Lang (2001). "Bayesian inference for generalized additive mixed models based on Markov random field priors". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 50.2, pp. 201–220.

Fassò, Alessandro (2023). "To what extent airborne particulate matters are influenced by ammonia and nitrogen oxides?" In: *arXiv preprint arXiv:2310.09302*.

Fassò, Alessandro and Francesco Finazzi (2011). "Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data". In: *Environmetrics* 22.6, pp. 735–748.

Fassò, Alessandro, Francesco Finazzi, and Ferdinand Ndongo (2016). "European population exposure to airborne pollutants based on a multivariate spatio-temporal model". In: *Journal of agricultural, biological, and environmental statistics* 21, pp. 492–511.

Fassò, Alessandro, Paolo Maranzano, and Philipp Otto (2022). "Spatiotemporal variable selection and air quality impact assessment of COVID-19 lockdown". In: *Spatial Statistics* 49, p. 100549. ISSN: 2211-6753. DOI: https://doi.org/10.1016/j.spasta.2021.100549. URL: https://www.sciencedirect.com/science/article/pii/S2211675321000592.

Fassò, Alessandro, Jacopo Rodeschini, et al. (2023a). "Agrimonia: a dataset on livestock, meteorology and air quality in the Lombardy region, Italy". In: *Scientific Data* 10.1, p. 143.

— (May 2023b). *AgrImOnIA: Open Access dataset correlating livestock and air quality in the Lombardy region, Italy*. Version 3.0.0. DOI: 10.5281/zenodo.7956006. URL: https://doi.org/10.5281/zenodo.7956006.

Field, David A (1988). "Laplacian smoothing and Delaunay triangulations". In: *Communications in applied numerical methods* 4.6, pp. 709–712.

Finazzi, Francesco, E Marian Scott, and Alessandro Fassò (2013). "A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 62.2, pp. 287–308.

Fioravanti, Guido et al. (2021). “Spatio-temporal modelling of PM_{10} daily concentrations in Italy using the SPDE approach”. In: *Atmospheric Environment* 248, p. 118192.

Fox, Eric W, Jay M Ver Hoef, and Anthony R Olsen (2020). “Comparing spatial regression to random forests for large environmental data sets”. In: *PloS one* 15.3, e0229509.

Francq, Christian and Jean-Michel Zakoian (2019). *GARCH models: structure, statistical inference and financial applications*. John Wiley & Sons.

Fritsch, F. N. and R. E. Carlson (1980). “Monotone Piecewise Cubic Interpolation”. In: *SIAM J. Numer. Anal.* 17(2), 238–246. http://www.jstor.org/stable/2156610. ISSN: 00361429.

Fuentes, Montserrat and Adrian E Raftery (2005). “Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models”. In: *Biometrics* 61.1, pp. 36–45.

Furrer, Reinhard (2005). “Covariance estimation under spatial dependence”. In: *Journal of multivariate analysis* 94.2, pp. 366–381.

Furrer, Reinhard, Marc G Genton, and Douglas Nychka (2006). “Covariance tapering for interpolation of large spatial datasets”. In: *Journal of Computational and Graphical Statistics* 15.3, pp. 502–523.

Fusta Moro, Alessandro and Alessandro Fassò (2025). “AQCLIM GRINS Dataset - The Italian Daily Dataset on Air Quality”. In: *Zenodo* https://doi.org/10.5281/zenodo.15699805. Accessed: 26 July 2025. URL: https://doi.org/10.5281/zenodo.15699805.

Fusta Moro, Alessandro, Matteo Salis, et al. (2022). “Ammonia emissions and fine particulate matter: some evidence in Lombardy”. In: *Book of Short Papers of the ASA Conference 2022-Data-Drive Decision Making*, pp. 1–6.

Gelfand, Alan E, Li Zhu, and Bradley P Carlin (2001). “On the change of support problem for spatio-temporal data”. In: *Biostatistics* 2.1, pp. 31–45.

Genton, Marc G and William Kleiber (2015). “Cross-covariance functions for multivariate geostatistics”. In: *Statistical Science*.

