

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 and for all his help and our many interesting discussions. His constant guidance, encouragement, and ability to reveal new perspectives have been invaluable throughout my academic journey. He shaped how I think about research.

I am also grateful to Francesco Finazzi. Our many discussions helped me solve technical problems and gave 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 want to thank Philipp Otto for his great ideas, enthusiasm, and help along the way, especially in connection with our joint work on the paper ‘Spatiotemporal Modelling of PM_{2.5} Concentrations in Lombardy (Italy): A Comparative Study’. My thanks also go to the fellow PhD student Lorenzo Tedesco for numerous comments on this thesis and, in particular, for our joint work ‘Multivariate Low-Rank State-Space Model with SPDE Approach for High-Dimensional Data’.

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.

I am excited to complete my PhD in such a beautiful year, the Jubilee year. 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	1
Chapter 1. Introduction	3
1.1 Motivating examples	5
1.1.1 Air pollution and livestock	5
1.1.2 Big dataset and multivariate analysis	9
1.2 Outline of the thesis	11
Chapter 2. Fundamentals of continuous-space and space-time processes	13
2.1 Univariate continuous-space processes	14
2.2 Multivariate continuous-space processes	18
2.3 Spatio-temporal processes	22
2.4 Statistical modelling of multivariate spatio-temporal processes	25
2.4.1 Univariate spatial modelling	28
2.4.2 Multivariate spatial modelling	33
2.4.3 Spatio-temporal modelling	34
Chapter 3. Fundamentals of discrete-space processes	39
3.1 Simultaneous autoregressive and conditional autoregressive	40
3.2 Gaussian Markov random field	42
3.2.1 SPDE approach	44
Chapter 4. Geostatistical dataset harmonisation: A dataset on livestock, meteorology and air quality in the Lombardy region, Italy	49
4.1 Overview	50
4.2 Source data description	51
4.3 Data harmonisation and processing	56

4.3.1	Validation of the harmonisation procedure	62
4.4	Gridded dataset	63
4.5	Data and code availability	67
 Chapter 5. Case study of analysis of livestock-related PM pollution based on spatio-temporal model		
		69
5.1	Preliminary analysis	70
5.2	Spatio-temporal modelling of $PM_{2.5}$ concentrations in Lombardy: A comparative study	73
5.2.1	Spatio-temporal statistical models and machine learning techniques . . .	75
5.2.2	Comparative study results	79
5.2.3	Model interpretation	84
5.3	Heteroskedastic spatio-temporal modelling of $PM_{2.5}$	88
5.3.1	The multivariate heteroskedastic hidden dynamic geostatistical model . .	90
5.3.2	Matrix representation of heteroskedastic HDGM and estimation formulas	92
5.4	Scenario analysis of livestock-related $PM_{2.5}$ pollution based on a heteroskedastic spatio-temporal model	96
5.4.1	Overview	97
5.4.2	Estimated model	97
5.4.3	Scenario analysis	100
5.4.4	Discussion of the results	103
5.5	Multivariate analysis	104
 Chapter 6. Multivariate low-rank state-space model with an SPDE approach for high-dimensional data		
		107
6.1	Big n problem and computational details	108
6.2	Efficient representations of the continuous-space Gaussian process	109
6.2.1	Likelihood approximations	109
6.2.2	Low-rank methods	111
6.2.3	Markov approximations	112
6.3	Multivariate state-space model framework	112
6.4	Multivariate low-rank SPDE state-space model	114
6.4.1	Identifiability, Convergence, and Approximation Bounds	116
6.5	Parameters estimation	120

6.5.1	Choice of the mesh	122
6.5.2	Approximation of the precision matrix	123
6.5.3	Asymptotic distribution and Wald confidence intervals	125
6.6	Simulations	126
6.6.1	Experimental design	126
6.6.2	Estimation procedure	128
6.6.3	Simulation results	128
6.7	Analysis of a multivariate spatio-temporal air quality dataset	129
6.7.1	Data description	130
6.7.2	Preliminary analysis	132
6.7.3	Application of the LR-SSM	134
6.8	Comparison with the INLA framework	139
6.9	Discussion	141
6.10	Software implementation	143
Chapter 7. Conclusions and future work		147
Conclusions		147
7.1	Main contributions of the thesis	147
7.2	Summary	151
Appendices		153
Appendix A. Appendix for Chapter 4		155
A.1	Data availability, abundance and standards	155
A.2	Validation of AQ variables	156
A.2.1	Validation of WE variables	159
A.3	Validation of EM and LI variables	160
A.3.1	Validation of LA variables	160
Appendix B. Appendix for Chapter 6		163
B.1	Notation	163
B.2	Proof of Theorem 1	164
B.3	Weak convergence of the low-rank approximation	166
B.3.1	Proof of Corollary 1	169
B.3.2	Proof of Theorem 2	169

B.4 Proof of Theorem 3	170
B.5 EM algorithm	171
B.6 Boundary effects	177
List of Figures	184
List of Tables	188
References	189

Abstract

This thesis proposes a framework based on geostatistical models for analysing 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’s 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 thus considered 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 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) a 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 and the estimation is naturally formulated within a state-space framework, where the unobserved state process can be inferred 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 states are assumed to be Gaussian Processes (GPs). 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), $PM_{2.5}$ and PM_{10} . 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.

However, with the modern set of larger spatial datasets, exact inference of GPs 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. Many 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 GPs inference to modern spatio-temporal datasets.

To deal with the large spatio-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 the fact that the Gaussian process, under general conditions, can be viewed as a solution to a certain stochastic partial differential equation. 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 the method’s accuracy and efficiency. 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 including 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), Banerjee et al. (2008), and Lawson (2018).

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 the ability to understand the underlying processes and 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 is based on Gaussian random fields, also known as Gaussian processes (GPs; C. E. 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 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 examples

1.1.1 Air pollution and livestock

Air pollution is 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 PM ($PM_{2.5}$), defined as a mixture of airborne particles with a diameter smaller than $2.5 \mu m$ (see also Jerrett et al., 2005, for a review), has been identified as the most hazardous air pollutant (European Environment 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 (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 (EEA, 2021). It was estimated that in 2018, around 90% of the world's population was exposed to polluted air (EEA, 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 (EEA, 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 $PM_{2.5}$ 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 depicts 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 PM during winter contains a substantial secondary component, often exceeding 50%. This proportion can be even 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

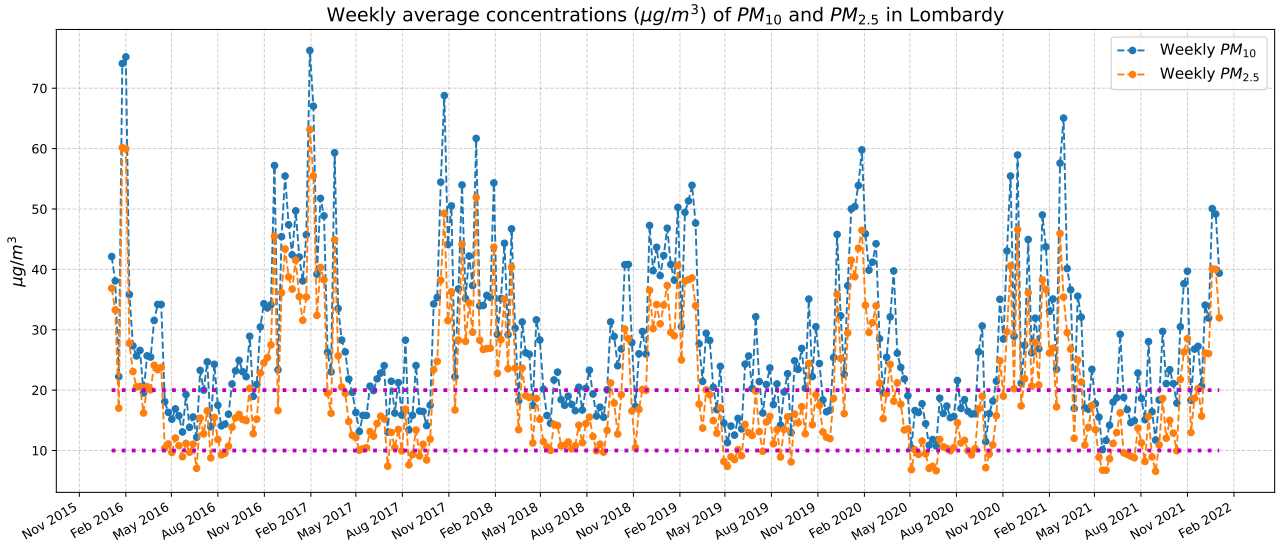


Figure 1: Time series of weekly averaged PM_{10} and $PM_{2.5}$ concentrations in the Lombardy region from January 2016 to December 2021. The weekly averages are computed from 105 stations for PM_{10} and 49 stations for $PM_{2.5}$. Pink horizontal lines indicate the WHO annual thresholds of $20 \mu\text{g}/\text{m}^3$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

limit the dispersion of pollutants; it is also characterised by weak ventilation, favouring the stagnation and accumulation of fine PM. 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 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 (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 total mass concentration of $PM_{2.5}$. Consequently, incorporating livestock data into air quality models can improve the understanding of, and inform strategies to mitigate, the impact of livestock on $PM_{2.5}$ concentrations.

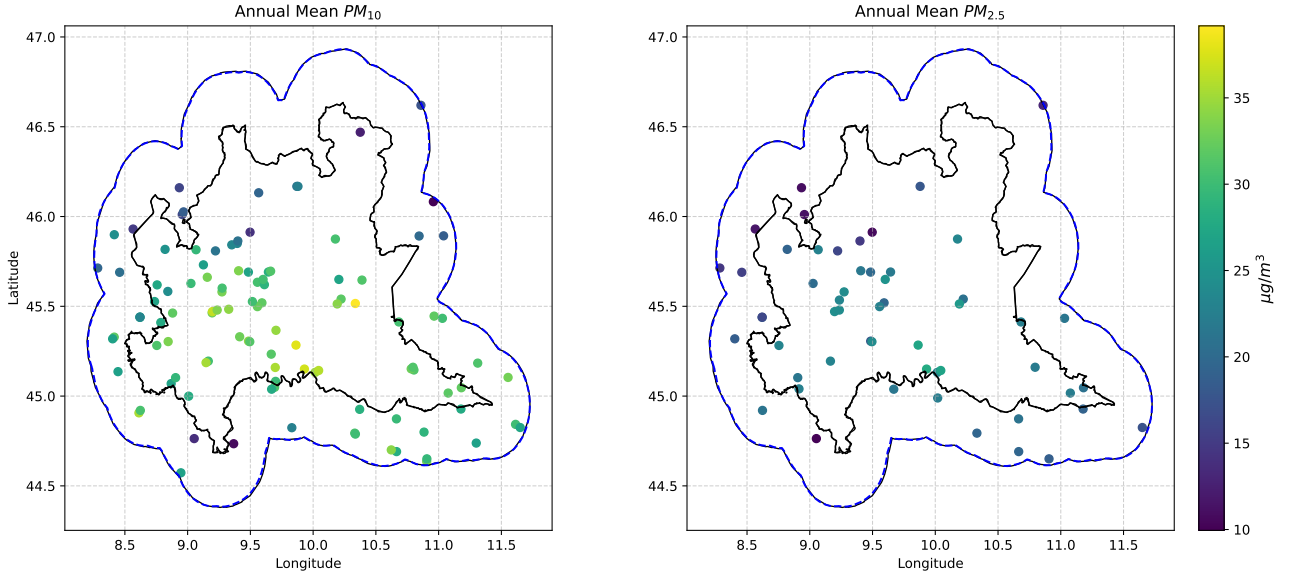


Figure 2: Average concentration of PM_{10} and $PM_{2.5}$ ($\mu\text{g}/\text{m}^3$) by 105 and 49 stations in the Lombardy region. The black line shows the Lombardy regional borders. While the blue solid line depicts the buffer of 0.3° that defines the neighbouring area, useful to avoid boundary effects.

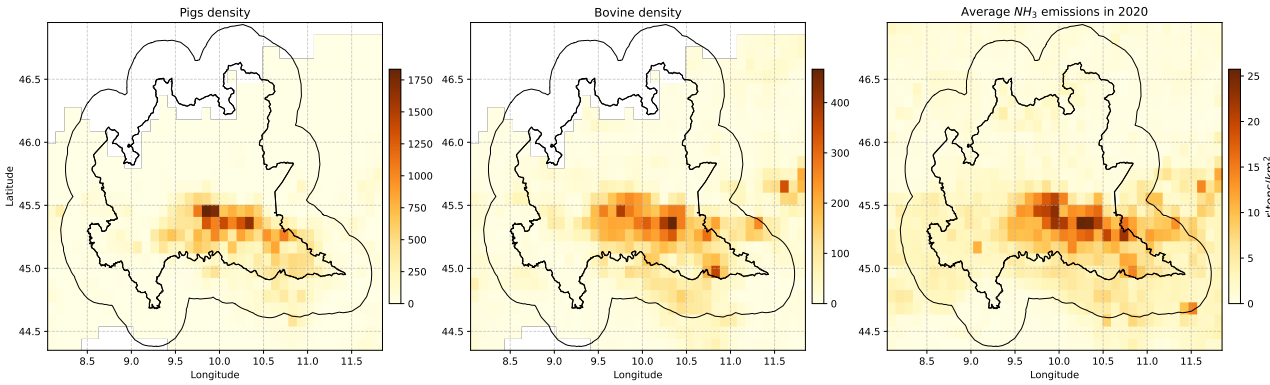


Figure 3: Density of pigs and bovines ($\text{number}/\text{km}^2$) in the left and centre panels, and ammonia (NH_3) emissions (t/km^2) over the Lombardy region in 2020 on the right. The black line indicates the augmented Lombardy region, obtained by applying a 0.3° buffer around the regional borders.

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 one 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?

The motivating example and the scientific questions raised above will be addressed in Chapters 4 and 5. In particular, Chapter 4 details the harmonisation process used to build the Agrimonia dataset, which integrates information on air quality, meteorology, emissions, livestock, and land and soil use at the same spatial and temporal resolution. In Chapter 5, we develop a spatio-temporal model suitable for studying and quantifying the relationship between ammonia and $PM_{2.5}$ concentrations. Finally, maps that show the sensitivity of $PM_{2.5}$ concentration to ammonia emission will be developed.

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 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. In Chapter 5 we will consider the multivariate analysis of PM_{10} and $PM_{2.5}$ to point out the role of the cross-dependence. 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 can all the sources be fused 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. Chapter 6 will discuss the ‘big n problem’, and introduce some of the approximations used to deal with computational constraints, including likelihood approximations, low-rank methods and the Markov approximations. Since the increasing dimensionality of spatio-temporal data makes exact Gaussian field inference computationally prohibitive, in Chapter 6 we will develop a low-rank approximation model based on the SPDE to achieve scalable inference for large multivariate spatio-temporal datasets, which is considered one of the main contributions of the thesis.

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 dimensions) 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 control the approximation error?
- How can one 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

This thesis is structured 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, Sections 2.1 and 2.2 present the definitions of stochastic processes for univariate and multivariate outcomes, respectively. The discussion of multivariate outcomes gives an understanding of the complexity of the multivariate modelling, and introduces the linear coregionalisation model, which will be used throughout the thesis. Section 2.3 introduces spatio-temporal processes, focusing on continuous-space and discrete-time formulations, and presents the discrete-time Markov process assumption, which is one of the fundamental assumptions underlying state-space models (SSM). Finally, Section 2.4 describes statistical models for data assumed to be generated by these processes, with a particular focus on the state-space modelling framework, which forms the core of the thesis.

Chapter 3 introduces spatial variation over a discrete-space domain modelling. Specifically, Section 3.1 introduces a range of modelling approaches that are particularly useful for analysing 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 central role in modern spatial statistics. A GMRF representation through the stochastic partial differential equation (SPDE) is then discussed in Section 3.2.1, and provides the central idea for developing the low-rank SSM presented in Chapter 6.

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 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 $PM_{2.5}$ concentrations, three classes of models, including machine learning and statistical approaches, are compared in Section 5.2. 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. The analysis is framed as a model-based “what-if” scenario rather than a formal counterfactual analysis, as causal identification assumptions are not explicitly enforced. 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 (SSM) 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 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 feature 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 $d = 1, 2, 3, \dots$, 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 for modelling 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 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 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 $\{z(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$. The distribution of the stochastic process $z(\mathbf{s})$ is defined by its finite-dimensional joint (multivariate) distributions

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

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

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

An important example of a stochastic process is the GP (Williams and C. E. Rasmussen (2006), see the Definition 2). Due to its flexibility and since the joint distribution is a multivariate normal and is 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 $\{z(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ is a Gaussian process if its finite-dimensional joint distribution is defined as a multivariate Gaussian distribution; that is, $F_{z(\cdot)}(y_1, \dots, y_n; \mathbf{s}_1, \dots, \mathbf{s}_n) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for every finite collection $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ 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 $\{\mathbf{s}_1, \dots, \mathbf{s}_n\} \subset \mathcal{D}$ to any other set $\{\mathbf{s}_1^*, \dots, \mathbf{s}_m^*\} \subseteq \mathcal{D}$.

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

$$F_{z(\cdot)}(y_1, \dots, y_n; \mathbf{s}_1 + \mathbf{h}, \dots, \mathbf{s}_n + \mathbf{h}) = F_{z(\cdot)}(y_1, \dots, y_n; \mathbf{s}_1, \dots, \mathbf{s}_n).$$

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 $\mu(\mathbf{s}) = E[z(\mathbf{s})]$ exist for all $\mathbf{s} \in \mathcal{D}$ and the $V[z(\mathbf{s})]$ is finite $\forall \mathbf{s} \in \mathcal{D}$; under these conditions, it is possible to 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 $\mathbf{h} \in \mathcal{D}$ and any locations $\mathbf{s}, \mathbf{s}' \in \mathcal{D}$ holds:

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

where $c(\cdot) : \mathcal{D} \rightarrow \mathbb{R}$ 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 $\mathbf{s}$ and $\mathbf{s}'$. It follows that a strictly stationary process with finite variance is also weakly stationary. The inverse is not true since a weakly stationary field is generally not strictly stationary.

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

When the covariance between two points depends only on their distance (i.e. $\text{Cov}[z(\mathbf{s}), z(\mathbf{s}')] = c(\|\mathbf{s} - \mathbf{s}'\|)$), 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

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

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 of the Matérn covariance function is largely 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, especially 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.

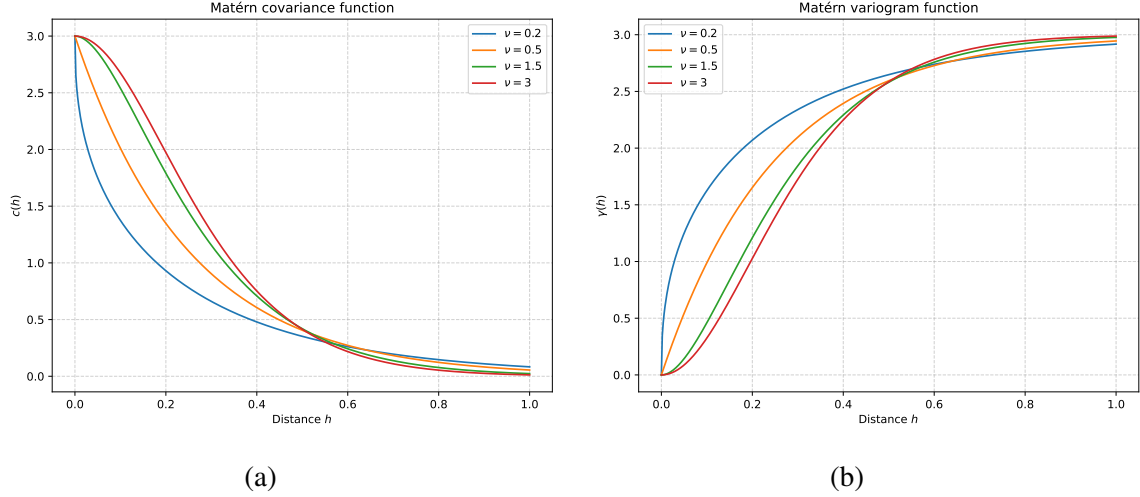


Figure 4: Matérn covariance in the left and the associated variogram functions in the right. $\nu = 0.5$ corresponds to the exponential covariance while larger values of ν yield smoother sample realisation.

Other classes of covariance 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 $z(\mathbf{s})$ and $z(\mathbf{s}')$ 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 $\nu \rightarrow +\infty$, 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 $z(\mathbf{s})$, spatial variation can be studied by considering the increments between values at pairs of points, $z(\mathbf{s} + \mathbf{h}) - z(\mathbf{s})$ for $\mathbf{s} + \mathbf{h}, \mathbf{s} \in \mathcal{D}$. First proposed by Georges Matheron (1963), the variogram $2\gamma(\mathbf{h})$ is defined as

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

If a process has a constant mean value function (if it exists) and a variogram only depending on $\mathbf{h}$, 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 $\gamma(\mathbf{h}) = c(0) - c(\mathbf{h})$, as shown in Figure 4b. In general, the reverse is not true since the variogram is not necessarily bounded. The correlation function $-1 \leq \rho(\mathbf{h}) \leq 1$ is given by $\rho(\mathbf{h}) = c(\mathbf{h})/c(0)$.

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

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

where $\mathcal{N}(\mathbf{h}) = \{(\mathbf{s}, \mathbf{s}') : \mathbf{s} - \mathbf{s}' = \mathbf{h}\}$ is the set of pair location at distance $\mathbf{h}$, and $|\mathcal{N}(\mathbf{h})|$ is the cardinality of the set. Although the estimator $\hat{\gamma}(\mathbf{h})$ 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):

$$\hat{\vartheta} = \arg \min_{\vartheta} \left\{ \sum_{j=1}^K |\mathcal{N}(\mathbf{h}(j))| \left(\frac{\hat{\gamma}(\mathbf{h}(j))}{\gamma(\mathbf{h}(j); \vartheta)} - 1 \right)^2 \right\}, \quad (2.4)$$

evaluated at some lags $\mathbf{h}(1), \dots, \mathbf{h}(K)$. $\hat{\gamma}(\mathbf{h}(j))$ is the method-of-moments estimator of $\gamma(\mathbf{h}(j))$. The fitting criterion Eq. (2.4) is sensible from the viewpoint that the more pairs of observations $|\mathcal{N}(\mathbf{h}(j))|$ there are, the more weight the residual at lag $\mathbf{h}(j)$ receives in the overall fit. Furthermore, 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. The minimisation Eq. (2.4) can be carried out using 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 $\mathbf{z}(\mathbf{s}) = (z_1(\mathbf{s}), \dots, z_p(\mathbf{s}))'$, $\mathbf{s} \in \mathcal{D}$, where $\{z_i(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ is the i -th process at location $\mathbf{s}$.

The definition of weak stationarity in the case of a multivariate process is straightforwardly obtained by extending Definition 4. Indeed, the process $\mathbf{z}(\mathbf{s})$ is said to be weak stationary if for all $\mathbf{s}, \mathbf{s}' \in \mathcal{D}$ holds $E[z_i(\mathbf{s})] = \mu_i$ and $\text{Cov}(z_i(\mathbf{s}), z_j(\mathbf{s}')) = c_{ij}(\mathbf{s} - \mathbf{s}')$, 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 $\mathbf{C}(\mathbf{s}, \mathbf{s}') = [c_{ij}(\mathbf{s} - \mathbf{s}')]_{i,j=1}^p$, where $c_{ij}(\cdot) : \mathcal{D} \rightarrow \mathbb{R}$ is a valid

covariance function and needs to be described to fully specify the multivariate random field. It is isotropic if $\text{Cov}(z_i(\mathbf{s}), z_j(\mathbf{s}')) = c_{ij}(\|\mathbf{s} - \mathbf{s}'\|)$. 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

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

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

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

That is, the quantity in Eq. (2.5) corresponds to the covariance of direct increments, while the quantity in Eq. (2.6) represents the variance of cross-increments. Cressie and Wile (1998) reviewed the differences between the two formulations and argued that Eq. (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 (Eq. 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 (Eq. 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 functions 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 challenging 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.

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

$$c_{ij}(\mathbf{s}, \mathbf{s}') = v_{ij}\rho(\mathbf{s}, \mathbf{s}'), \quad (2.7)$$

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

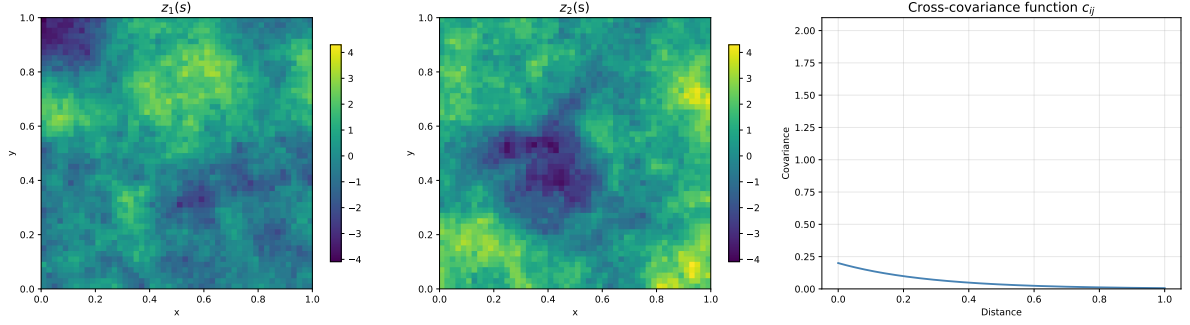


Figure 5: An example of the bivariate spatial process $\mathbf{z}(\mathbf{s}) = (z_1(\mathbf{s}), z_2(\mathbf{s}))$ defined on $\mathcal{D} = [0, 1]^2$, with an intrinsic multivariate covariance model $\mathbf{C}(\mathbf{h}) = \mathbf{V}\rho(\mathbf{h}; \vartheta)$, where $v_{ii} = 1$ and $v_{ij} = 0.2$ 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 $\rho(\mathbf{h}; \vartheta)$, parametrised by $\vartheta = \sigma^2 = 2$, $\nu = 0.5$, $k = 0.2$.

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 $\mathbf{C}(\mathbf{h})$ involves describing the relations between variables by the variance-covariance matrix $\mathbf{V}$ (sometimes called coregionalisation matrix) and the relations between points in space by a spatial correlation function $\rho(\mathbf{h}) : \mathcal{D} \rightarrow [-1, 1]$, such that $\mathbf{C}(\mathbf{h}) = \mathbf{V}\rho(\mathbf{h})$. 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 on spatial scale; they are intrinsically correlated by

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

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 $\mathbf{V}$ and a Matérn correlation function $\rho(\mathbf{h}; \vartheta)$. In Figure 5, $\mathbf{V}$ has $v_{ii} = 1$ and $v_{ij} = 0.2$, while in Figure 6, $\mathbf{V}$ 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.

The intrinsic correlation model implies an even cross-covariance function because $\rho(\mathbf{h})$ 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 $c_{ij}(\mathbf{h})$.

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

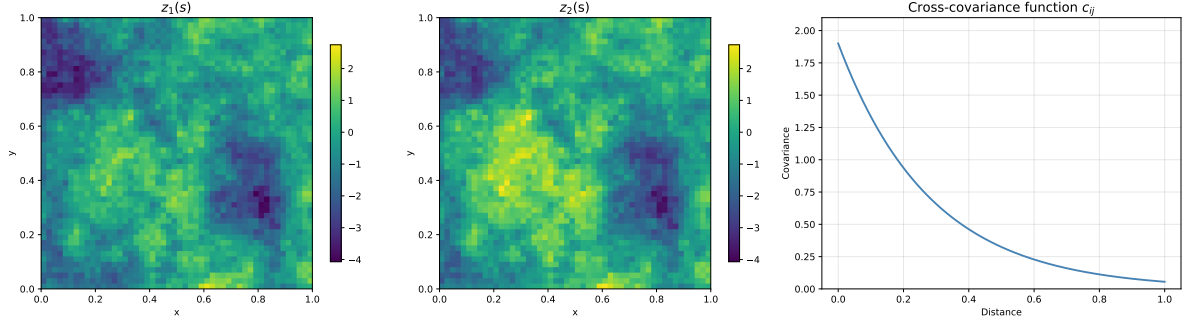


Figure 6: An example of the bivariate spatial process $\mathbf{z}(\mathbf{s}) = (z_1(\mathbf{s}), z_2(\mathbf{s}))$ defined on $\mathcal{D} = [0, 1]^2$, with an intrinsic multivariate covariance model $C(\mathbf{h}) = \mathbf{V}, \rho(\mathbf{h}; \vartheta)$, where $v_{ii} = 1$ and $v_{ij} = 1.9$ for $i, j = 1, 2$. 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 $\rho(\mathbf{h}; \vartheta)$, parametrised by $\vartheta = \sigma^2 = 2, \nu = 0.5, k = 0.2$.

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

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

where $c_{ij}(\mathbf{h}_k; \vartheta)$ is a valid parametric model for $c_{ij}(\mathbf{h}_k)$, and $\|\cdot\| : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^+$ 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. The minimisation is subject to the constraint that $\mathbf{V}$ 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 Eq. (2.5) and Eq. (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 involves representing the multivariate random field as a linear combination of $K > 0$ independent p -variate stationary processes, $\mathbf{w}_k$ for $k = 0, \dots, K$, with mean 0 and multivariate (separable) covariogram $\mathbf{V}_k \rho_k(\mathbf{h}, \vartheta_k)$, where $\rho_k(\mathbf{s}; \vartheta_k)$ is a valid correlation function and $\mathbf{V}_k$ is a $p \times p$ positive definite matrix which models the (non-spatial) covariance between the processes i and j . In addition, the $\mathbf{w}_k$ processes are spatially uncorrelated in the sense that for any $k \neq j$, $\text{Cov}[\mathbf{w}_k(\mathbf{s}), \mathbf{w}_j(\mathbf{s}')] = 0$ for all

$\mathbf{s}, \mathbf{s}' \in \mathcal{D}$. The LCM is defined as a linear mixture, that is:

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

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

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

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

With many processes, the number of parameters can quickly become unwieldy, and the resulting estimation difficult. H. Zhang (2007) described the MLE of the spatial LCM based on an EM algorithm. Importantly, H. Zhang (2007) showed 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. $z(\cdot) : \mathcal{D} \rightarrow \mathbb{R}$), as well as multivariate processes, $\mathbf{z}(\cdot) : \mathcal{D} \rightarrow \mathbb{R}^d$. However, in real applications, 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. Therefore, this section aims 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 $\{z(\mathbf{h}, t) : \mathbf{s} \in \mathcal{D}, t \in \mathcal{T} \subset \mathbb{R}^+\}$ 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

$$F_{z(\cdot, \cdot)}(y_1, \dots, y_n; (\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)) = P(z(\mathbf{s}_1, t_1) \leq y_1, \dots, z(\mathbf{s}_n, t_n) \leq y_n),$$

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

Definition 9 extends naturally to 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 $(\mathbf{h}, \delta) \in \mathcal{D} \times \mathcal{T}$ and for every collection $\{(\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)\} \subset \mathcal{D} \times \mathcal{T}$, the finite-dimensional joint distribution is such that

$$F_{z(\cdot, \cdot)}(y_1, \dots, y_n; (\mathbf{s}_1 + \mathbf{h}, t_1 + \delta), \dots, (\mathbf{s}_n + \mathbf{h}, t_n + \delta)) = F_{z(\cdot, \cdot)}(y_1, \dots, y_n; (\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)).$$

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 $(\mathbf{h}, \delta) \in \mathcal{D} \times \mathcal{T}$

$$E[z(\mathbf{s} + \mathbf{h}, t + \delta)] = E[z(\mathbf{s}, t)],$$

$$\text{Cov}[z(\mathbf{s} + \mathbf{h}, t + \delta), z(\mathbf{s}, t)] = \text{Cov}(z(\mathbf{0}, 0), z(\mathbf{h}, \delta)) = c(\mathbf{h}, \delta),$$

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

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

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

and

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

where $\vartheta = \{\vartheta_s, \vartheta_t\}$ and $c_s(\mathbf{s}; \vartheta_s)$, $c_t(t; \vartheta_t)$ 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 $g(\cdot) : \mathbb{R}^+ \rightarrow \mathbb{R}$ and a positive function with a completely monotone derivative $\psi(\cdot) : \mathbb{R}^+ \rightarrow \mathbb{R}$, then any function of the following form

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

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

However, spatio-temporal processes are often not observed continuously along time, but observed into ordered integer-spaced intervals such that $\mathcal{T} = \{0, 1, \dots, T\}$, where T is the temporal length. In this case, the problem reduces to the analysis of a collection of temporally varying spatial processes $\{z_t(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$ for $t \in \mathcal{T}$: meaning different continuous-space processes 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 $z(\mathbf{s}, t) = z_t(\mathbf{s})$. 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, $\text{Cov}(z(\mathbf{s}, t_i), z(\mathbf{s}, t_j)) = 0$ 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 $z(\mathbf{s})$ characterised only by the spatial covariance function $c_s(\mathbf{s}; \vartheta_s)$. Otherwise, the potential temporal dependence between the points belonging to different surfaces at subsequent times must be considered.

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; Calulli 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 an ordered sequence of random variables $(z_1, z_2, \dots, z_T)$ 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: $P(z_t \mid z_{t-1}, z_{t-2}, \dots, z_0) = P(z_t \mid z_{t-1})$ if both conditional probabilities are well defined; that is, if the joint probability $P_{z_1, \dots, z_T}(\cdot) > 0$. The countable set $\mathcal{S} = \{z_1, z_2, \dots, z_T\}$ 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. The Markov property is usually a fundamental assumption in dynamical models, as it implies conditional independence. This conditional independence structure allows the joint likelihood of the states and observations to be factorised into a product of low-dimensional densities. As a consequence, inference and parameter estimation become computationally tractable.

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) $\{z(\mathbf{s}, t) : \mathcal{D} \times \mathcal{T}\}$ indexed both in space and time (see Definition 9). For any given ω in the sample Ω , the mapping function $z(\mathbf{s}, t) : \mathcal{D} \times \mathcal{T} \rightarrow \mathbb{R}$ 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) provided a rigorous account of the mathematical theory underlying linear kriging. In this framework, spatial prediction corresponds to the case $g(z(\cdot)) = z(s_0)$, where s_0 is an unobserved location. The kriging predictor of $z(s_0)$ 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). The formal definition of BLUP is provided in Definition 13.

Definition 13. (BLUP) A predictor $\hat{z}(s_0) \in \Theta$ where Θ is the class of the linear predictor it 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. $E[\hat{z}(s_0)] = E[z(s_0)] = z(s_0)$ and $\hat{z}(s_0) = \arg \min_{\hat{z}(s_0) \in \Theta} E[(\hat{z}(s_0) - z(s_0))^2]$).

We have already introduced the GP, which is a common assumption in the analysis of spatio-temporal processes. GPs are fully specified by their mean function and covariance function $c(\cdot)$, which together characterise the behaviour of the process. Sections 2.1–2.3 describe the most commonly used covariance functions in spatial and spatio-temporal modelling. These are summarised in Table 1 for ease of reference. The categories describe complementary aspects of covariance modelling and should be interpreted as a taxonomy of properties and construction principles rather than as mutually exclusive classes. 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 $E(z(s_0)|z)$, where 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, multivariate spatial processes and 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 cokriging.

Category	Defining property	Examples and remarks
Stationary	$c(s, s') = c(h)$ with $h = s - s'$, $s, s' \in \mathbb{R}^d$	Includes isotropic and anisotropic stationary models. Common examples are the Matérn, exponential, and squared exponential covariance functions.
Non-stationary	$c(s, s')$ depends explicitly on both s and s'	Allows spatially varying variance, smoothness, or correlation range. Examples include spatial deformation models, kernel convolution approaches, and non-stationary Matérn models.
Isotropy / Anisotropy	Isotropic: $c(h) = c(\ h\)$; anisotropic: directional dependence	Anisotropy is commonly introduced via linear transformations of the spatial coordinates or geometric deformation of the domain.
Compactly supported	$c(h) = 0$ for $\ h\ > r$ for some $r > 0$	Examples include Wendland, spherical, and Bohman covariance functions; yield sparse covariance matrices and enable scalable inference for large datasets.
Long-range dependence	Polynomial or slow decay of $c(h)$ as $\ h\ \rightarrow \infty$	Fractional Matérn and power-law covariance models; useful for processes exhibiting persistent spatial dependence.
Separable spatio-temporal	$c((s, t), (s', t')) = c_s(s, s'; \theta_s) c_t(t, t'; \theta_t)$	Where $c_s(\cdot; \theta_s)$ and $c_t(\cdot; \theta_t)$ are valid spatial and temporal covariance functions, respectively (stationary or non-stationary). Separable models are computationally convenient but restrictive, as they cannot capture dynamic space–time interactions such as transport or advection effects.
Non-separable spatio-temporal	Joint dependence on space and time	Includes the Gneiting class and physically motivated covariance models; allows for realistic interactions between spatial and temporal dependence.
Periodic	$c(t, t') = c(t - t' \bmod p)$ for period $p > 0$	Periodic and quasi-periodic kernels; commonly combined with Matérn-type covariances to model seasonal behaviour.
Multivariate	Valid cross-covariance for vector-valued processes	Linear Model of Coregionalization (LMC) and Intrinsic Correlation Model; widely used for multivariate spatial and spatio-temporal data.
SPDE-based covariance	Covariance structure implied by a stochastic partial differential equation	Matérn fields constructed via SPDEs; yield Gaussian Markov random field (GMRF) representations with sparse precision matrices.

Table 1: Main classes of covariance functions and construction principles used in spatial and spatio-temporal modelling. The categories describe complementary properties and are not mutually exclusive.