Gneiting, Tilmann (2002). “Nonseparable, stationary covariance functions for space–time data”. In: *Journal of the american statistical association* 97.458, pp. 590–600.

Goulard, Michel and Marc Voltz (1992). “Linear coregionalization model: tools for estimation and choice of cross-variogram matrix”. In: *Mathematical Geology* 24.3, pp. 269–286.

Grange, S. K., J. Sintermann, and C. Hueglin (2023). “Meteorologically normalised long-term trends of atmospheric ammonia (NH_3) in Switzerland/Liechtenstein and the explanatory role of gas-aerosol partitioning”. In: *Science of The Total Environment* 900, p. 165844. ISSN: 0048-9697.

DOI: https://doi.org/10.1016/j.scitotenv.2023.165844. URL: https://www.sciencedirect.com/science/article/pii/S0048969723044698.

Grange, Stuart K, Jörg Sintermann, and Christoph Hueglin (2023). "Meteorologically normalised long-term trends of atmospheric ammonia (NH3) in Switzerland/Liechtenstein and the explanatory role of gas-aerosol partitioning". In: *Science of the Total Environment*, p. 165844.

Granier, C et al. (2019). "The Copernicus atmosphere monitoring service global and regional emissions". In: *Reading, United Kingdom: Copernicus Atmosphere Monitoring Service*. DOI: https://doi.org/10.24380/d0bn-kx16.

Guinness, Joseph (2019). "Spectral density estimation for random fields via periodic embeddings". In: *Biometrika* 106.2, pp. 267–286.

— (2022). "Nonparametric spectral methods for multivariate spatial and spatial–temporal data". In: *Journal of Multivariate Analysis* 187, p. 104823.

Guo, Peng-Tao et al. (2015). "Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach". In: *Geoderma* 237, pp. 49–59.

Guttorp, Peter and Tilmann Gneiting (2006). "Studies in the history of probability and statistics XLIX on the Matérn correlation family". In: *Biometrika* 93.4, pp. 989–995.

Hamilton, James D (2020). *Time series analysis*. Princeton university press.

Handcock, Mark S and James R Wallis (1994). "An approach to statistical spatial-temporal modeling of meteorological fields". In: *Journal of the American Statistical Association* 89.426, pp. 368–378.

Harvey, Andrew C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781107049994.

Hastie, Trevor and Robert Tibshirani (1987). "Generalized additive models: some applications". In: *Journal of the American Statistical Association* 82.398, pp. 371–386.

Heaton, Matthew J et al. (2019). "A case study competition among methods for analyzing large spatial data". In: *Journal of Agricultural, Biological and Environmental Statistics* 24, pp. 398–425.

Henschel, Susann and Gabrielle Chan (2013). "Health risks of air pollution in Europe—HRAPIE project". In: *Copenhagen, Denmark: World Health Organization*.

Hersbach, H. et al. (2018). *ERA5 hourly data on single levels from 1979 to present. Copernicus climate change service (C3S). Climate Data Store (CDS), https://doi.org/10.24381/cds.adbb2d47*. Journal Article.

Hsieh, Cho-jui et al. (2011). "Sparse Inverse Covariance Matrix Estimation Using Quadratic Approximation". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor et al.

Vol. 24. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2011/file/2ba8698b79439589fdd2b0f7218d8b07-Paper.pdf.

Huang, Guowen, Duncan Lee, and Marian Scott (2015). "An integrated Bayesian model for estimating the long-term health effects of air pollution by fusing modelled and measured pollution data: A case study of nitrogen dioxide concentrations in Scotland". In: *Spatial and Spatio-temporal Epidemiology* 14, pp. 63–74.

Huang, Hsin-Cheng and Noel Cressie (1996). "Spatio-temporal prediction of snow water equivalent using the Kalman filter". In: *Computational Statistics & Data Analysis* 22.2, pp. 159–175.

Hulshof, Catherine M. and Maria Natalia Umana (2023). "Power laws and plant trait variation in spatio-temporally heterogeneous environments". In: *Global Ecology and Biogeography* 32.2, pp. 310–323. DOI: https://doi.org/10.1111/geb.13620. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/geb.13620. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/geb.13620.