2.4.1 Univariate spatial modelling

Let $\{z(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ be a real valued random process (see Definition 1). We are interested in making inferences on $z(\mathbf{s})$ given a finite number of observations $\mathbf{y} = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))'$. The set of observed locations $\mathcal{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$, with $\mathbf{s}_i \in \mathcal{D}$ for $i = 1, \dots, n$, is assumed known and fixed, as well as the domain $\mathcal{D}$. The measurements $\mathbf{y}$ 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):

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

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 a covariance function $c(\mathbf{h}) = c(\mathbf{h}; \vartheta)$ parametrised by ϑ . $\epsilon(\cdot)$ is an uncorrelated random noise, sometimes referred to as measurement error, with zero mean and marginal variance σ_ϵ^2 . $\epsilon(\cdot)$ and $z(\cdot)$ are uncorrelated; that is, $\text{Cov}(z(\cdot), \epsilon(\cdot)) = 0$. Under the SRE model (2.14), the process $y(\mathbf{s})$ is stationary with a constant mean. One possible generalisation of Eq. (2.14) is to assume that $E[z(\mathbf{s})] = \mu$ is no longer constant but is an unknown linear combination of a set of determinist covariates; that is $\mu(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}$, where $\mathbf{x}(\mathbf{s}) : \mathcal{D} \rightarrow \mathbb{R}^b$ 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 $y(\mathbf{s}) - \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}$ has zero mean by construction. Building upon these intuitions, the SRE model (2.14) can be expressed as

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

The parameter set of the model (2.15) is $\Pi = \{\boldsymbol{\beta}, \vartheta, \sigma_\epsilon^2\}$. 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. $\text{SNR} = \sigma^2/\sigma_\epsilon^2$). 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 $\mathbf{s}$. 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 $\mathbf{y}$ be the $n \times 1$

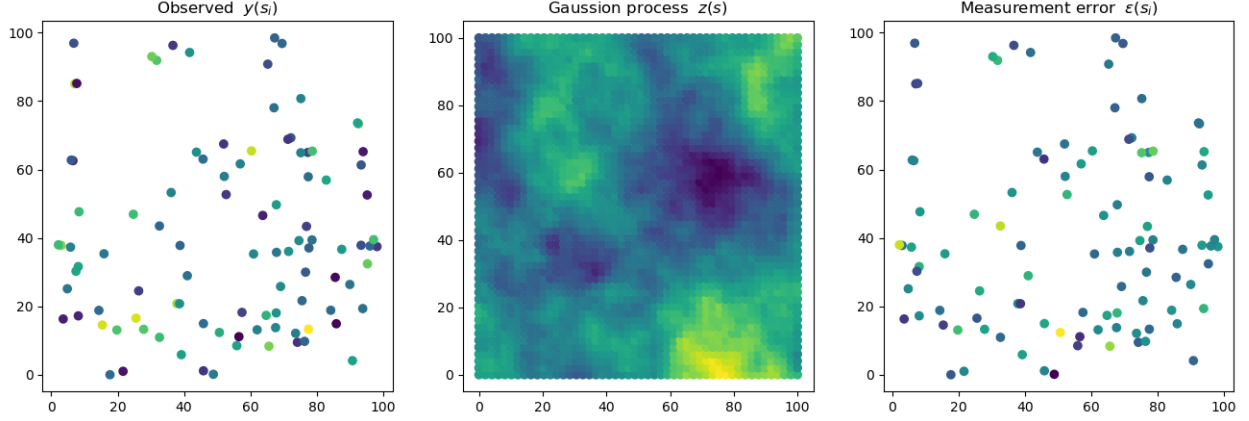


Figure 7: An example of the SRE data model (2.16). Left: the observed data $(y(s_1), \dots, y(s_n))$ with $n = 100$ and $\mathcal{D} = [0, 100]^2$. Centre: the GP $z(s)$ with $s \in \mathcal{D}$ and $c(\mathbf{h}; \vartheta)$ is the Matérn covariance function parametrised by $\vartheta = \{\sigma^2 = 1, \nu = 1, k = 40\}$. Right: the measurement error $(\epsilon(s_1), \dots, \epsilon(s_n))$.

vector of the observation and $\mathbf{z} = (z(s_1), \dots, z(s_n))'$ be the $n \times 1$ random vector. Since $z(s)$ is a GP, then $\mathbf{z} \sim N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (c(s_i - s_j; \vartheta))_{i,j=1}^n$. Finally, let $\boldsymbol{\epsilon} = (\epsilon(s_1), \dots, \epsilon(s_n))'$ be the corresponding noise vector. The SRE data model associated with the SRE model (2.15) becomes:

$$\mathbf{y} - \mathbf{X}'\boldsymbol{\beta} = \mathbf{z} + \boldsymbol{\epsilon}. \quad (2.16)$$

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 GP regression. Figure 7 shows an example of the SRE data model where $\mathbf{y} = (y(s_1), \dots, y(s_n))'$ with $n = 100$ and $\mathcal{D} = [0, 100]^2$, and $c(\mathbf{h}; \vartheta)$ 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. However, in this work, we focus on inference through the ML framework. If $\vartheta, \sigma_\epsilon^2$ are known, the ML estimation (MLE) of $\boldsymbol{\beta}$ 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. (*Generalised least square SRE estimation*) As an example, we illustrate here the GLS method to estimate the SRE model. Assume that the dataset $\{(y_i, \mathbf{s}_i)\}_{i=1}^n$ satisfies the SRE data model (2.16). Let $\mathbf{X}' = \text{stack}(x(\mathbf{s}_1)', \dots, x(\mathbf{s}_n)')$ be the $n \times b$ design matrix of the observed covariates. Assume that the random vector $\mathbf{z}$ is with $E[\mathbf{z}|\mathbf{X}] = 0$ and $\text{Cov}[\mathbf{z}|\mathbf{X}] = \Sigma(\vartheta)$, where $\Sigma(\vartheta) = (c(\mathbf{s}_i - \mathbf{s}_j; \vartheta))_{i,j=1}^n$ where ϑ is known. Assume that the measurement errors ϵ satisfy $E[\epsilon|\mathbf{X}] = 0$ and $\text{Cov}[\epsilon|\mathbf{X}] = \sigma_\epsilon^2 I_n$. The observed vector $\mathbf{y} = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))'$ then satisfies the SRE data model (2.16), with $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}'\beta$, where $\beta \in \mathbb{R}^b$ is a vector of unknown parameters. The GLS method estimates β by minimising the squared Mahalanobis distance of the residual vector $\mathbf{y} - \mathbf{X}'\beta$. The resulting estimator is

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

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

$$\begin{aligned} 2 \log L(\mathbf{y} | \mathbf{X}; \Pi) - n \log(2\pi) - \log |\Sigma(\vartheta) + \sigma_\epsilon^2 I_n| - (\mathbf{y} - \mathbf{X}'\beta)' (\Sigma(\vartheta) + \sigma_\epsilon^2 I_n)^{-1} (\mathbf{y} - \mathbf{X}'\beta), \\ \hat{\Pi}_{MLE} = \arg \max (\log L(\mathbf{y} | \mathbf{X}; \Pi)). \end{aligned} \quad (2.18)$$

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 in Eq. (2.4), while β^0 can be estimated using the GLS. Sometimes the log-likelihood is also written as $\log L(\mathbf{y} | \mathbf{X}; \Pi) = \log L_\Pi(\mathbf{y} | \mathbf{X})$ to emphasise the dependence on the distributional assumptions specified by Π .

Theoretical discussions on the MLE of the 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 MLE behaves as $n \rightarrow \infty$, 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) demonstrated 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 $\vartheta = \{\nu = 1/2, \sigma^2, 1/\theta\}$, they showed 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 the MLE of σ^2/θ not only displays reduced variance and more symmetric distributions but also retains 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 the 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 $\{(y_i, \mathbf{s}_i)\}_{i=1}^n$ satisfies the SRE data model (2.16) with a known parameter set $\Pi = \{\beta, \sigma_\epsilon^2, \vartheta\}$ and $\mathcal{S}$ the set of spatial indices. Let $\mathbf{z} \sim N(0, \Sigma(\vartheta))$, where $\Sigma(\vartheta) = [c(\mathbf{s}_i - \mathbf{s}_j; \vartheta)]_{i,j=1}^n$, and $c(\cdot; \vartheta)$ is a covariance function.

The measurement errors $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 I_n)$. Consider now a location set $\mathcal{S}^* = \{\mathbf{s}_0^*, \dots, \mathbf{s}_N^*\}$. Since $y(\cdot)$ is a GP, then $\mathbf{y}^* = \mathbf{y}(\mathcal{S}^*)$ and $\mathbf{y} = \mathbf{y}(\mathcal{S})$ are jointly normal, where $\mathbf{y}(\mathcal{S}) = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))'$. That is,

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

where the generic covariance matrix $\boldsymbol{\Sigma}_{(y,y^*)} = [\text{Cov}(y(\mathbf{s}_i), y(\mathbf{s}_j^*))]_{\mathbf{s}_i \in \mathcal{S}, \mathbf{s}_j^* \in \mathcal{S}^*} = [c(\mathbf{s}_i - \mathbf{s}_j^*; \boldsymbol{\theta})]_{\mathbf{s}_i \in \mathcal{S}, \mathbf{s}_j^* \in \mathcal{S}^*}$.

For the GP, the kriging predictor is the conditional expectation, $E(\mathbf{y}^* | \mathbf{y})$, i.e.

$$\begin{aligned} \mathbf{y}^* | \mathbf{y} &\sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}^* | \mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{y}^* | \mathbf{y}}) \\ \boldsymbol{\mu}_{\mathbf{y}^* | \mathbf{y}} &= \boldsymbol{\mu}_{\mathbf{y}^*} + \boldsymbol{\Sigma}_{\mathbf{y}^* \mathbf{y}} (\boldsymbol{\Sigma} + \sigma_\epsilon^2 I_n)^{-1} (\mathbf{y} - \boldsymbol{\mu}_y)' \\ \boldsymbol{\Sigma}_{\mathbf{y}^* | \mathbf{y}} &= \boldsymbol{\Sigma}_{\mathbf{y}^* \mathbf{y}^*} - \boldsymbol{\Sigma}_{\mathbf{y}^* \mathbf{y}} (\boldsymbol{\Sigma} + \sigma_\epsilon^2 I_n)^{-1} \boldsymbol{\Sigma}'_{\mathbf{y}^* \mathbf{y}} \end{aligned} \quad (2.20)$$

Note that for the locations in both sets (i.e. $\{\mathbf{s} : \mathbf{s} \in \mathcal{S} \cap \mathcal{S}^*\}$), 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 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 $\mathbf{X}'(\cdot)\boldsymbol{\beta}$ to $\mathbf{X}'(\cdot)\boldsymbol{\beta}(\cdot)$, where $\boldsymbol{\beta}(\cdot) : \mathcal{D} \rightarrow \mathbb{R}^b$ 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 known 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, $\mathcal{V}_R = \{\mathbf{v}_i\}_{i=1}^R \subset \mathcal{D}$, that span the domain. The finer the grid of the knots, the greater the local flexibility of the resulting function. Within this framework, the SRE model with spatially varying coefficients becomes

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

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

is the block diagonal matrix operator, and $\text{stack}(\cdot)$ is the stacking operator. The corresponding data model in Eq. (2.21) follows directly, and estimation may be performed within either the GLS or ML 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 $\mathbf{z}(\mathbf{s}) = (z_1(\mathbf{s}), \dots, z_q(\mathbf{s}))'$, $\mathbf{s} \in \mathcal{D}$ given a finite number of observations $\mathbf{y}$. 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 datasets. 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 $\mathcal{S}_1 = \dots = \mathcal{S}_p$, 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, which 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 pose a problem for inferring the cross variogram or covariance model. Indeed, the empirical cross-variogram in Eq. (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 $c_{ii}(\cdot)$ and the cross-covariance $c_{ij}(\cdot)$ such that the multivariate covariance $\mathbf{C}(\cdot) = \text{Cov}[z_1(\cdot), \dots, z_p(\cdot)] = [c_{ij}(\cdot; \boldsymbol{\vartheta})]_{i,j=1}^p$. Two popular choices of multivariate covariance are the intrinsic correlation model and the LCM 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,

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

where $\mathbf{y}(\mathbf{s}) = (y_1(\mathbf{s}), \dots, y_p(\mathbf{s}))'$ is the p -dimensional vector of observations, and $\mathbf{X}(\mathbf{s})\boldsymbol{\beta}$ is the deterministic mean structure with design matrix $\mathbf{X}(\mathbf{s}) = \text{blockdiag}(\mathbf{x}_1(\mathbf{s}), \dots, \mathbf{x}_p(\mathbf{s}))$, where $\mathbf{x}_i(\mathbf{s}) = (x_{i,1}(\mathbf{s}), \dots, x_{i,b_i}(\mathbf{s}))$ is a $(1 \times b_i)$ vector of the covariates for the i -process. The coeffi-

cient vector $\beta = \text{stack}(\beta_1, \dots, \beta_p)$, where each β_i , for $i = 1, \dots, p$, is a column vector of dimension b_i corresponding to the regression covariates associated with $y_i(s)$. W is a $p \times q$ loading matrix that linearly projects the latent random effects $z(s, t)$ into the observation space. $z(\cdot) = (z_1(\cdot), \dots, z_p(\cdot))'$ is a zero-mean, intrinsically stationary p -variate process with $\text{Cov}[z_1(\cdot), \dots, z_p(\cdot)] = C(\cdot)$. $\epsilon(\cdot) = (\epsilon_1, \dots, \epsilon_p)'$ is a p -variate measurement error with $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for all $i, j = 1, \dots, p$ and $\text{Cov}(z(\cdot), \epsilon(\cdot)) = 0$. From the model (2.22), the corresponding data model follows directly, and estimation may be performed within either the GLS or ML framework. If $z(s)$ is a GP, then the cokriging estimator is $E(y(s_0)|y)$.

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 $y(s, t) = (y_1(s, t), \dots, y_p(s, t))' \in \mathbb{R}^p$ be a p -dimensional spatio-temporal random field defined over a spatial domain $\mathcal{D}$ and discrete time steps $t = 1, \dots, T$, where $s \in \mathcal{D}$ indexes the spatial location. The multivariate SRE model (2.22) can be extended to account for the temporal correlations as,

$$y(s, t) - X'(s, t)\beta = Wz(s, t) + \epsilon(s, t), \quad (2.23)$$

where $X(s, t) = \text{blockdiag}(x_1(s, t), \dots, x_p(s, t))$ is a $b \times p$ design matrix, $x_i(s, t)$ is a column vector of dimension b_i associated with $y_i(s, t)$ for $i = 1, \dots, p$, and $b = \sum_{i=1}^p b_i$. The measurement errors are IID and $\epsilon \sim N(0, R)$. $\{z(s, t) : s \in \mathcal{D}, t \in \mathcal{T}\}$ 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 by taking the sum or the product of a purely spatial and a purely temporal covariance function, as in Eqs. (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 in Eq. (2.23), we impose the Markovian assumption on the latent process $\mathbf{z}(\mathbf{s}, t) = (z_1(\mathbf{s}, t), \dots, z_q(\mathbf{s}, t))$, which evolves over time according to the linear equation:

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

where $\mathbf{F}$ denotes a $q \times q$ transition matrix describing the system’s dynamics. The vector $\boldsymbol{\eta}(\mathbf{s}, t) = (\eta_1(\mathbf{s}, t), \dots, \eta_q(\mathbf{s}, t))$ represents a q -dimensional zero-mean GP, which is temporally independent but spatially correlated. Its multivariate covariance is given by $\text{Cov}(\boldsymbol{\eta}(\mathbf{s}, t)) = [c_{ij}(\mathbf{s} - \mathbf{s}'; \boldsymbol{\vartheta})]_{i,j=1}^q = \mathbf{Q}$. A popular choice of multivariate covariance is the LCM, see Section 2.2. The process starts from an initial state $\mathbf{z}(\mathbf{s}, 0)$, which is a GP with a 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 $\mathbf{z}_t(\mathbf{s}) = \mathbf{F}\mathbf{z}_{t-1}(\mathbf{s}) + \boldsymbol{\eta}_t(\mathbf{s})$, which essentially captures the dynamic transition from the state (i.e. the spatial process) at time $t-1$, $\mathbf{z}_{t-1}(\mathbf{s})$ to the state at time t , $\mathbf{z}_t(\mathbf{s})$.

Equation (2.23) is referred to as the measurement equation, while Eq. (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 SSM (Shumway, Stoffer, and Stoffer, 2000) with parameters:

$$\Pi = \{\boldsymbol{\beta}, \mathbf{F}, \mathbf{W}, \mathbf{Q}, \mathbf{R}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\}. \quad (2.25)$$

This formulation is particularly appealing due to its generality and flexibility, and can capture both observed and latent components of the time series. The model is called linear since the measurement and state equations depend linearly on the latent state. The matrices $\mathbf{W}$ and $\mathbf{F}$ specify the linear mappings from the state to the observations and from the previous state to the next state, respectively. Given the observed data $\mathbf{y}_{1:T} = (\mathbf{y}'_1, \dots, \mathbf{y}'_T)'$, where $\mathbf{y}_t = (y(\mathbf{s}_1, t), \dots, y(\mathbf{s}_m, t))'$, we wish to make inference on the random variable $\mathbf{z}(\mathbf{s}_0, t)$, regardless of whether there is a datum $\mathbf{z}(\mathbf{s}_0, t)$ available or not. Inference on the random variable $\mathbf{z}(\mathbf{s}_0, t')$, where $t' \in \{1, 2, \dots, T\}$, from data $\mathbf{y}_{1:t'}$ is known as filtering. Inference on the random variable $\mathbf{z}(\mathbf{s}_0, t')$, where $t' \in \{1, 2, \dots, T\}$, from data $\mathbf{y}_{1:T}$ is known as smoothing, and inference on the random variable $\mathbf{z}(\mathbf{s}_0, t')$, where $t' \in \{T+1, T+2, \dots\}$, from data $\mathbf{y}_{1:T}$ 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,

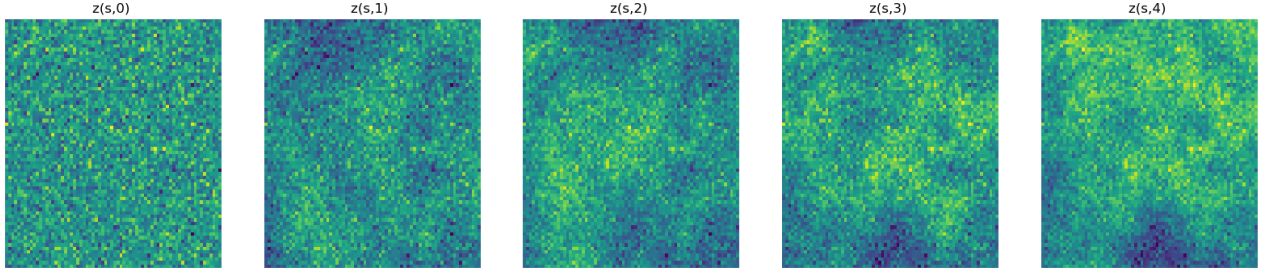


Figure 8: Simulation of the univariate state equation Eq. (2.24). Sample data $\{z(s_1, t), \dots, z(s_n, t)\}$ for $t \in \{0, \dots, 4\}$ are show from left to right. $z(s, 0) \sim N(0, I_n)$, $f = 0.2$. The innovation covariance matrix $\mathbf{Q}$ is given by the stationary Matérn function parametrised by $\vartheta = \{\sigma^2 = 1, \nu = 1, \theta = 0.5\}$.

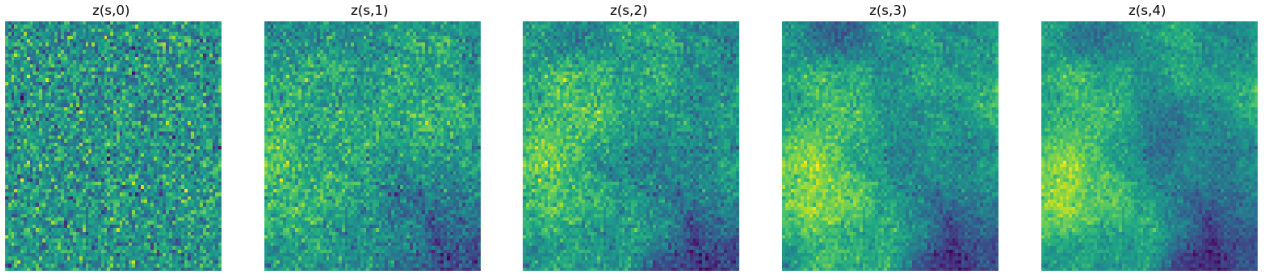


Figure 9: Simulation of the univariate state equation Eq. (2.24). Sample data $\{z(s_1, t), \dots, z(s_n, t)\}$ for $t \in \{0, \dots, 4\}$ are show from left to right. $z(s, 0) \sim N(0, I_n)$, $f = 0.9$. The innovation covariance matrix $\mathbf{Q}$ is given by the stationary Matérn function parametrised by $\vartheta = \{\sigma^2 = 1, \nu = 1, \theta = 0.5\}$.

the eigenvalues of the transition matrix $\mathbf{F}$ control the temporal correlation of the process: larger eigenvalues correspond to a stronger correlation between the current and past states. Figures 8 and 9 illustrate a univariate simulation of the Markovian process $z(s, t)$ for $t = 1, \dots, 4$ with $\mathbf{F} = 0.2$ and $\mathbf{F} = 0.9$, respectively. The initial state is set as $z(s, 0) \sim N(0, I_n)$, and the innovation covariance function is given by the stationary Matérn function. As expected, the process with $\mathbf{F} = 0.9$ 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 Calculi et al. (2015) for extensions to multivariate heterotopic data.

Given the parameter set Π , inference on the unobserved latent states $z(s, t)$ can be efficiently performed using the Kalman filter (KF) and the Kalman smoother (KS; Hamilton (2020)). Specifically, the

KS yields the state estimate $\mathbf{z}_t^T = \mathbb{E}_\Pi(\mathbf{z}_t | \mathbf{y}_{1:T})$ and the corresponding uncertainty $\mathbf{P}_t^T = \text{Var}_\Pi(\mathbf{z}_t | \mathbf{y}_{1:T})$. The notations $\mathbb{E}_\Pi(\cdot)$ and $\text{Var}_\Pi(\cdot)$ mean the expected value and the variance given the known set of parameters Π and hence the distribution from which the data are sampled. Since $\varepsilon(\mathbf{s}, t)$ and $\boldsymbol{\eta}(\mathbf{s}, t)$ are Gaussian and the dynamic system is linear, inference on $\mathbf{z}(\mathbf{s}, t)$ via the KF yields the best linear estimator in the minimum mean-square-error sense (Humpherys, Redd, and West, 2012). The optimality of the KF is limited to linear systems. More complex systems, however, may be nonlinear. Nonlinearity may arise in the measurement equation, the state equation, or both. The most common extensions of the KF for nonlinear systems are the Extended KF (Ribeiro, 2004) and the Unscented KF (Julier and Uhlmann, 2004). Regardless of Gaussianity, if the process and measurement covariances are known, the KF remains the best linear estimator in the minimum mean-square-error sense, although better nonlinear estimators may exist. While nonlinear extensions of the KF exist, this thesis focuses on the linear SSM.

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 $L_\Pi(\mathbf{y}, \mathbf{z})$ (or equivalently on the complete-data log-likelihood function $\log L_\Pi(\mathbf{y}, \mathbf{z})$, since the logarithm is monotone). At each iteration $k = 1, 2, \dots$, 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 observations $\mathbf{y}$,

$$Q(\Pi | \Pi^{(k)}) = \mathbb{E}_{\Pi^{(k)}} [\log L(\mathbf{y}, \mathbf{z} | \mathbf{y}, \mathbf{X}; \Pi)]. \quad (2.26)$$

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

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

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 $Q(\Pi | \Pi^{(k)})$ rather than directly improving the observed log-likelihood $\log L(\mathbf{y}; \Pi)$. 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 $\log L(\mathbf{y}; \Pi) - \log L(\mathbf{y}; \Pi^{(k)}) \geq Q(\Pi | \Pi^{(k)}) - Q(\Pi^{(k)} | \Pi^{(k)})$. In words, the maximisation at the M-step chose Π ($\Pi^{(k+1)}$) to improve $Q(\Pi | \Pi^{(k)})$, which in turn leads to an increase in the observed-data log-likelihood $\log L(\mathbf{y}; \Pi)$. As a result, the EM algorithm produces a sequence of parameter estimates that converges to a (local) maximum of the likelihood. At convergence $\Pi = \Pi_{\text{MLE}}$. 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) discussed modifications to the EM algorithm to accommodate fully restricted parameter matrices. However, it is not clear how they accounted 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 whether they give the ML estimates under the SSM 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 limits our choices for parameterisation. In general, the EM algorithm is useful for spatio-temporal SSMs 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 Chapter 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 this 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. 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 $\mathbf{s} \in \mathcal{D}$ is the collection of indices $N(\mathbf{s}) \subseteq \mathcal{D} \setminus \{\mathbf{s}\}$. The set $N(\mathbf{s})$ is the set of indices that are neighbours of index $\mathbf{s}$.*

Definition 14 defines a neighbourhood without explicitly addressing how to construct the set $N(\mathbf{s})$. 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. Given a lattice $\mathcal{D} \subset \mathbb{R}^2$, the first-order neighbourhood of a site $\mathbf{s} \in \mathcal{D}$, denoted $N_1(\mathbf{s})$, 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 $N_2(\mathbf{s})$, 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 $d : \mathcal{D} \rightarrow \mathbb{R}^+$ 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 Section 3.1, we introduce two fundamental models for discrete-space data: the simultaneous autoregressive (SAR) and conditional autoregressive (CAR) models.

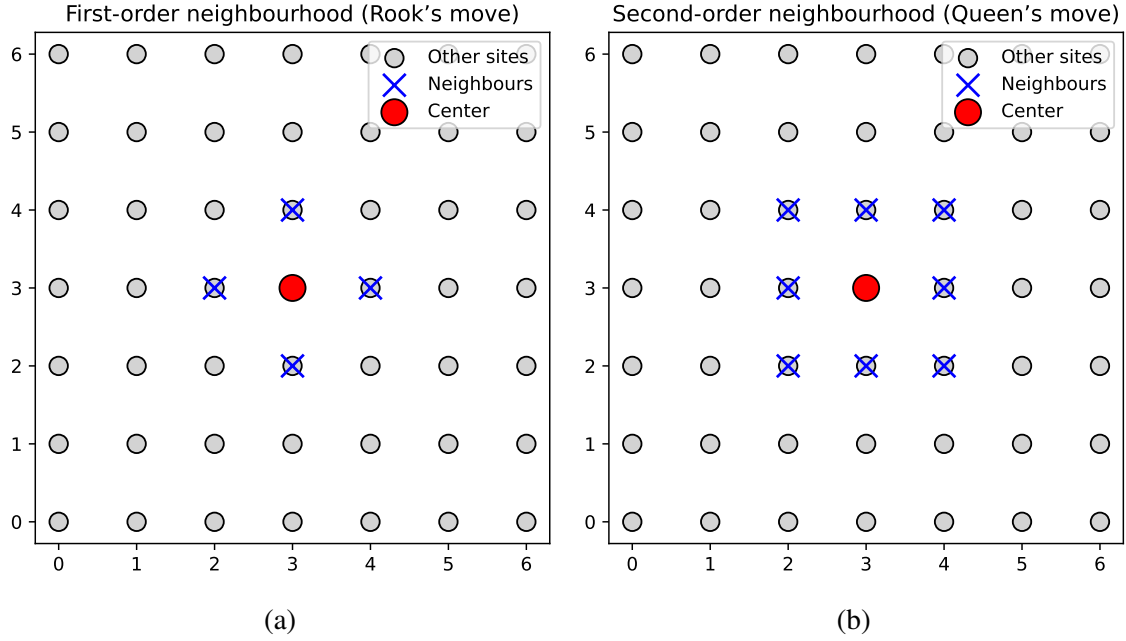


Figure 10: Neighbourhood structures for regular lattice data in $\mathbb{R}^2$: first-order neighbourhood (rook's move) on the left, and second-order neighbourhood (queen's move) on the right.

3.1 Simultaneous autoregressive and conditional autoregressive

In this section, we introduce the SAR and 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 $\mathcal{D} = \{(x_i, y_j) : i = 1, \dots, n\}$ be the discrete lattice.

We start by considering the SAR model, which was introduced by Whittle (1954). Suppose that the process $\{\epsilon(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ is a GP of IID random variables. Then, the SAR model defines the process $\{z(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ by specifying the linear system

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

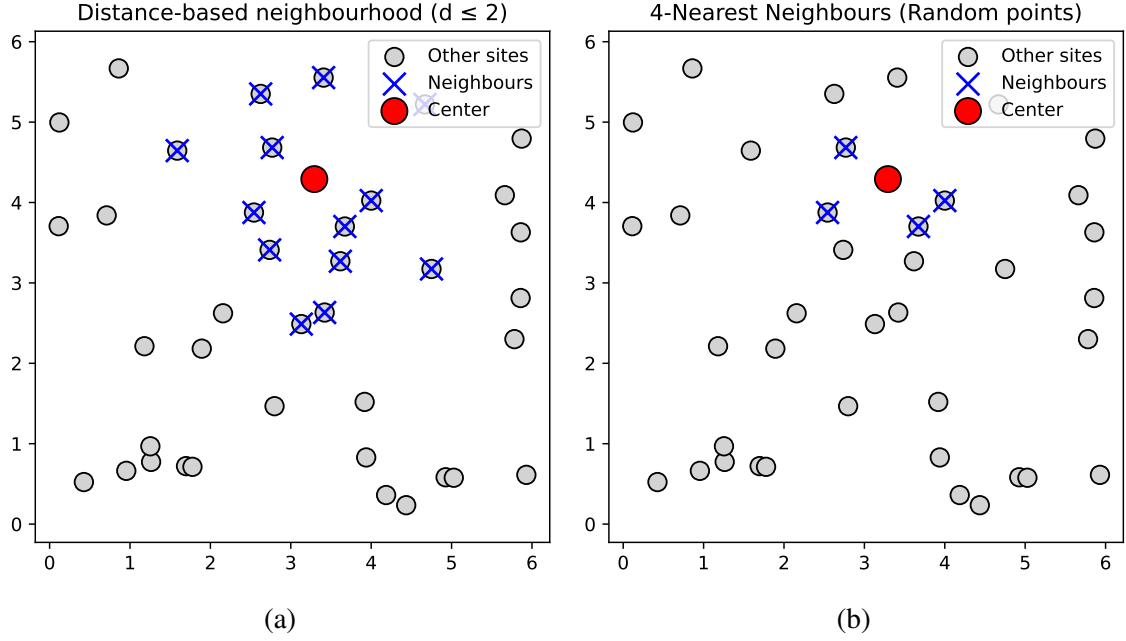


Figure 11: Neighbourhood structures for irregular lattice data in $\mathbb{R}^2$: distance-based neighbourhoods, on the left, and k-nearest neighbours on the right.

where $\mathbf{z} = (z(s_1), \dots, z(s_n))'$ is a realisation of the process $z(s)$ at the location set $\mathcal{S} = \{s_1, \dots, s_n\}$. $\mathbf{X}'\beta$ is the fixed effect term, $\epsilon \sim N(0, \Lambda)$ with diagonal covariance matrix, $\mathbf{B} = [b_{ij}]_{i,j=1}^n$ is a matrix of spatial dependence, such that $b_{ij} > 0$ if $z(s_i)$ is thought to depend positively with $z(s_j)$, and if $z(s_i)$ is thought to not depend on $z(s_j)$ then $b_{ij} = 0$. It is suppose that $b_{ii} = 0$ for $i = \{1, \dots, n\}$. The SAR model in Eq. (3.1) holds assuming $(\mathbf{I}_n - \mathbf{B})$ is non singular. Since $\mathbf{z}$ is a linear combination of ϵ , then $\mathbf{z} \sim N(\mathbf{X}'\beta, (\mathbf{I}_n - \mathbf{B})^{-1}\Lambda(\mathbf{I}_n - \mathbf{B}')^{-1})$. The model's parameter set is $\Pi = \{\beta, \text{diag}(\Lambda), \mathbf{B}\}$ and can be estimated through the MLE, as shown in Example 4.

Example 4. (SAR model estimation) This example presents the MLE for the model in Eq. (3.1) and assuming a parametric form for $\mathbf{B}$ given by $\mathbf{B} = \lambda\mathbf{W}$ and $\epsilon \sim N(0, \sigma_\epsilon^2 \mathbf{I}_n)$. The matrix $\mathbf{W} = [w_{ij}]$ 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 $(\mathbf{I}_n - \lambda\mathbf{W})(\mathbf{z} - \mathbf{X}\beta) = \epsilon$. The parameter set $\Pi = \{\beta, \lambda, \sigma_\epsilon^2\}$ can then be estimated using the ML methods.

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

$$P(z(s) | \{z(u) : u \in \mathcal{D}\}) = P(z(s) | \{z(u) : u \in \mathcal{N}(s)\}). \quad (3.2)$$

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

$$z(\mathbf{s}) \perp \{z(\mathbf{u}) : \mathbf{u} \in \mathcal{D} \setminus \{\mathbf{s}\}\} | \{z(\mathbf{u}) : \mathbf{u} \in \mathcal{N}(\mathbf{s})\}, \quad (3.3)$$

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

One reason for the popularity of 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 had been used for both lattices and spatially motivated graphs, they were only spatially consistent for regular lattices, which limited their applicability. However, Finn Lindgren, Håvard Rue, and Lindström (2011) derived the CAR models in a new way that 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

The Gaussian Markov random fields (GMRFs; see Håvard Rue and Held (2005) for extensive details) arise when CAR models are defined for GPs. 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. In Chapter 6, we leverage the GMRF representation of a continuous-space GP to develop a computationally-efficient method suitable for high-dimensional datasets.

The Markov property in Eq. (3.3) implies that if $(z(\mathbf{s}_1), \dots, z(\mathbf{s}_n))' \sim N(0, \mathbf{Q}^{-1})$, 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 computational gain comes from the fact that the zero pattern of the precision matrix $\mathbf{Q}$ relates directly to the notion of neighbours. The neighbourhood of $\mathbf{s}_i$ typically consists of all points that, in some sense, are close to $\mathbf{s}_i$. In theory, there are no restrictions on the size of the neighbourhood, and one could, for example, have $\mathcal{N}(\mathbf{s}_i) = \mathcal{D} \setminus \mathbf{s}_i$,

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. GMRFs have very good computational properties, which are of major importance in inferential methods (Havard Rue and Held, 2005).

Although GMRFs have very good computational properties, there are reasons why current statistical models based on GMRFs are relatively simple, in particular when applied to area data from regions or counties. First, there has been no good way to parameterise the precision matrix of a GMRF to achieve predefined behaviour in terms of correlation between two sites and to control marginal variances. Finn 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 CAR models, removing the lattice constraint and allowing for non-stationary extensions. However, the most important implication of the work in Finn Lindgren, Håvard Rue, and Lindström (2011) is that it provides a spatially consistent method for approximating continuous-space GPs using GMRFs, which enables the use of sparse matrix techniques for GMRF inference (see Figure 12). The results apply to GPs for which a GMRF with a local neighbourhood structure provides a sufficiently accurate representation, thereby preserving the interpretability of the parameters and the validity of the inferential results.

In particular, Finn Lindgren, Håvard Rue, and Lindström (2011) established an explicit link between GPs with Matérn covariance functions defined on $\mathbb{R}^d$ and their corresponding GMRF representations. The goal is to approximate a GP with a covariance matrix Σ using a GMRF with a 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 a Matérn GP. The GMRF representation of a continuous-space GP will be used in Chapter 6 to develop a scalable spatio-temporal low-rank model, where we apply the SPDE method within the SSM framework. An important observation is that the resulting modelling strategy does not involve having to construct explicit formulae for the covariance functions, which are instead only defined implicitly through the SPDE. Additional details on the GMRF representation through the SPDE approach are provided in the next section.

Example 5. (*Advantages of sparse matrix*) In Gaussian models, likelihood evaluation involves the inversion of the covariance matrix, which has a complexity of $O(n^3)$, see Example 2. As a result,

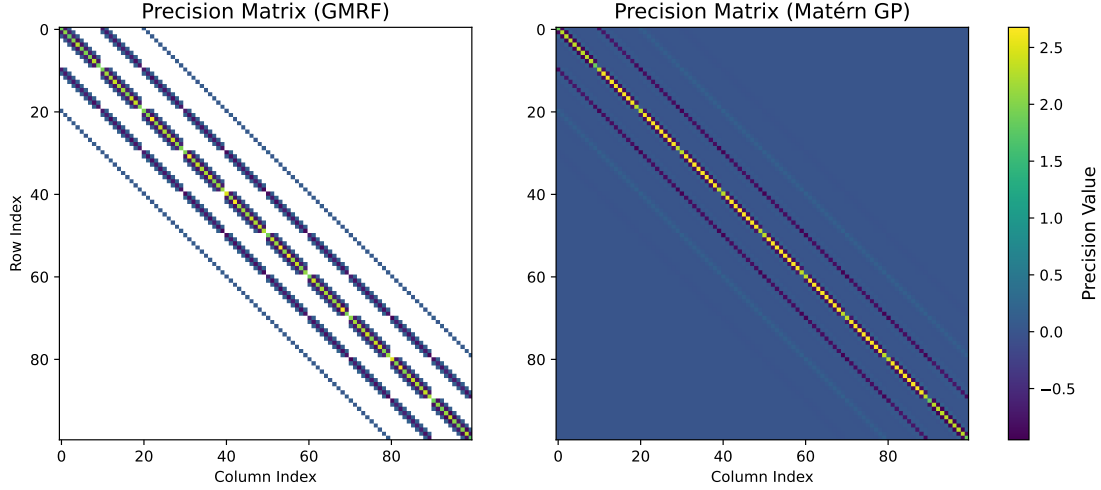


Figure 12: Comparison of precision matrices from two spatial models. Left: Gaussian Markov Random Field (GMRF) precision matrix obtained from the SPDE discretisation. Right: Matérn Gaussian Process (GP) precision matrix. Zero entries are shown in white, highlighting the sparsity pattern of the GMRF compared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

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.

3.2.1 SPDE approach

As introduced above, Finn Lindgren, Håvard Rue, and Lindström (2011) shows that, by using an approximate stochastic weak solution to (linear) stochastic partial differential equations, it is possible, for certain Gaussian fields in the Matérn class, to establish an explicit link between GPs and GMRFs for any triangulation of $\mathbb{R}^d$, formulated through a basis-function representation. The key consequence is that modelling can be carried out using GPs while computations are performed using GMRFs, thereby combining the strengths of both approaches.