Humpherys, Jeffrey, Preston Redd, and Jeremy West (2012). "A Fresh Look at the Kalman Filter". In: *SIAM Review* 54.4, pp. 801–823. DOI: 10.1137/100799666. URL: https://doi.org/10.1137/100799666.

Ignaccolo, Rosaria, Jorge Mateu, and Ramon Giraldo (2014). "Kriging with external drift for functional data for air quality monitoring". In: *Stochastic environmental research and risk assessment* 28, pp. 1171–1186.

Inness, A. et al. (2019). "The CAMS reanalysis of atmospheric composition". In: *Atmos. Chem. Phys.* 19.6, pp. 3515–3556. DOI: https://doi.org/10.5194/acp-19-3515-2019. URL: https://acp.copernicus.org/articles/19/3515/2019/.

Italian Ministry of Health (2021). *Banca Dati Nazionale (BDN) dell'Anagrafe Zootecnica Istituita dal Ministero della Salute presso il CSN dell'Istituto "G. Caporale" di Teramo. Italian veterinary information system* https://www.vetinfo.it/j6_statistiche/index.html#/.

Jensen, Jens Ledet and Niels Væver Petersen (1999). "Asymptotic normality of the maximum likelihood estimator in state space models". In: *The Annals of Statistics* 27.2, pp. 514–535.

Jerrett, Michael et al. (2005). "A review and evaluation of intraurban air pollution exposure models". In: *Journal of Exposure Science & Environmental Epidemiology* 15.2, pp. 185–204.

Jiang, Jiming and Thuan Nguyen (2007). *Linear and generalized linear mixed models and their applications*. Vol. 1. Springer.

Jona Lasinio, Giovanna, Gianluca Mastrantonio, and Alessio Pollice (2013). "Discussing the "big n problem"". In: *Statistical Methods & Applications* 22, pp. 97–112.

Jones, Richard H and Yiming Zhang (1997). "Models for continuous stationary space-time processes". In: *Modelling longitudinal and spatially correlated data*. Springer, pp. 289–298.

Jordan, Michael I. (2013). "On statistics, computation and scalability". In: *Bernoulli* 19.4, pp. 1378–1390. DOI: 10.3150/12-BEJSP17. URL: https://projecteuclid.org/journals/bernoulli/volume-19/issue-4/On-statistics-computation-and-scalability/10.3150/12-BEJSP17.full.

Kahaner, David, Cleve Moler, and Stephen Nash (1989). *Numerical Methods and Software*. USA: Prentice-Hall, Inc. ISBN: 0136272584.

Kammann, EE and Matthew P Wand (2003). "Geoadditive models". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 52.1, pp. 1–18.

Kaufman, Cari G, Mark J Schervish, and Douglas W Nychka (2008). "Covariance tapering for likelihood-based estimation in large spatial data sets". In: *Journal of the American Statistical Association* 103.484, pp. 1545–1555.

Kiesewetter, Gregor et al. (2015). "Modelling PM2.5 impact indicators in Europe: Health effects and legal compliance". In: *Environmental Modelling & Software* 74, pp. 201–211.

Knape, Jonas (2016). "Decomposing trends in Swedish bird populations using generalized additive mixed models". In: *Journal of Applied Ecology* 53.6, pp. 1852–1861.

Kneib, Thomas, Felix Knauer, and Helmut Küchenhoff (2011). "A general approach to the analysis of habitat selection". In: *Environmental and Ecological Statistics* 18, pp. 1–25.

Krige, Daniel G (1951). "A statistical approach to some basic mine valuation problems on the Witwatersrand". In: *Journal of the Southern African Institute of Mining and Metallurgy* 52.6, pp. 119–139.

Kulkarni, Padmavati et al. (2022). "Which model to choose? Performance comparison of statistical and machine learning models in predicting $PM_{2.5}$ from high-resolution satellite aerosol optical depth". In: *Atmospheric Environment* 282, p. 119164.

Kyriakidis, Phaedon C and André G Journel (1999). "Geostatistical space–time models: a review". In: *Mathematical geology* 31.6, pp. 651–684.