Whittle (1954) shows that, a one-dimensional GP $z(s)$ with Matérn covariance parametrised by $\{k, \sigma^2, \nu\}$, see Eq. (2.1), is a stationary solution to the linear SPDE:

$$(\kappa^2 - \Delta)^{\alpha/2} x(s) = \mathcal{W}(s), \quad s \in \mathcal{D} \subset \mathbb{R}^d, \quad \kappa > 0, \quad (3.4)$$

where $\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$ is the Laplacian operator and $\alpha = \nu + d/2$, $\kappa > 0$ is a scaling parameter, and ν control the smoothness of the field. $\mathcal{W}$ denotes spatial Gaussian white noise with unit variance, and

$x(\mathbf{s})$ has marginal variance:

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

According to Whittle (1954), the random field is Markovian when $\alpha = \nu + d/2$ is an integer. When d is even, we get Matérn models with $\nu \in \mathbb{N}$. On the other hand, when d is odd, we get Matérn models with $\nu \in \frac{1}{2}\mathbb{N}$.

To construct a GMRF representation of the Matérn GP, we begin by partitioning the domain $\mathcal{D} \subset \mathbb{R}^2$ into a set of non-overlapping triangular elements (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 $\mathcal{V}_R = \{\mathbf{r}_1, \dots, \mathbf{r}_R\}$ and $\mathcal{E}_R$ being the set of the edges of the triangles. A special triangulation is the so-called Delaunay triangulation (Hjelle and Dæhlen, 2006), which maximises the minimum interior angles of the mesh elements, ensuring smooth transitions between regions of different mesh resolution. The triangulation procedure results in $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is a set of vertices with cardinality $R < \infty$, and $\mathcal{E}$ is the set of edges.

To obtain a numerical solution of Eq. (3.4), the latent field is approximated by a finite expansion, known as the finite element (FE) representation, in terms of predefined basis functions with Gaussian-distributed weights. Specifically,

$$z(\mathbf{s}) = \sum_{k=1}^R \psi_k(\mathbf{s})w_k = \boldsymbol{\psi}'_R(\mathbf{s})\mathbf{w}, \quad (3.5)$$

where $\boldsymbol{\psi}_R(\mathbf{s}) = (\psi_1(\mathbf{s}), \dots, \psi_R(\mathbf{s}))'$ are piecewise linear basis functions, and $\mathbf{w}$ are Gaussian weights. The basis functions are defined such that $\psi_k(\mathbf{s})$ takes a value of 1 at node $\mathbf{r}_k$ and 0 at all other nodes, thereby preserving the partition of the unit overall the domain $\mathcal{D}$. Given this construction of the basis functions, the weights w_r can be interpreted as the values of the field at the mesh nodes, while values within the interior of each triangle are obtained by linear interpolation. Consequently, Eq. (3.5) yields a continuously indexed but finite-dimensional approximation to the solution of the SPDE.

The solution to the SPDE in Eq (3.4) is obtained by satisfying the stochastic weak formulation of the SPDE given by

$$\langle \phi_j, (\kappa^2 - \Delta)^{\alpha/2} x \rangle \stackrel{d}{=} \langle \phi_j, \mathcal{W} \rangle, \quad \forall j = 1, \dots, m, \quad (3.6)$$

for suitable test functions $\{\phi_j\}_{j=1}^m$. The inner product is defined as

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

The finite-dimensional solution is obtained by substituting the FE representation in Eq. (3.5) into Eq. (3.6) and by deriving the distribution of the weights that satisfy the resulting weak formulation for a

chosen set of test functions, with $m = R$. For $\alpha = 1$, the test functions are chosen as $\phi_k = (k^2 - \Delta)^{1/2} \psi_k$, whereas for $\alpha = 2$, the test functions are $\phi_k = \psi_k$. These choices correspond to the least-squares and Galerkin formulations, respectively. By replacing the FE representation in Eq. (3.5) into the stochastic weak formulation of the SPDE in Eq. (3.6), and considering for simplicity the case $\alpha = 2$, we obtain the following system of linear equations:

$$\begin{aligned} \langle \psi_j, (\kappa^2 - \Delta) \sum_{k=1}^R \psi_k w_k \rangle &\stackrel{d}{=} \langle \psi_j, \mathcal{W} \rangle, \quad \forall j = 1, \dots, R \\ \sum_{k=1}^R w_k \langle \psi_j, (\kappa^2 - \Delta) \psi_k \rangle &\stackrel{d}{=} \langle \psi_j, \mathcal{W} \rangle \\ \sum_{k=1}^R w_k [\kappa^2 \langle \psi_j, \psi_k \rangle - \langle \psi_j, \Delta \psi_k \rangle] &\stackrel{d}{=} \langle \psi_j, \mathcal{W} \rangle \end{aligned} \quad (3.7)$$

Applying Green's first identity, the inner product $\langle \phi_j, -\Delta \psi_k \rangle$ can be expressed as

$$\langle \phi_j, -\Delta \psi_k \rangle = \langle \nabla \phi_j, \nabla \psi_k \rangle - \langle \phi_j, \partial_n \psi_k \rangle_{\partial \mathcal{D}},$$

where $\partial_n \psi(s) = \mathbf{n}(s) \nabla \psi(s)$ is the normal derivative of a function $\psi(\cdot)$ on the boundary $\partial \mathcal{D}$ and $\mathbf{n}(s)$ denotes the unit length outward normal vector at the point s on the boundary. Homogeneous Neumann boundary conditions are imposed by selecting basis functions with zero normal derivatives on $\partial \mathcal{D}$. Under these conditions, the boundary term vanishes. Define the mass and stiffness matrices as

$$\begin{aligned} C_{jk} &= \langle \psi_j, \psi_k \rangle, \\ G_{jk} &= \langle \nabla \psi_j, \nabla \psi_k \rangle. \end{aligned}$$

Then, for $\alpha = 2$, the finite-dimensional representation of the SPDE leads to the linear system

$$\sum_{k=1}^R w_k [\kappa^2 C_{jk} - G_{jk}] \stackrel{d}{=} \epsilon_j, \quad \forall j = 1, \dots, R, \quad (3.8)$$

where ϵ_j is the projection of the white noise of the base ψ_j . Since the $\mathcal{W}$ is a white noise, then $(\epsilon_1, \dots, \epsilon_R) \sim \mathcal{N}(\cdot, \cdot)$ where $\mathbb{E}[\epsilon_j] = 0$ and $\text{Cov}[\epsilon_j, \epsilon_k] = \langle \phi_j, \phi_k \rangle = C_{jk}$. Using matrix notation, the linear system Eq. (3.8) becomes

$$(\kappa^2 \mathbf{C} + \mathbf{G}) \mathbf{w} = \boldsymbol{\epsilon}. \quad (3.9)$$

It follows that $\mathbf{w} \sim \mathcal{N}(0, \mathbf{K}_\kappa^{-1} \mathbf{C}^{-1} \mathbf{K}_\kappa^{-1})$ where $\mathbf{K}_\kappa = (\kappa^2 \mathbf{C} + \mathbf{G})$ or equivalently the precision matrix for $\mathbf{w}$ is:

$$\mathbf{Q}_\kappa = \mathbf{K}_\kappa \mathbf{C}^{-1} \mathbf{K}_\kappa. \quad (3.10)$$

While C^{-1} is dense, one can replace C with a diagonal matrix $\tilde{C}$, where $\tilde{C}_{ii} = \langle \psi_i, 1 \rangle$, to obtain a sparse precision Q_κ (Finn Lindgren, Håvard Rue, and Lindström, 2011). Asymptotic results of this approximate GMRF representation of the Matérn field are in Finn Lindgren, Håvard Rue, and Lindström (2011).

Since the solution to the SPDE is a GMRF, the sparsity in Q_{α, κ^2} provides great computational benefits, see for instance Figure 12 and Example 5. In Chapter 6, we leverage the computational advantages of the GMRF representation of a continuous-space Gaussian process within the SSM framework. Specifically, the continuous-index GP for the innovation term $\eta(s, t)$ (see Eq. (2.24)) is approximated using the finite element representation, resulting in a low-rank formulation of the SSM.

From a traditional point of view, the most surprising extension within the SPDE framework is how we can model non-stationarity. Many applications require non-stationarity. The SPDE approach has the additional huge advantage that the resulting (non-stationary) GP is a GMRF, which allows for swift computations and can additionally be defined on a manifold. In the SPDE defined in Eq. (3.6), the parameters κ^2 and the innovation variance are constant in space. In general, it is possible to allow both parameters to depend on the coordinate s (Finn Lindgren, Håvard Rue, and Lindström, 2011).

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

To support studies quantifying the effect of the agricultural and livestock sectors on the Lombardy air quality, this chapter presents the development of a spatio-temporal dataset for 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 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 rest of this chapter is organised as follows. Section 4.1 provides an overview of explanatory variables useful to model 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. The dataset details are reported 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 PM_{2.5} concentrations (L. Zhang et al., 2017).

We consider 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 km²). Including livestock data is essential to capture the impact of ammonia (NH₃) emissions on air quality, as livestock farming is the major source of NH₃ emissions (up to 95%). Therefore, including livestock 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 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 2 and are described in the following subsections, with details on spatio-temporal resolution and data availability.

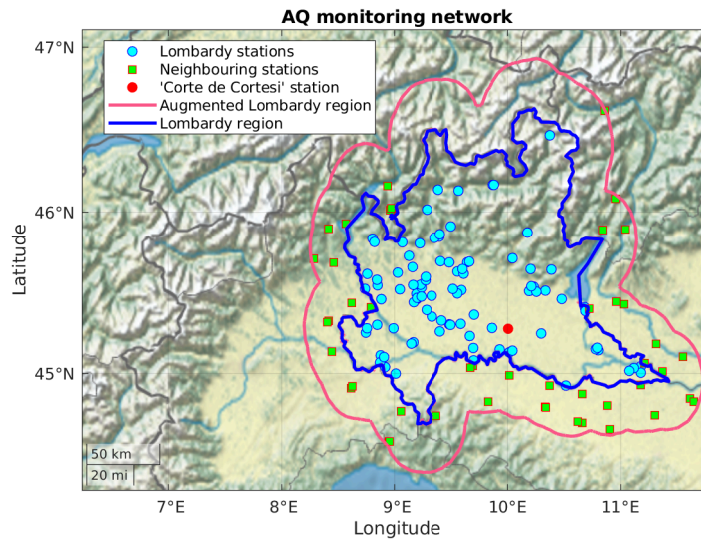


Figure 13: AQ network of $S = 141$ stations in the augmented Lombardy region (pink boundaries): $S_1 = 93$ stations are inside the Lombardy region (blue boundaries) and $S_2 = 48$ in the 0.3° buffer defining the neighbouring area between blue and pink boundaries. The station named ‘Corte de Cortesi’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

Air quality

Air quality data are pollutant concentrations ($\mu\text{g}/\text{m}^3$) sampled at $S = 141$ ground-level monitoring stations, irregularly located over the augmented Lombardy region, as shown in Figure 13. For the

Dimension	Source	Spatial coverage	Temporal resolution
Air quality (AQ)	EEA – Air pollution section EEA (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 2: Input sources dataset used to build the Agrimonia dataset.

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 EEA from another 48 stations. Inside Lombardy, the variables listed in Table 3, 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 come from the EEA’s open-access service. Table 3 summarises the pollutants selected, their sources, and the number of sensors available for each pollutant.

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

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 3: 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 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. The 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 4 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):

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

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)	<i>Categ.</i>

Table 4: 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 5 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 5: EM variables selected from the CAMS-anthropogenic emissions dataset. Spatial resolution of (0.1×0.1) degrees. Unit is $kg/(m^2s)$.

Livestock

The Agrimonia dataset focuses on data on pigs and bovines because of their role in NH₃ emissions. Indeed, for Lombardy, about 87% of the annual NH₃ 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.

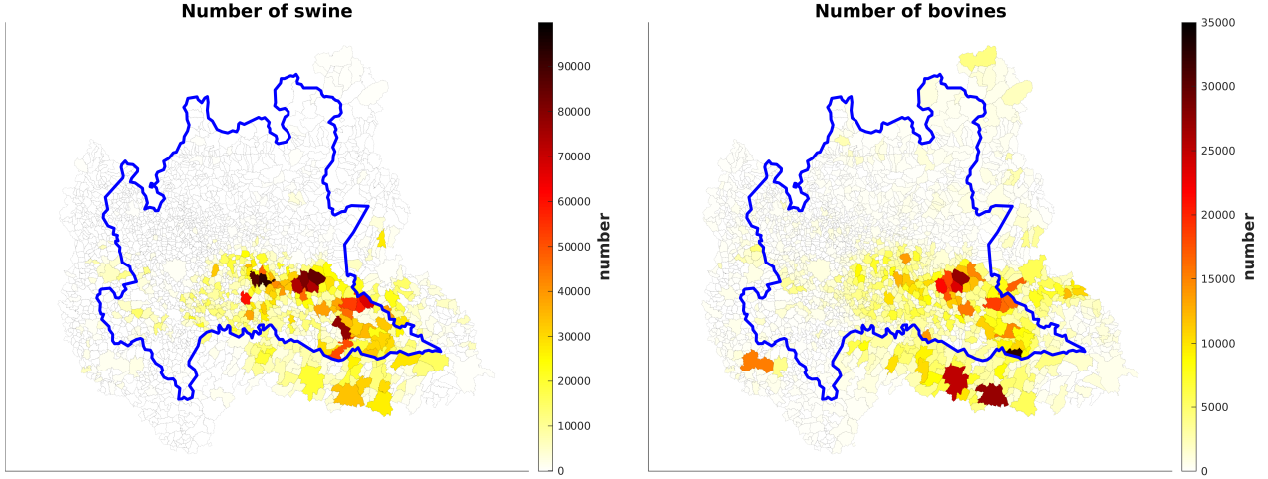


Figure 14: Number of swine (left) and bovines (right) in the augmented Lombardy region and neighbouring area, aggregated at the municipal level on 31 December 2021.

Land

The land cover and 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 6. 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 how humans manage and utilise these surfaces, for example, for agricultural, residential, or industrial purposes. In addition, soil use provides complementary information on 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. Here, we introduce 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	<i>Categ.</i>
SIARL (soil use)	Soil cultivation type	Classification of soil use with 21 classes	<i>Categ.</i>

Table 6: LA variables selected from the ERA5, CLC and SIARL datasets. *Categ.* stands for Categorical.

Air quality

The AQ data, from the 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 3).

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 Algorithm 1. For each time series, missing values are then imputed using an SSM (Durbin and Koopman, 2012) and the relative KS (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 7 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 8). 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 the variable type (Cameletti, Ignaccolo, and Bande, 2011). Table 8 lists the weather variables in the Agrimonia dataset.

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 $\{\sigma_d\}$

```

1 Preprocessing;
2 foreach  $y_t \in \{y_1, \dots, y_T\}$  do
3   if  $y_t$  not validated or  $y_t < 0$  then
4      $y_t \leftarrow \text{NaN};$ 

5 Impute missing values using Kalman smoother;
6  $\hat{y}_t, \hat{\sigma}_t \leftarrow \text{SSM}(y_t);$ 

7 Daily Aggregation;
8 foreach day  $d$  do
9   if number of consecutive hourly missing values  $> 6$  then
10     $\bar{y}_d \leftarrow \text{NaN};$ 
11  else
12     $\bar{y}_d \leftarrow \frac{1}{n_d} \sum_{t \in d} \hat{y}_t;$ 
13     $\sigma_d \leftarrow \sqrt{1/24^2 \sum_{t \in d} \hat{\sigma}_t^2};$ 
14 return  $\{\bar{y}_d, \sigma_d\}$ 

```

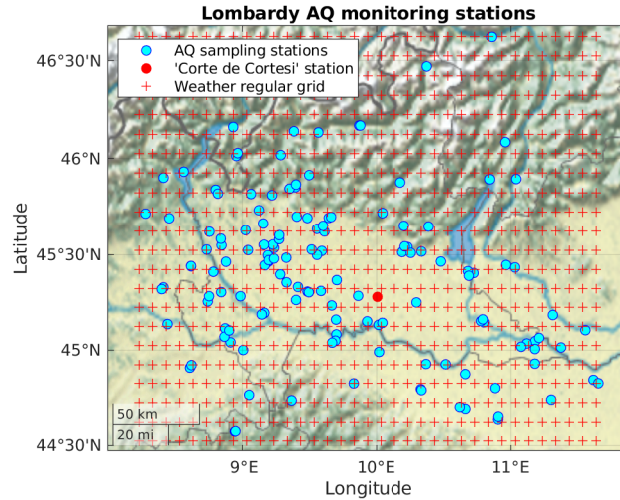


Figure 15: The irregularly located AQ monitoring stations (cyan circles) and the weather grid centres (red '+' symbols). The station named 'Corte de Cortesi' is marked as a red circle.

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 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.

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

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 7: 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 3 only.

To avoid oscillations, overshoots, edge effects, and negative values, we use piecewise cubic Hermite interpolating polynomials (PCHIP; Fritsch and Carlson (1980)). As discussed in more detail in the technical validation section (see Appendix A), 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 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 daily time series of swine and bovine density for each monitoring station. See Table 10 for a summary of the livestock variables in the Agrimonia dataset.

Variable name	Description	Temporal aggregation function	Units
WE_temp_2m	Temperature of air at 2 m	Daily mean	$^{\circ}\text{C}$
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	<i>Categ.</i>
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	<i>Categ.</i>
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	<i>Categ.</i>
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 8: 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 9: 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.

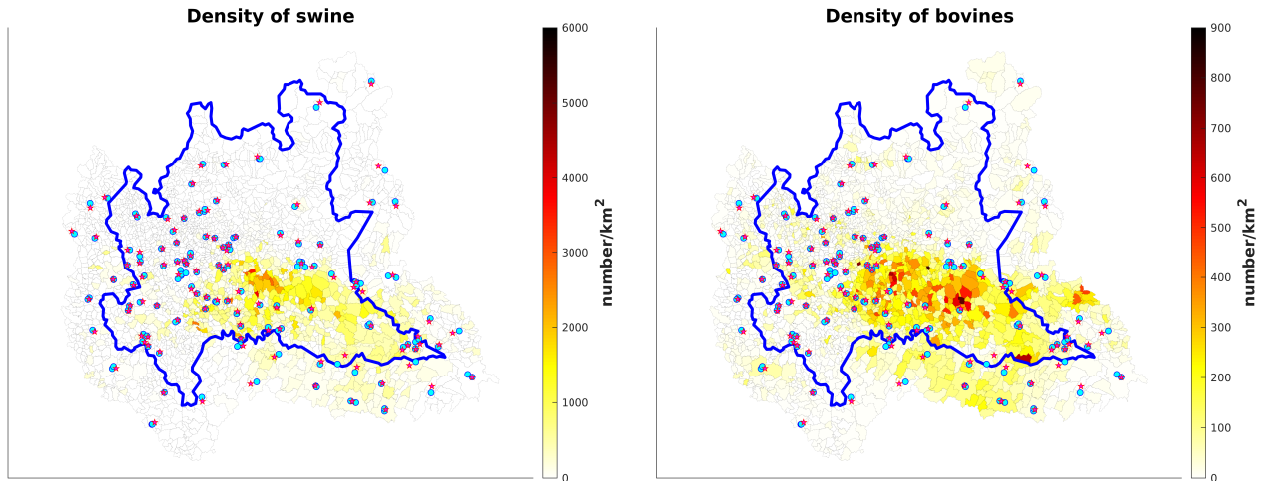


Figure 16: Livestock density over the augmented Lombardy region on 31 December 2021. Left: swine density, right: bovine density. The AQ stations (cyan circle) are spread randomly over the studied area. Each station is associated with the municipality centroid to which it belongs (red start). Based on 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 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 Appendix A). The CLC provides categorical data in polygons, while SIARL does so on a regular grid. In both cases, each AQ station

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

Table 10: 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).

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	<i>Categ.</i>
LA_soil_use	SIARL Lombardy - Lombardy soil use across 21 sectors	None	None	<i>Categ.</i>

Table 11: 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.

4.3.1 Validation of the harmonisation procedure

To assess the harmonisation procedure, Appendix A details the validation analyses performed on the Agrimonia dataset. Table 12 summarises the main validation results and methodological assumptions for each data component, highlighting the treatment of missing values, spatial and temporal representativeness, interpolation strategies, and simplifying assumptions adopted during the harmonisation.

Component	Validation objective	Method and outcome
AQ	Missing data imputation validation	Performance of the imputation method for hourly and bi-hourly time series is assessed by comparing daily averages from imputed data with agency-provided daily values (not officially validated). The resulting RMSE is low. For bi-hourly stations, daily means computed with and without imputation show small RMSE and low uncertainty, indicating no systematic bias.
WE	Meteorological representativeness	Meteorological inputs are derived from ERA5, providing variables averaged over $0.25^\circ \times 0.25^\circ$ grid cells. ERA5 is used to ensure spatial coverage and is not intended to reproduce local station-level conditions.
EM & LI	Temporal interpolation of discontinuous processes	Emissions and livestock variables are driven by discontinuous farm-level processes and population dynamics. Interpolated time series are interpreted as smooth approximations of irregular step changes. PCHIP splines are adopted to prevent oscillations, overshoots, boundary artefacts, and negative values.
LA	Land-use consistency over time	Land use is derived from CORINE Land Cover (2018) and assumed constant over 2016–2021. This is a minor approximation given slow land-use changes and is consistent with the fixed station-type classification reported in Table 3.

Table 12: Summary of validation procedures and methodological assumptions adopted for each component of the Agrimonia dataset.

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 $\mathbf{x}(\cdot) : \mathcal{D} \rightarrow \mathbb{R}^b$ 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 issue is known as a 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 (Gelfand, Zhu, and Carlin, 2001).

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). 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 1’053 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 the joint modelling of environmental and agricultural factors.

Weather

Weather variables are provided with different spatial resolutions (see Table 2). To associate the weather variable with the cells of the AGC dataset, we used the IDW interpolation using the four 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 NH_3 emissions. The seasonal behaviour of NH_3 emissions is determined by seasonal practices of agricultural activities responsible for strong variations in the NH_3 emissions levels. In particular, the NH_3 emissions are higher in the spring and summer periods because these periods coincide with the widespread application of manure and synthetic fertilisers. Moreover, elevated temperatures in those months enhance ammonia volatilisation, increasing emissions. In contrast, winter sees much lower emissions 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 NH_3 concentration on the AQ. However, areas presenting high levels of ammonia emissions are characterised by high concentrations of PM_{10} and $PM_{2.5}$.

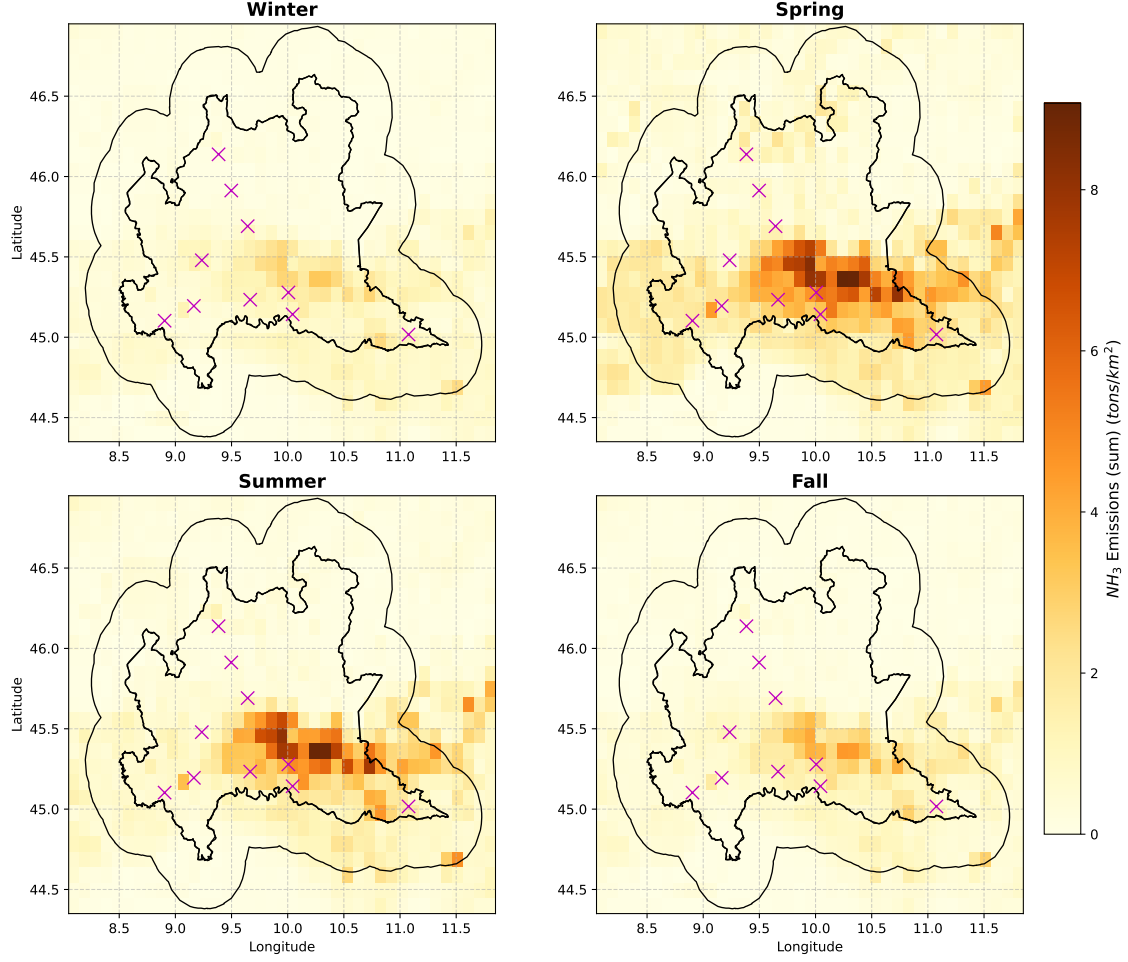


Figure 17: Season variability of the NH_3 emission in (tons/km^2) in 2020. From left to right, from top to bottom, the seasons Winter, Spring, Summer and Fall are shown. The NH_3 air quality network is depicted by magenta crosses.

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 must be reshaped from the area level to the AGC.

Consider the AGC output grid (0.1×0.1) such that $\mathcal{D} = \cup_i^K \mathcal{A}_i$, where $\mathcal{A}_i$ is the area associated 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 , denote the density level for the area $\mathcal{A}_i$. x_i cannot be measured directly, as the livestock density is provided only at the municipal level; that is, $\{x(s_1), \dots, x(s_n)\}$, with municipality centroids at coordinates $\{s_1, \dots, s_n\}$. Let $\{x(s_1), \dots, x(s_n)\}$ be a realisation of the latent stochastic process $x(s)$, with $s \in \mathcal{D}$ representing the true livestock density, which is continuous

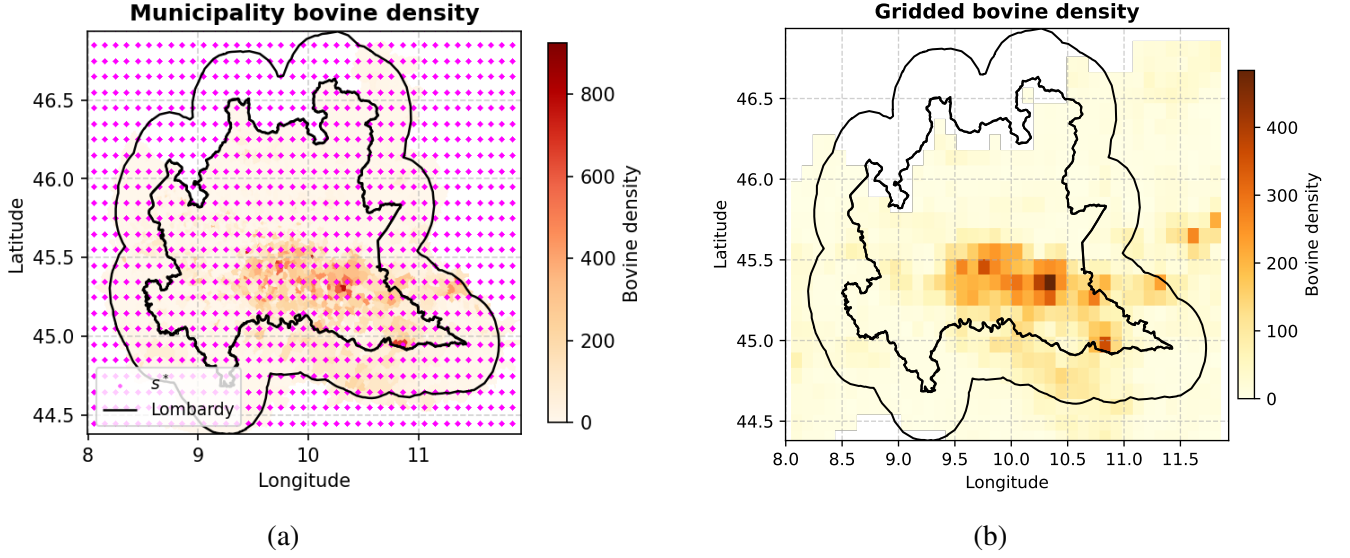


Figure 18: Left: bovine density at the municipal level. The magenta dots represent the centroids of the AGC regular grid (0.1×0.1) degrees, while the Lombardy boundaries are depicted by the solid black line. Right: resulting bovine density organised on a regular grid, computed by the discretisation of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

in space. The average density over the area $\mathcal{A}_i$ would be given by

$$x_i = \int_{s \in \mathcal{A}_i} x(s)w(s)ds, \quad (4.2)$$

where $w(s)$ is a weight for a generic spatial point $s \in \mathcal{A}_i$ such that $\int_{s \in \mathcal{A}_i} w(s)ds = 1$. The stochastic integral in Eq. (4.2) can be computed through Monte Carlo integration using a set of additional points, denoted by 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:

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

where $\mathcal{N}_i = \{s^* : s^* \cap \mathcal{A}_i = s^*\}$ is the set of points inside area $\mathcal{A}_i$. $s^* \in \mathcal{D}$ are drawn from a regular lattice of (0.01×0.01) degrees, 10 times lower than the AGC grid, such that each $\mathcal{A}_i$ is discretised with 100 points (i.e. $|\mathcal{N}_i| = 100$). The estimated mean values $\hat{x}(s^*)$ 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 $w(s) = 1/|\mathcal{N}_i|$. 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 the quantity $\hat{x}(s^*) = E[x(s^*)|\{x(s_1), \dots, x(s_n)\}]$ can then be computed using such a model.

Land cover and altitude

Land variables are provided from the Copernicus Land Monitoring Service as a shape file with a polygon associated with categorical variables that 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 the discretisation in Eq. (4.3). Since the expected outcome is a categorical value, the mode is used to obtain the most probable realisation among $s(s^*)$. Future developments could also 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. To this end, the discretisation in Eq. (4.3) was used.

4.5 Data and code availability

The output dataset has been built by joining the daily time series related to the air quality (AQ), weather (WE), emission (EM), livestock (LI) and land (LA) variables discussed in the previous sections and referred to the same AQ monitoring station for the Lombardy region augmented by the 0.3° buffer, depicted in Figure 13. The dataset and the metadata files are available on the Zenodo repository (Fassò, Rodeschini, et al., 2023b) as follows:

- *Agrimonia_Dataset_version.csv*: this is the Agrimonia output dataset joining the daily time series, at station locations, related to the AQ (see Table 7), WE (see Table 8), EM (see Table 9), LI (see Table 10) and LA (see Table 11) variables. In order to simplify the access to variables in the Agrimonia dataset, the variable name starts with the dimension of the variable, e.g., the name of the variables related to the AQ dimension starts with ‘AQ_’. Missing data are denoted by the ‘NaN’ value. The dataset has $S = 141$ monitoring stations and $T = 2192$ days between 1st January 2016 and 31st December 2021, totalling $S \times T = 309072$ rows, with the unique identifier given by the pair station code and date in “YYYY-MM-DD” format. It contains 41 columns, including the following block: row header (station code, latitude and longitude, date and altitude), AQ (7 columns), WE (16 columns), EM (7 columns), LI (2 columns) and LA (4 columns).

This file is also made available in the *.mat* and *.Rdata* format for MATLAB and R software, respectively.

- *AGC_Dataset.csv*: this is the Agrimonia gridded output dataset (AGC) that includes daily information for the period from 2016 to 2020 of almost all variables within the Agrimonia dataset on an equidistant grid of $(0.1^\circ \times 0.1^\circ)$ resulting in $n = 1053$ cells. The AGC does not include AQ variables, as they come from the monitoring stations that are irregularly spread over the area considered. 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 the joint modelling of environmental and agricultural factors. This file is also made available in the *.mat* and *.Rdata* format for MATLAB and R software, respectively.
- *Metadata_Agrimonia.csv*: this is the main Agrimonia metadata file and provides further information for the sources used, variables imported, transformations applied and Agrimonia variables.
- *Metadata_monitoring_network_registry_v2_0_1.csv*: it contains details about the AQ monitoring stations, including station type, NUTS3 code, environment type, altitude, monitored pollutants and others. Each row represents a single sensor.
- *Metadata_AQ_imputation_uncertainty.csv*: it contains the estimate of the daily uncertainty due to missing data imputation for the AQ time series. In particular, for each AQ variable, days without missing hours have zero uncertainty, days with one or more imputed hours have a number resulting from the propagation of the uncertainty in the daily averaging, and days with a 'NaN' in the concentrations have a 'NaN' also in the uncertainty.
- *Metadata_LA_CORINE_labels.csv*: it contains labels and descriptions associated with the CLC land variables (column CORINE code).
- *Metadata_LA_SIARL_labels.csv*: it contains labels and descriptions associated with the SIARL land variables (column SIARL code).

Details on the data format and standards used to build the Agrimonia dataset are provided in Appendix A.1. The Agrimonia dataset can be reproduced using the code and associated documentation available at the GitHub repository: https://github.com/AgrImOnIA-project/AgrImOnIA_Data. Specifically, Air Quality and Livestock data are pre-processed using MATLAB, while Weather, Emissions, and Land data are pre-processed using R. Spatial interpolations and the final data merging are also performed in R.

Chapter 5

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

This chapter presents different spatio-temporal modelling approaches useful for relating air quality 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 detailed 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 geo-statistical models (HDGMs), generalised additive mixed models (GAMMs), and the random forest spatio-temporal kriging models (RFSTK). All three models employed in this study have been specifi-

cally 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 modelling heteroskedasticity in two main frameworks: (i): time-varying unstructured flexible error variance and (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 emissions, thereby assessing the sensitivity of fine PM 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 stochastic 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 variable Season, which takes four categories: winter, spring, summer and autumn. Table 13 summarises the variables used in this study along with the main statistics.

Name	min	mean	max	std
PM _{2.5} [$\mu\text{g}/\text{m}^3$]	1.00	21.58	182.00	16.63
Wind speed [m/s]	0.56	2.59	11.93	1.33
Temperature [$^\circ\text{C}$]	-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 [$\text{mg}/(\text{m}^2\text{day})$]	0.05	3.34	45.06	5.53
NH ₃ [$\text{mg}/(\text{m}^2\text{day})$]	0.13	11.55	72.03	12.00
NO _x [$\text{mg}/(\text{m}^2\text{day})$]	0.75	14.06	118.60	16.73
Rain	0	0.02	1	0.15
Urban	0	0.75	1	0.43

Table 13: 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 shown 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 $10\mu\text{gm}^{-3}$ and $40\mu\text{gm}^{-3}$ across the year. Thus, throughout the year, the median concentration was beyond the threshold of $5\mu\text{gm}^{-3}$ considered hazardous by the WHO 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, 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.

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. Similarly, the seasonal behaviour

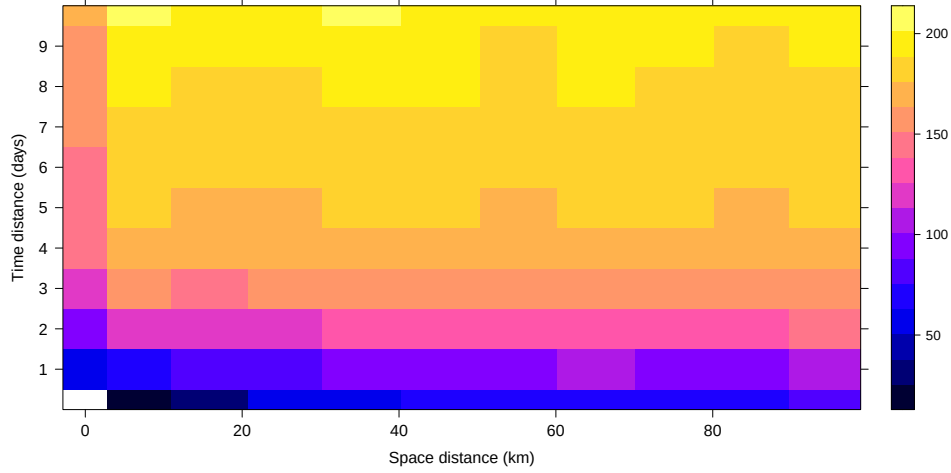


Figure 19: Spatio-temporal variogram computed on $\text{PM}_{2.5}$ daily observations at 45 stations from 2016 to 2020.