Lark, RM (2003). "Two robust estimators of the cross-variogram for multivariate geostatistical analysis of soil properties". In: *European Journal of Soil Science* 54.1, pp. 187–202.

Lawson, Andrew B (2018). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. Chapman and Hall/CRC.

Lee, Duncan and Gavin Shaddick (2010). "Spatial modeling of air pollution in studies of its short-term health effects". In: *Biometrics* 66.4, pp. 1238–1246.

Lewbel, Arthur (2019). “The identification zoo: Meanings of identification in econometrics”. In: *Journal of Economic Literature* 57.4, pp. 835–903.

Li, Tizheng, Yuping Wang, and Ke Fang (2024). “A semiparametric dynamic higher-order spatial autoregressive model”. In: *Statistical Papers* 65.2, pp. 1085–1123.

Lindgren, Finn, Håvard Rue, and Johan Lindström (2011). “An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 73.4, pp. 423–498.

Liu, Haitao et al. (2020). “When Gaussian process meets big data: A review of scalable GPs”. In: *IEEE transactions on neural networks and learning systems* 31.11, pp. 4405–4423.

Liu, Ying, Guofeng Cao, et al. (2018). “Improve ground-level $PM_{2.5}$ concentration mapping using a random forests-based geostatistical approach”. In: *Environmental pollution* 235, pp. 272–282.

Liu, Ying, Naizhuo Zhao, et al. (2019). “Revisiting the estimations of $PM_{2.5}$-attributable mortality with advancements in $PM_{2.5}$ mapping and mortality statistics”. In: *Science of the Total Environment* 666, pp. 499–507.

Lombardy Region, Sistema Informativo Agricoltura Regione Lombardia (SIARL) (2019). *Carta Uso Agricolo - Dati SIARL dal 2012 al 2019. Territorial Information of Region Lombardy,* https://www.geoportale.regione.lombardia.it/metadati?p_p_id=detailSheetMetadata_WAR_gptmetadataportlet&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&_detailSheetMetadata_WAR_gptmetadataportlet_uuid=%7B83483117-8742-4A1F-A16E-3A48AEE2EBE2%7D.

Lonati, Giovanni and Stefano Cernuschi (2020). “Temporal and spatial variability of atmospheric ammonia in the Lombardy region (Northern Italy)”. In: *Atmospheric Pollution Research* 11.12, pp. 2154–2163.

Lovarelli, Daniela, Cecilia Conti, et al. (2020). “Describing the trend of ammonia, particulate matter and nitrogen oxides: The role of livestock activities in northern Italy during Covid-19 quarantine”. In: *Environmental research* 191, p. 110048.

Lovarelli, Daniela, Davide Fugazza, et al. (2021). “Comparison of ammonia air concentration before and during the spread of COVID-19 in Lombardy (Italy) using ground-based and satellite data”. In: *Atmospheric Environment* 259, p. 118534.

Lu, Meng, Joaquin Cavieres, and Paula Moraga (2023). “A Comparison of Spatial and Nonspatial Methods in Statistical Modeling of NO_2: Prediction Accuracy, Uncertainty Quantification, and Model Interpretation”. In: *Geographical Analysis*.

Mardia, Kanti V and Roger J Marshall (1984). “Maximum likelihood estimation of models for residual covariance in spatial regression”. In: *Biometrika* 71.1, pp. 135–146.

Martin, Buhmann, MJ Buhmann, and J Ablowitz (2003). “Radial basis functions: theory and implementations”. In: *Cambridge University (ISBN: 0-521-63338-9.)*, p. 5.

Matérn, Bertil (1960). *Spatial variation*. Vol. 36. Springer Science & Business Media.

Matheron, G (1962). “Precision of exploring a stratified formation by boreholes with rigid spacing—application to a bauxite deposit”. In: *Mining Research*. Elsevier, pp. 407–422.

Matheron, Georges (1963). “Principles of geostatistics”. In: *Economic Geology* 58.8, pp. 1246–1266.

McLachlan, Geoffrey J and Thriyambakam Krishnan (2008). *The EM algorithm and extensions*. John Wiley & Sons.