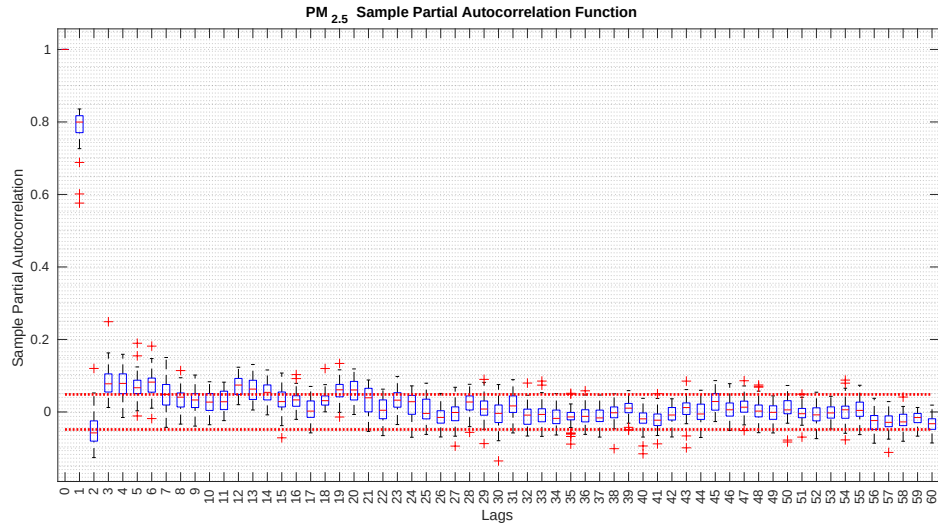


Figure 20: Partial Auto-Correlation Function (PACF) calculated by stations for the first 60 lags (days) and summarised through box plots. The i -th box plot summarises the partial autocorrelation calculated for each station at the i -th lag.

of NH_3 emissions is determined by seasonal practices of agricultural activities responsible for strong variations in the NH_3 emission levels. As shown in Figure 17, the NH_3 emissions are higher in the spring and summer periods. Table 14 summarises the daily temperature average, emissions flow, and $\text{PM}_{2.5}$ 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 15. 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 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.

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 14: Seasonal average of daily Temperature [$^{\circ}\text{C}$], NH₃, NO_x and SO₂ emissions [mg/m^2] and PM_{2.5} concentrations [$\mu\text{g}/\text{m}^3$].

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 15: Total NH₃ 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 also account for the spatial correlation. Indeed, the spatio-temporal variogram shown in Figure 19 highlights the 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 $PM_{2.5}$ concentrations, we compare three different classes of models: a geostatistical model, a semiparametric model, and a machine learning technique. More precisely, HDGM, GAMM, and 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. The HDGM incorporates a latent variable to capture spatio-temporal dependence, while external factors are included linearly within the model. Conversely, the 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, the 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, the 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 Gijssman, 2021). The 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, the GAMM has been employed in various areas, such as ecology (Knappe, 2016; Kneib, Knauer, and Küchenhoff, 2011), psychology (Bono, Alarcón, and Blanca, 2021), and economics (Fahrmeir and Lang, 2001). 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 Drăguț 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, including generalised additive models (Bertaccini, Dukic, and Ignaccolo, 2012), Bayesian hierarchical models based on the SPDE approach (Cameletti, Finn Lindgren, et al., 2013; Fioravanti et al., 2021), or RF (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 RF, are adept at capturing nonlinearities and interactions. However, 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 the RF 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 $\{z(\mathbf{s}, t) : \mathbf{s} \in \mathcal{D}, t = 1, 2, \dots, T\}$, as in Definition 9, where $\mathcal{D}$ is the spatial domain that contains a set of locations $\{\mathbf{s}_i : i = 1, \dots, n\}$ (ground-level measurement stations) and the discrete temporal domain is $t = 1, \dots, T$ (daily). Furthermore, we assume that $z(\mathbf{s}, t)$ may be influenced by external variables. Using the SRE model in Eq. (2.15), we can express the generic univariate spatio-temporal process as:

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

where $S(\mathbf{s}, t)$ are the large-scale components, $z(\mathbf{s}, t)$ are small-scale spatio-temporal random effects, and $\varepsilon(\mathbf{s}, t)$ 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 was 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 $S(\mathbf{s}, t)$. Using the Wilkinson notation, $S(\mathbf{s}, t)$ can be expressed as follows: $\log(PM_{2.5}) = 1 + \text{Altitude} + \text{Wind_speed} + \text{Temperature} + \text{Precipitation} + \text{RH} + \text{BLH} + \text{LVI} + \text{HVI} + \text{Pigs} + \text{Bovine} + \text{Month}$.

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 GP with a spatial covariance function. Under these conditions, the HDGM can be cast into the SSM in Eqs. (2.23)–(2.24).

Let $y(\mathbf{s}, t)$ be the response variable observed at site $\mathbf{s} \in \mathbb{S}^2$, 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:

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

where $S(\mathbf{s}, t) = \mathbf{x}'(\mathbf{s}, t)\boldsymbol{\beta}$ is the regression term with fixed effect coefficients $\boldsymbol{\beta}$, $\mathbf{x}(\mathbf{s}, t)$ 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 $\epsilon(\mathbf{s}) \sim N(0, \sigma_\epsilon^2)$ is independent in space and time, $z(\mathbf{s}, t)$ is a unit-variance Markovian scalar process ruled by the transition coefficient f , the innovation term $\eta(\mathbf{s}, t)$ is a zero-mean GP, $GP(0, \rho(\|\mathbf{s} - \mathbf{s}'\|; \theta))$, independent in time where ρ is a valid spatial correlation function, and $\|\mathbf{s} - \mathbf{s}'\|$ is the geodetic distance between $\mathbf{s}$ and $\mathbf{s}' \in \mathbb{S}^2$. The model parameters set is $\Pi = \{\boldsymbol{\beta}, f, \theta, \alpha, \sigma_\epsilon^2\}$, which is estimated using the MLE 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 (Calculi et al., 2015), which makes the EM algorithm an efficient method for obtaining the MLE.

Generalised additive mixed model

Compared to generalised additive models (GAMs; Hastie and Tibshirani 1987), GAMMs 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

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

with $\mathbf{x}'(\mathbf{s}, t)\boldsymbol{\beta}$ being a linear parametric regression term of the first j -covariates with a parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_b)'$, including the intercept term β_0 . Moreover, $\sum_{j=1}^b \alpha_j(x_j(\mathbf{s}, t))$ is an additive term with nonlinear functions $\alpha_j(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$. $\alpha_j(\cdot)$ can be estimated along with the other model coefficients, for instance 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,

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

where g is the parameter representing the temporal dependence, where zero indicates no temporal correlation. The spatial dependence is modelled as a smooth surface $\eta(\mathbf{s})$, 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 the HDGM as given by Eq. (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 an RF (Breiman, 2001) algorithm. In other words, the third hybrid model combines an RF for the large-scale component $S(\mathbf{s}, t)$ with that of a spatio-temporal kriging model for $z(\mathbf{s}, t)$. The idea traces back to the combination of RF and kriging, the so-called random forest residual kriging, which, even if only considering spatial dependence, showed promising results compared to a 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 broad 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 $\mathbf{y}_j^*$ 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 $\hat{y}_j(\mathbf{s}_i)$ 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

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

for $i = 1, \dots, n$ and $t = 1, \dots, T$. $\hat{y}_j(\mathbf{s}_i)$ is the prediction of the j -th decision tree at location $\mathbf{s}_i$. It is essential to emphasise that for regression trees within the RF 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. RFs 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 whether a covariate is helpful to explain the response variable. In particular, the HDGM and GAMM use the MLE to estimate the model parameters, as well as the associated uncertainty. The GAMM is more flexible than the 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 the 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 the HDGM fixed effects to the spline of the GAMM and the partial dependence plot (PDP) within the 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 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:

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

	HDGM		GAMM		RFSTK	
	LS	FM	LS	FM	LS	FM
RMSE [$\mu g m^{-3}$]	11.92	1.814	11.456	8.468	8.245	5.361
MAE [$\mu g m^{-3}$]	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 16: In-sample performance of the three models assessed by RMSE (in $\mu g m^{-3}$), and the adjusted coefficient of determination R^2 . The fit of the only large-scale component (LS) is compared to the full model (FM).

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 (LOSOCV) scheme.

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 16. The HDGM generally had better prediction capabilities and lower computational costs than the other models when considering the FM. 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 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 to model the extremes.

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

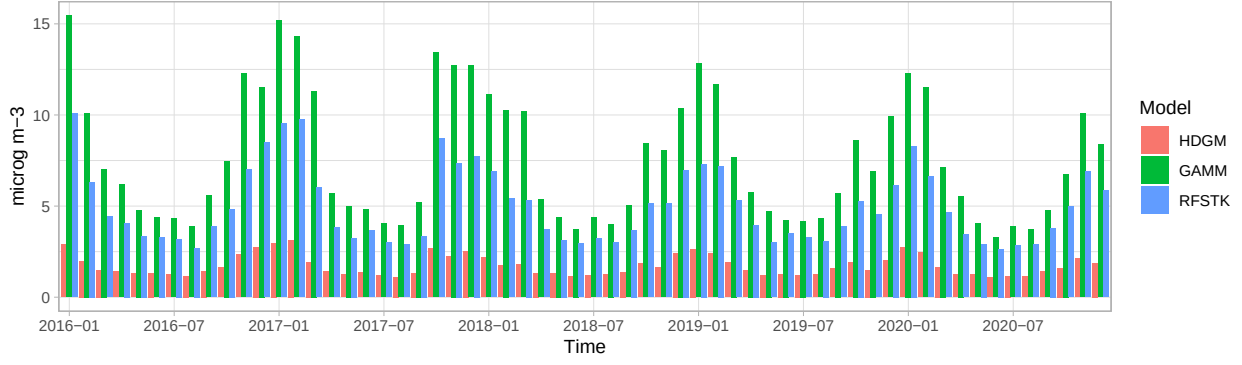


Figure 21: In-sample residual diagnostics for the three models. Standard deviation of residuals grouped by months for each model.

(2023) for much simpler mean models.

Furthermore, we investigated the spatial and temporal dependence of the residuals, estimating temporal autocorrelation functions (ACFs) and spatio-temporal variograms. Different patterns were observed across the three models. While the 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, the 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 LOSOCV scheme, which is a variation of the commonly used leave-one-out cross-validation (LOOCV) 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 an LOOCV provides an assessment of model performance but may overlook the temporal correlation of errors. By adopting an 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-

	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 17: Prediction performance indices evaluated with the LOSOCV scheme [μgm^{-3}]

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 17. The HDGM is confirmed to be the best model, but compared to the in-sample residuals in Table 16, 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 HDGM's generalisation ability. This could be due to the linear structure in the large-scale component. While a more flexible model (e.g., RF 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 the 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 tends to have a higher RMSE compared with the GAMM. The GAMM, which does not include strong, time-varying spatial interactions, was able to capture this difference in air quality better than the HDGM and RFSTK. In another situation where there is a clear spatial correlation among

observations, the RFSTK model performed the worst, while the HDGM showed the best performance. The reason may lie in RFSTK's 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, the 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 carefully select the model according to the local conditions. The best model for one location may not be the best for another. Moreover, model averaging could additionally improve the predictions.

As a result, and considering the cross-validation performance, the HDGM is confirmed to outperform the other models on the considered dataset. The HDGM, which belongs to the SSM framework, is fully parameterised. In particular, the dynamical representation of the temporal component, discussed in Section 2.4.3, allows the model to capture the transition of the latent state from time $t - 1$ to t , thereby capturing serial dependence in the temporal process. The gains from this dynamical specification can be substantial when temporal dependence is strong, which is consistent with the partial autocorrelation function shown in Figure 20, indicating a strong first-order temporal correlation in the underlying process. The HDGM explicitly model the measurement error, which allows the model to separate the variability due to the underlying signal (the “true underlying process”) from the variability due to noise, leading to more accurate estimation and prediction. Furthermore, the spatial covariance structure enables the use of a kriging-based predictor to estimate the response at locations included in the validation set. The estimated covariance parameter is statistically significant, indicating that the data provide strong evidence that the corresponding spatial effect is non-zero. The ability to model spatial variability is further supported by the residual analysis, which does not reveal remaining strong spatial patterns. However, compared to machine learning or semi-parametric approaches, a key limitation of the HDGM lies in its strong parametric assumptions. In particular, the use of a stationary covariance function limits the model's ability to capture spatial non-linearities, such as abrupt changes over short spatial distances.

Model comparison discussion

A comprehensive interpretation of each estimated model is detailed 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 the HDGM and GAMM (with January as the worst month for air quality) and ‘month’ receiving the highest

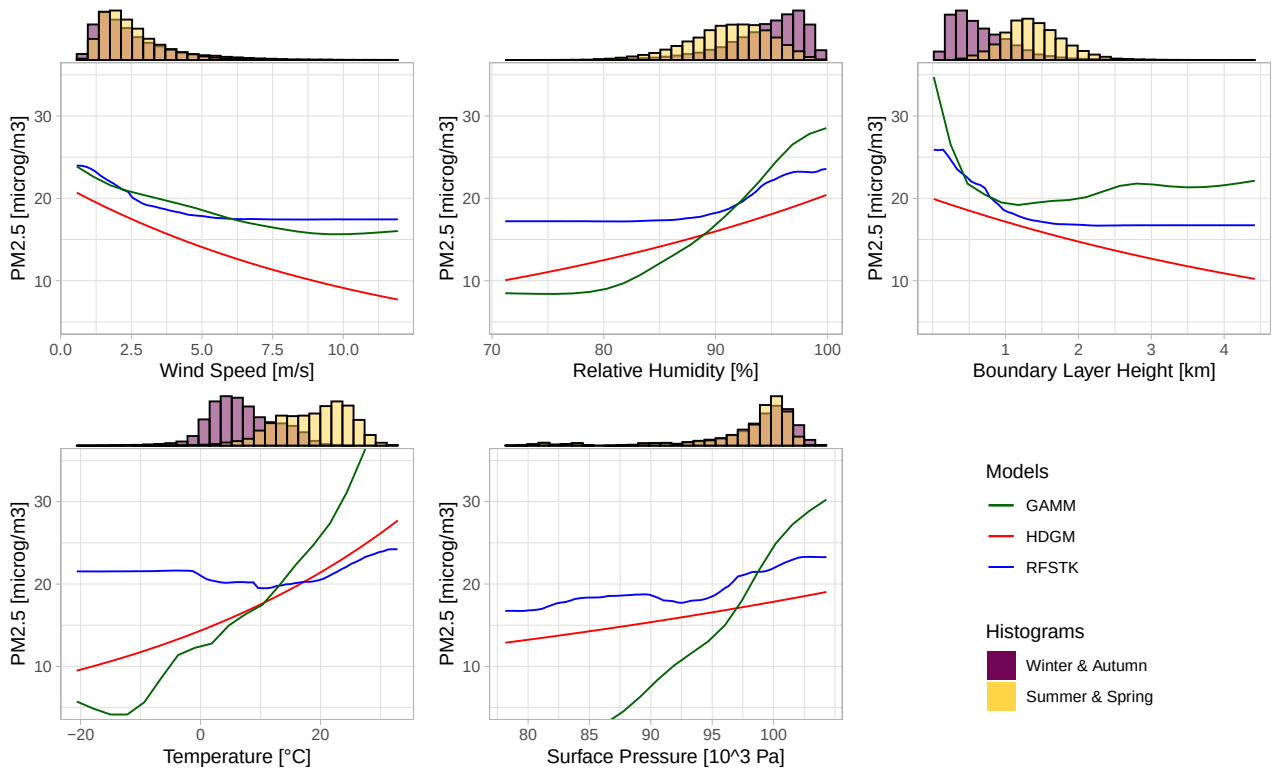


Figure 22: PDP calculated on the large scale of all three models. In yellow, the empirical histogram is for cold months, while in purple for hot months.

ranking in the VIF of the 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 the 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 the 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.

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. The RFSTK model can capture these interactions more effectively than the other two models, which is why its PDP is flatter. This suggests that the RFSTK model 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 techniques can be used to improve the performance and interpretability of classic geostatistical approaches, such as the HDGM or GAMM. This is because machine learning 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, such as the GAMM. The comparison of PDPs also highlights the complementary nature of machine learning techniques and classic approaches. Machine learning techniques are better at capturing complex relationships, while traditional approaches are easier to interpret and allow for straightforward uncertainty estimation. Thus, we advise combining machine learning techniques and classic approaches for modelling and predicting PM concentrations.

5.2.3 Model 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 18 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

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 18: Estimated coefficients of the large-scale component of the HDGM model

vegetation (e.g. trees). The small-scale effect of the HDGM is defined by a latent variable $z(s, t)$, 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.

GAMM Results

The estimated coefficients of our second model, the GAMM, are presented in Table 19. 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

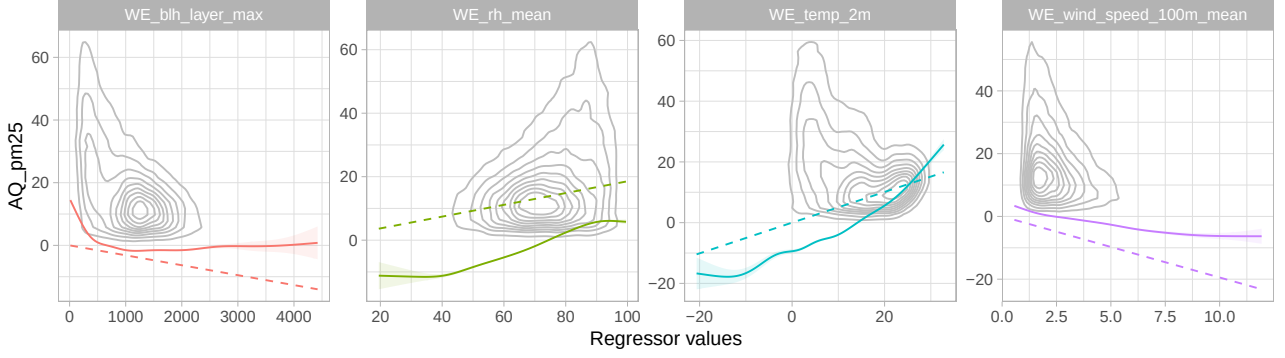


Figure 23: Regression splines (continuous lines), including their 95% confidence intervals (coloured shadows), and HDGM regression coefficients (dashed lines) for relevant weather regressors (i.e., boundary layer height, relative humidity, temperature, and wind speed). Grey contour lines represent the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

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 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 $\eta(s)$ is displayed in Figure 24. This smoothing spline $\eta(s)$ 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 $\hat{\theta} = 1.16^\circ$, 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.

<i>A: Linear effects</i>	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
<i>B: Nonlinear effects</i>	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 19: 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).