McMillan, Nancy J et al. (2010). “Combining numerical model output and particulate data using Bayesian space–time modeling”. In: *Environmetrics: The official journal of the International Environmetrics Society* 21.1, pp. 48–65.

Megaritis, AG et al. (2013). “Response of fine particulate matter concentrations to changes of emissions and temperature in Europe”. In: *Atmospheric Chemistry and Physics* 13.6, pp. 3423–3443.

Merk, Miryam S and Philipp Otto (2020). “Estimation of anisotropic, time-varying spatial spillovers of fine particulate matter due to wind direction”. In: *Geographical Analysis* 52.2, pp. 254–277.

Meyer, Hanna et al. (2018). “Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation”. In: *Environmental Modelling & Software* 101, pp. 1–9.

Muñoz Sabater, J. (2021). *ERA5-Land hourly data from 1981 to present. Copernicus climate change service (C3S). Climate Data Store (CDS)* https://doi.org/10.24381/cds.e2161bac. Journal Article.

Mur, Jesus and Ana Angulo (2009). “Model selection strategies in a spatial setting: Some additional results”. In: *Regional Science and Urban Economics* 39.2, pp. 200–213. ISSN: 0166-0462. DOI: https://doi.org/10.1016/j.regsciurbeco.2008.05.018. URL: https://www.sciencedirect.com/science/article/pii/S0166046208001051.

Myers, Donald E (1991). “Pseudo-cross variograms, positive-definiteness, and cokriging”. In: *Mathematical Geology* 23.6, pp. 805–816.

— (1992). “Kriging, cokriging, radial basis functions and the role of positive definiteness”. In: *Computers & Mathematics with Applications* 24.12, pp. 139–148.

Nag, Pratik, Ying Sun, and Brian J. Reich (2023). “Spatio-temporal DeepKriging for interpolation and probabilistic forecasting”. In: *Spatial Statistics*, p. 100773. ISSN: 2211-6753. DOI: https:

//doi.org/10.1016/j.spasta.2023.100773. URL: https://www.sciencedirect.com/science/article/pii/S2211675323000489.

Nenes, A. et al. (2020). "Aerosol pH and liquid water content determine when particulate matter is sensitive to ammonia and nitrate availability". In: *Atmos. Chem. Phys.* 20(5), pp. 3249–3258. ISSN: 1680-7324. DOI: https://doi.org/10.5194/acp-20-3249-2020. URL: https://acp.copernicus.org/articles/20/3249/2020/.

Nowak, Gen and AH Welsh (2020). "Improved prediction for a spatio-temporal model". In: *Environmental and Ecological Statistics* 27, pp. 631–648.

Nychka, Douglas et al. (2015). "A multiresolution Gaussian process model for the analysis of large spatial datasets". In: *Journal of Computational and Graphical Statistics* 24.2, pp. 579–599.

Otto, Philipp, Osman Doğan, and Süleyman Taşpınar (2023). "A dynamic spatiotemporal stochastic volatility model with an application to environmental risks". In: *Econometrics and Statistics*.

Otto, Philipp, Alessandro Fassò, and Paolo Maranzano (2024). "A review of regularised estimation methods and cross-validation in spatiotemporal statistics". In: *Statistic Surveys* 18, pp. 299–340.

Otto, Philipp, Alessandro Fusta Moro, et al. (Feb. 2024). "Spatiotemporal Modelling of $PM_{2.5}$ Concentrations in Lombardy (Italy): A Comparative Study". In: *Environmental and Ecological Statistics*. ISSN: 1573-3009. DOI: 10.1007/s10651-023-00589-0. URL: https://doi.org/10.1007/s10651-023-00589-0.

Otto, Philipp, Andreas Piter, and Rik Gijsman (2021). "Statistical analysis of beach profiles–A spatiotemporal functional approach". In: *Coastal engineering* 170, p. 103999.

Otto, Philipp, Wolfgang Schmid, and Robert Garthoff (2018). "Generalised spatial and spatiotemporal autoregressive conditional heteroscedasticity". In: *Spatial Statistics* 26, pp. 125–145. ISSN: 2211-6753. DOI: https://doi.org/10.1016/j.spasta.2018.07.005. URL: https://www.sciencedirect.com/science/article/pii/S2211675318300794.

Papritz, A, HR Künsch, and R Webster (1993). "On the pseudo cross-variogram". In: *Mathematical Geology* 25.8, pp. 1015–1026.

Pernigotti, Denise et al. (2012). "Impact of meteorological modelling on air quality: Summer and winter episodes in the Po valley (Northern Italy)". In: *International Journal of Environment and Pollution* 50.1-4, pp. 111–119.

Piter, Andreas, Philipp Otto, and Hamza Alkhatib (2022). "The Helsinki Bike-Sharing System—Insights Gained from a Spatiotemporal Functional Model". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 185.3, pp. 1294–1318.

Pozzer, Andrea et al. (2017). “Impact of agricultural emission reductions on fine-particulate matter and public health”. In: *Atmospheric Chemistry and Physics* 17.20, pp. 12813–12826.

Qi, Yanjun (2012). “Random forest for bioinformatics”. In: *Ensemble machine learning: Methods and applications*, pp. 307–323.

Quarteroni, Alfio and Alberto Valli (2008). *Numerical approximation of partial differential equations*. Vol. 23. Springer Science & Business Media.

Rasmussen, Carl Edward (2003). “Gaussian processes in machine learning”. In: *Summer school on machine learning*. Springer, pp. 63–71.

Rodeschini, Jacopo, Alessandro Fassò, et al. (2024). “Scenario analysis of livestock-related PM2. 5 pollution based on a new heteroskedastic spatiotemporal model”. In: *Socio-Economic Planning Sciences* 96, p. 102053.

Rodeschini, Jacopo, Lorenzo Tedesco, et al. (2025). “Multivariate Low-Rank State-Space Model with SPDE Approach for High-Dimensional Data”. In: arXiv: 2509.12825 [stat.ME]. URL: https://arxiv.org/abs/2509.12825.

Rue, Havard and Leonhard Held (2005). *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC.

Schmidt, Jonathan et al. (2023). “The rank-reduced Kalman filter: Approximate dynamical-low-rank filtering in high dimensions”. In: *Advances in Neural Information Processing Systems* 36, pp. 61364–61376.

Sekulić, Aleksandar et al. (2020). “Random forest spatial interpolation”. In: *Remote Sensing* 12.10, p. 1687.

Shao, Yanchuan et al. (2020). “Estimating daily ground-level PM2.5 in China with random-forest-based spatiotemporal kriging”. In: *Science of The Total Environment* 740, p. 139761.

Shumway, Robert H, David S Stoffer, and David S Stoffer (2000). *Time series analysis and its applications*. Vol. 3. Springer.

Smith, Brian J., Jun Yan, and Mary Kathryn Cowles (2008). “Unified Geostatistical Modeling for Data Fusion and Spatial Heteroskedasticity with R Package ramps”. In: *Journal of Statistical Software* 25.10. DOI: 10.18637/jss.v025.i10. URL: https://www.jstatsoft.org/index.php/jss/article/view/v025i10.

Song, Yan, Wenlin Dai, and Marc G Genton (2024). “Large-scale low-rank gaussian process prediction with support points”. In: *Journal of the American Statistical Association*, pp. 1–12.

Stafoggia, Massimo et al. (2019). “Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model”. In: *Environment international* 124, pp. 170–179.

Stein, Michael L (1999). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

— (2005). “Space–time covariance functions”. In: *Journal of the American Statistical Association* 100.469, pp. 310–321.

Stein, Michael L, Zhiyi Chi, and Leah J Welty (2004). “Approximating likelihoods for large spatial data sets”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 66.2, pp. 275–296.

Stoffer, David S and Kent D Wall (1991). “Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman filter”. In: *Journal of the American Statistical Association* 86.416, pp. 1024–1033.

Strobl, Carolin et al. (2008). “Conditional variable importance for random forests”. In: *BMC bioinformatics* 9, pp. 1–11.

Tedesco, Lorenzo, Jacopo Rodeschini, and Philipp Otto (2025). “Computational Benchmark Study in Spatio-Temporal Statistics With a Hands-On Guide to Optimise R”. In: *Environmetrics* 36.5, e70017.

Thunis, P. et al. (2021). “Non-linear response of $PM_{2.5}$ to changes in NO_x and NH_3 emissions in the Po Basin (Italy): consequences for air quality plans”. In: *Atmos. Chem. Phys.* 21(12), pp. 9309–9327. DOI: https://doi.org/10.5194/acp-21-9309-2021. URL: https://acp.copernicus.org/articles/21/9309/2021/.

Vecchia, Aldo V (1988). “Estimation and model identification for continuous spatial processes”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 50.2, pp. 297–312.

Veratti, Giorgio et al. (2023). “Impact of NO_x and NH_3 Emission Reduction on Particulate Matter across Po Valley: A LIFE-IP-PREPAIR Study”. In: *Atmosphere* 14.5, p. 762.

Viscarra Rossel, Raphael A, Richard Webster, and Darren Kidd (2014). “Mapping gamma radiation and its uncertainty from weathering products in a Tasmanian landscape with a proximal sensor and random forest kriging”. In: *Earth Surface Processes and Landforms* 39.6, pp. 735–748.

Wackernagel, H (2003). *Multivariate Geostatistics: An Introduction with Applications*. Springer.

Wall, Kent D and David S Stoffer (2002). “A state space approach to bootstrapping conditional forecasts in ARMA models”. In: *Journal of Time Series Analysis* 23.6, pp. 733–751.

Wang, Lei, Wei Wu, and Hong-Bin Liu (2019). “Digital mapping of topsoil pH by random forest with residual kriging (RFRK) in a hilly region”. In: *Soil Research* 57.4, pp. 387–396.

Wang, Yaqiong, Francesco Finazzi, and Alessandro Fassò (2021). “D-STEM v2: A Software for Modeling Functional Spatio-Temporal Data”. In: *Journal of Statistical Software* 99.10, pp. 1–29. DOI: 10.18637/jss.v099.i10. URL: https://www.jstatsoft.org/index.php/jss/article/view/v099i10.

Whittle, Peter (1954). “On stationary processes in the plane”. In: *Biometrika*, pp. 434–449.

— (1963). “Stochastic-processes in several dimensions”. In: *Bulletin of the International Statistical Institute* 40.2, pp. 974–994.

WHO (2021). “WHO global air quality guidelines: particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide: executive summary”. In.

Wikle, Christopher K, Andrew Zammit-Mangion, and Noel Cressie (2019). *Spatio-temporal statistics with R*. Chapman and Hall/CRC.

Williams, Christopher KI and Carl Edward Rasmussen (2006). *Gaussian processes for machine learning*. Vol. 2. 3. MIT press Cambridge, MA.

Wood, Simon N (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Xu, Ke and Christopher K Wikle (2007). “Estimation of parameterized spatio-temporal dynamic models”. In: *Journal of Statistical Planning and Inference* 137.2, pp. 567–588.

Yang, Yang et al. (2019). “Short-term and long-term exposures to fine particulate matter constituents and health: a systematic review and meta-analysis”. In: *Environmental pollution* 247, pp. 874–882.

Zammit-Mangion, Andrew and Noel Cressie (2021). “FRK: An R package for spatial and spatio-temporal prediction with large datasets”. In: *Journal of Statistical Software* 98, pp. 1–48.

Zhan, Yu et al. (2018). “Satellite-based estimates of daily NO2 exposure in China using hybrid random forest and spatiotemporal kriging model”. In: *Environmental science & technology* 52.7, pp. 4180–4189.

Zhang, Hao (2004). “Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics”. In: *Journal of the American Statistical Association* 99.465, pp. 250–261.

— (2007). “Maximum-likelihood estimation for multivariate spatial linear coregionalization models”. In: *Environmetrics: The official journal of the International Environmetrics Society* 18.2, pp. 125–139.

Zhang, Liyuan et al. (2017). “Impact of air humidity fluctuation on the rise of PM mass concentration based on the high-resolution monitoring data”. In: *Aerosol and Air Quality Research* 17.2, pp. 543–552.