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 corre-

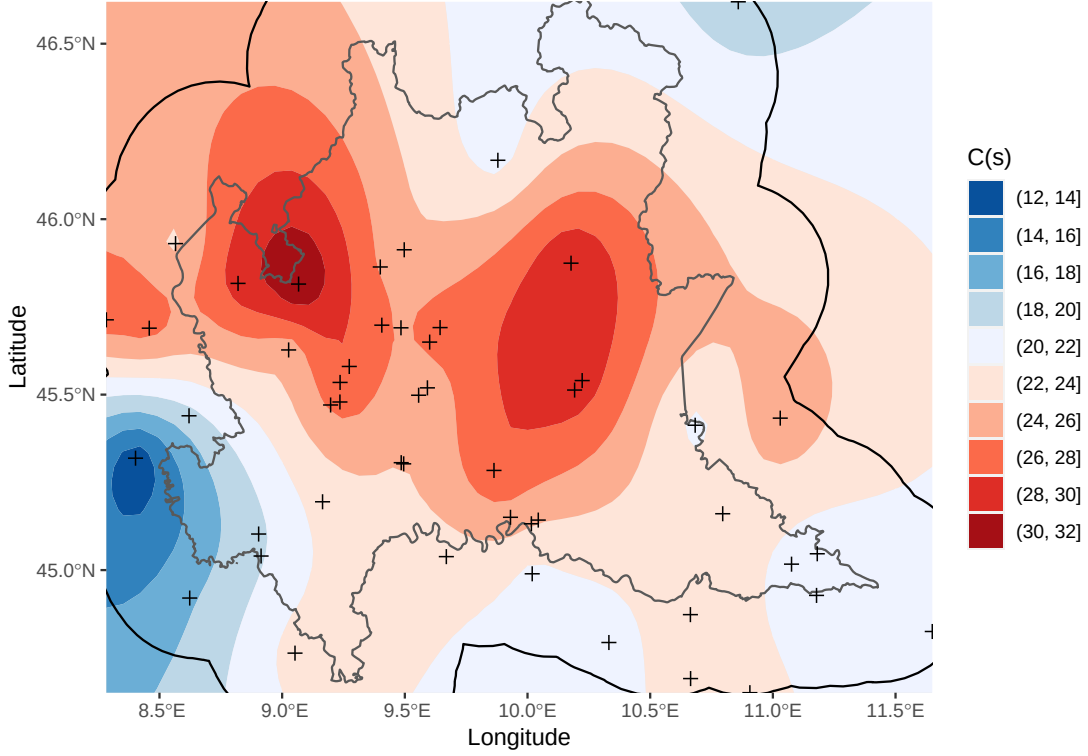


Figure 24: Estimated smoothing spline $\hat{\eta}(s)$ of GAMM, which corresponds to the $PM_{2.5}$ predicted on a regular grid using the large scale of GAMM, where all regressors are set to 0. Stations are marked with a black cross, Lombardy boundaries are shown in grey, and the black line marks the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

sponding variogram model is fitted to the RF residuals. For our analysis, we obtained exponential decay parameters of $\hat{\theta}_t = 0.78$ days and $\hat{\theta}_s = 0.48^\circ$ 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 16.

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

Figure 21 in Section 5.2 shows that all three models exhibit time-varying uncertainty and a clear temporal pattern throughout the year. When residuals are heteroskedastic, inference under a homoskedastic assumption is typically biased and inconsistent, potentially leading to invalid hypothesis tests. More generally, in likelihood-based models, heteroskedasticity may result in biased and inconsistent maximum likelihood estimates unless the likelihood is adjusted to account for the heteroskedastic structure. Figure 21, therefore, highlights the need for a heteroskedastic model to not underestimate predictive uncertainty during the winter months. Furthermore, the comparison in Section 5.2 shows that the dynamical formulation of the SSM achieves the best performance in the cross-validation study for $PM_{2.5}$ modelling in Lombardy (see Table 17).

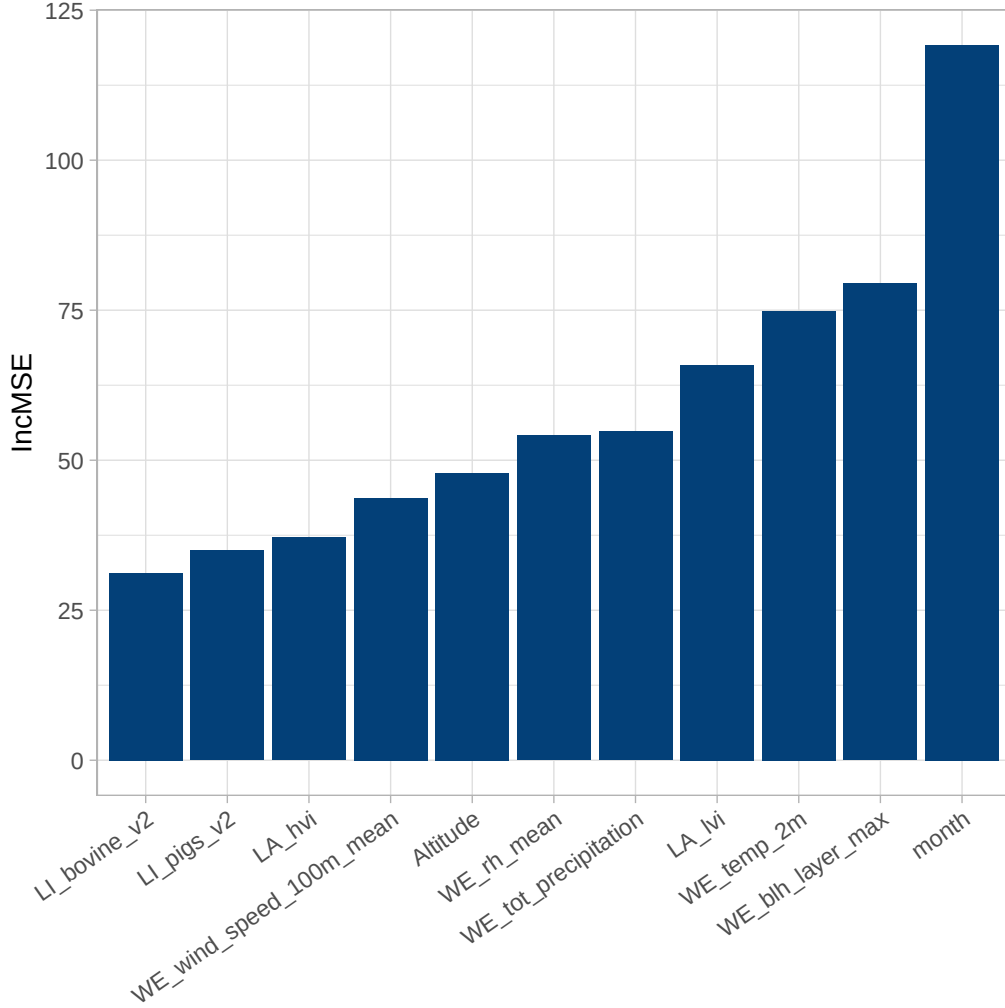


Figure 25: Variable importance factor (measured as IncMSE) for the 11 selected features in the large-scale component of the RFSTK.

In the SSMs, heteroskedasticity in the observation process affects both filtering and smoothing uncertainty, as well as predictive uncertainty. The KF relies on the assumed measurement error variance to compute the Kalman gain and the posterior state covariance. Misspecifying a time-varying observation variance as homoskedastic leads to incorrect uncertainty quantification. In particular, the posterior state covariance is typically underestimated during high-variance periods and overestimated during low-variance periods. Similarly, under a homoskedastic assumption, predictive uncertainty fails to adapt to changes in the noise level, resulting in predictive intervals with incorrect coverage properties. From a parameter estimation perspective, heteroskedasticity is often absorbed incorrectly into other components of the model, such as the state innovation variance or the transition dynamics. This can further distort uncertainty quantification. Consequently, even when state estimates remain approximately unbiased, standard errors and prediction intervals are generally unreliable.

Building on these insights, here, we focus on developing a novel multivariate heteroskedastic spatio-temporal model designed to estimate $PM_{2.5}$ concentration surfaces from a finite set of observations.

We base our proposed model on the SSM framework to explicitly account for heteroscedasticity, and since the preliminary analysis of $\text{PM}_{2.5}$ indicates time-varying variance, with spatial-varying variance being less pronounced, we consider modelling the temporal error variance using either an unstructured approach or a functional specification. In particular, we propose modelling heteroskedasticity in two main frameworks: (i): time-varying unstructured flexible error variance and (ii) deterministic skedastic function 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 EM algorithm.

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 via different sensors or processes (see, e.g., Smith, Yan, and Cowles (2008)). Considering the temporal dimension, alternative parameterisations could also be considered, such as modelling the variance as a function of covariates that explain its temporal behaviour (see, e.g., J. S. Benth, F. E. Benth, and Jalinskas (2007) and Fassò (2023)) or adopting specifications from the GARCH family models introduced by the Nobel Prize winner Robert F. Engle (1982). GARCH-like models are based on conditioning the error variance on the past observations (see, e.g., Francq and Zakoian (2019)). Considering the spatial dimension, deterministic spatial skedastic functions have been used extensively. For example, spatial skedastic functions were considered by Mur and Angulo (2009) in spatial econometric models and by Hulshof and Umana (2023) in ecology. Engle and Bollerslev's generalised conditionally heteroskedastic approach has recently been 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. However, here 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.

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 novel heteroskedastic extension of the HDGM (see Section 5.2.1). The HDGM in Eq. (5.2) is a two-level hierarchical model where the hierarchy is constructed by putting together two conditional submodels. The heteroskedastic extension concerns only the measurement error ϵ ; 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 $\mathbf{y}(\mathbf{s}, t)$ be the p -variate response variable observed at site $\mathbf{s} \in \mathbb{S}^2$, 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:

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

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

Functional heteroskedastic HDGM

Estimating daily heteroskedastic variance through unstructured models, like the model in Eq. (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. From an interpretability perspective, it is preferable to model a smoothly varying variance, for which a differentiable function is advantageous in understanding the underlying process. To address this, a more flexible approach involves modelling heteroskedasticity using a functional representation, in which the variance is expressed as a smooth function of time. In particular, representing the error variance as a linear combination of basis functions with unknown coefficients provides a flexible yet parsimonious framework, reducing the number of parameters to be estimated while retaining modelling flexibility. Specifically, let the multivariate functional variance be denoted as $\boldsymbol{\sigma}_\epsilon^2(t) = (\sigma_{\epsilon, 1}^2(t), \dots, \sigma_{\epsilon, p}^2(t))$, and let $\boldsymbol{\Psi}_i(t) = (\psi_{1, i}(t), \dots, \psi_{R, i}(t))$ be a $R \times 1$ where $\psi_{i, j}(t)$ is a known basis function $\psi_{i, j} : \mathbb{R}^+ \rightarrow \mathbb{R}$ evaluated at time t . The error variance for the i -components is modelled as

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

The definition of $\psi_{i,j}$ usually requires the choice of a reference set of points called knots $v_i \in \mathcal{V}_R$ for $i = 1, \dots, R$ that span the considered domain $\mathcal{T}$. The finer the grid of knots, the more locally flexible the resulting function.

The functional approach offers interpretability and ensures a continuous variance structure over time. More compactly, Eq. (5.8) can be written as $\log(\sigma_\epsilon^2(t)) = \Phi(t)\mathbf{c}_\epsilon$, where $\Phi(t) = \text{blockdiag}(\Psi_1(t), \dots, \Psi_p(t))$, and $\mathbf{c}_\epsilon = (c'_{\epsilon,1}, \dots, c'_{\epsilon,p})'$ is the vector of unknown parameters. The model parameters set is $\Pi = \{\beta, \text{diag}(\mathbf{F}), \vartheta, \text{diag}(\mathbf{A}), \mathbf{V}, \mathbf{c}_{\epsilon,1}, \dots, \mathbf{c}_{\epsilon,p}\}$, which is estimated using the MLE through the EM algorithm. To simplify the notation, hereinafter, we consider $\Psi_1 = \dots = \Psi_p$. Different basis functions Ψ_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 dynamics of variance. The HDGM relies on the KS 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 and estimation formulas

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

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

with $\mathbf{X}_t = \text{blockdiag}(\mathbf{X}_{1,t}, \dots, \mathbf{X}_{p,t})$, where $\mathbf{X}_{i,t} = \text{stack}(\mathbf{x}(s_{1,i}, t), \dots, \mathbf{x}(s_{m_i,i}, t))$, and $x(s, t)$ is the known function of the covariates. Vectors $\mathbf{z}_t$ and $\boldsymbol{\eta}_t$ are defined similarly to $\mathbf{y}_t$. Moreover, considering the unstructured error variance, ϵ_t is a p -dimensional random noise vector with independent components such that $\epsilon_t \sim N(0, \Sigma_{\epsilon,t})$, where $\Sigma_{\epsilon,t}$ is given by $\Sigma_{\epsilon,t} = \text{blockdiag}(\Sigma_1(t), \dots, \Sigma_p(t))$, where $\Sigma_i(t) = \sigma_{\epsilon,i}^2(t) \mathbf{I}_{m_i}$, and $\mathbf{I}_{m_i}$ is the identity matrix of order m_i . The distribution of the latent variable at $t = 0$, $\mathbf{z}_0 \sim N(\boldsymbol{\mu}_0, \Sigma_0)$. If all parameters Π are known, the unobserved temporal process $\mathbf{z}_t$ in the model in Eq. (5.9) is estimated for each time t through the KS technique, which is

initialised with the condition $\mathbf{z}_0$. The KS algorithm handles time-varying parameters in a natural way. Alongside the KS output $\mathbf{z}_t^T$, $\mathbf{P}_t^T$, and $\mathbf{P}_{t,t-1}^T$, the quantities $\mathbf{S}_{10} = \sum_{t=1}^T \mathbf{z}_t^T (\mathbf{z}_{t-1}^T)' + \mathbf{P}_{t,t-1}^T$, $\mathbf{S}_{00} = \sum_{t=1}^T \mathbf{z}_{t-1}^T (\mathbf{z}_{t-1}^T)' + \mathbf{P}_{t-1}^T$, and $\mathbf{S}_{11} = \sum_{t=1}^T \mathbf{z}_t^T (\mathbf{z}_t^T)' + \mathbf{P}_t^T$ 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 an initial value $\Pi^{(0)}$, the update formulas are:

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

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

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

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

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

$$\boldsymbol{\theta}^{(m+1)} = \underset{\boldsymbol{\theta}}{\text{argmax}} \left(T \log |\boldsymbol{\Sigma}_{\eta}^{(m)}| + \text{tr} \left[(\boldsymbol{\Sigma}_{\eta}^{(m)})^{-1} \left(\mathbf{S}_{11}^{(m)} - \mathbf{S}_{10}^{(m)} \mathbf{F}^{(m)} - \mathbf{F}^{(m)} \mathbf{S}_{10}^{\prime(m)} + \mathbf{F}^{(m)} \mathbf{S}_{00}^{(m)} \mathbf{F}^{(m)} \right) \right] \right), \quad (5.15)$$

$$(\sigma_{\epsilon,t}^2)^{(m+1)} = \frac{1}{n} \text{tr}(\boldsymbol{\Lambda}_t^{(m)}), \quad (5.16)$$

where $\boldsymbol{\Lambda}_t^{(m)} = \mathbb{E}_{\Pi^m}(\mathbf{e}_t | \mathbf{y}_{1:T}) \mathbb{E}_{\Pi^m}(\mathbf{e}_t | \mathbf{y}_{1:T})' + \text{Var}_{\Pi^m}(\mathbf{e}_t | \mathbf{y}_{1:T})$ with $\mathbb{E}_{\Pi^m}(\mathbf{e}_t | \mathbf{y}_{1:T}) = \mathbf{y}_t - \mathbf{X}_t \boldsymbol{\beta}^{(m)} - \mathbf{A}^{(m)} \mathbf{z}_t^T$ and $\text{Var}_{\Pi^m}(\mathbf{e}_t | \mathbf{y}_{1:T}) = \mathbf{A}^{(m)} \mathbf{P}_t^T \mathbf{A}^{(m)}$.

Note that the updated Eqs. (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 updated formulas for $\boldsymbol{\beta}$ and $\mathbf{A}$, in Eqs. (5.10) and (5.11), respectively, take into account the time-varying variance $\boldsymbol{\Sigma}_{\epsilon,t}$. The derivation of the updated Eqs. (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

$$\Sigma_{\epsilon,t} = \text{blockdiag} \left(\exp(\Psi_1(t)c_{\epsilon,1}I_{m_1}), \dots, \exp(\Psi_p(t)c_{\epsilon,p}I_{m_p}) \right). \quad (5.17)$$

The log-likelihood restricted to the vector $c_\epsilon = (c_{\epsilon,1}, \dots, c_{\epsilon,p})$ at the iteration m became $Q(c_\epsilon, c_\epsilon^{(m)}) = \sum_{t=1}^T \log |\Sigma_{\epsilon,t}^{(m)}| + \sum_{t=1}^T \text{tr}(\Sigma_{\epsilon,t}^{(m)})^{-1} \Lambda_t$. The MLE solution is obtained by maximising the $Q(c_\epsilon, c_\epsilon^{(m)})$ function. The solutions are given by the roots of the derivatives with respect to the c_ϵ , which are given by:

$$c_{\epsilon,i} = \text{solve} \left\{ \sum_{t=1}^T m_i \Psi_i(t) c_{\epsilon,i} - \text{tr}(\Psi_i(t) \exp(\Psi_i(t) c_{\epsilon,i}) I_{m_i} \Lambda_{i,t}) \right\}, \quad (5.18)$$

where solve is a numerical procedure to find the root of the non-linear system in the form $g(x) = 0$, where $g(x)$ 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.

It is worth noting that we did not conduct simulation studies aimed at assessing the sensitivity of parameter estimates to the chosen variance parameterisation, as the time-varying variance is treated as a nuisance parameter required to make correct inferences on spatial maps. Nevertheless, such simulations could be used in future work to evaluate the robustness of the proposed specification under variance misspecification, particularly as the sample size increases. We defer to future work the investigation of the sensitivity and robustness of the proposed parameterisation, with particular emphasis on the functional specification, and using the unstructured variance model as a benchmark for comparison.

Prediction and block kriging

Given the MLE $\hat{\Pi}$, the plug-in predictions at new locations $s_0 \in \mathcal{D}$ and times $t = 1, \dots, T$ are defined as

$$\hat{y}(s_0, t) = X(s_0, t) \hat{\beta} + A z_t^T(s_0), \quad (5.19)$$

with associated prediction variance

$$\Sigma_{\hat{y}}(s_0, t) = A \text{Var} [\hat{z}_t^T(s_0)] A', \quad (5.20)$$

where $z_t^T(s_0)$ denotes the smoothed state process evaluated at location s_0 , and $\text{Var}[z_t^T(s_0)]$ is given by the corresponding KS covariance $P_t^T(s_0)$. Here, the term ‘plug-in predictions’ refers to predictions where the MLE $\hat{\Pi}$ are considered fixed and known. From a computational perspective, obtaining $z_t^T(s_0)$ requires running the KF and KS using the estimated parameter set $\hat{\Pi}$, while augmenting the

state vector to account for the additional location s_0 . The computational complexity of this procedure is $O(m^3T)$ for each new site s_0 , where m is the dimension of the state vector. From a statistical standpoint, the estimate $z_t^T(s_0)$ can be interpreted as a kriging predictor, since the KF recursively computes conditional expectations under Gaussian assumptions, thereby providing the best linear unbiased prediction at new space-time locations.

Given the pointwise predictions for the j -variable $\hat{y}_j(s_0, t)$ with $s_0 \in \mathcal{D}$ and $t = 1, \dots, T$, one can make an inference on the underlying continuous spatial process. For visualisation purposes and to provide a dynamic map of the j -variable, it is often convenient to discretise the domain $\mathcal{D}$ into a collection of non-overlapping cells, such as a regular grid of pixels. More generally, 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 s_0 , but rather over a block $\mathcal{A}_i$. Formally, the block average of the process at time t is defined as

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

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

$$E[y_j(\mathcal{A}_i, t) \mid \mathbf{y}_{1:T}] = \frac{1}{|\mathcal{A}_i|} \int_{\mathcal{A}_i} E[y_j(\mathbf{u}, t) \mid \mathbf{y}_{1:T}] d\mathbf{u}. \quad (5.21)$$

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

$$\begin{aligned} E[y_j(\mathcal{A}_i, t) \mid \mathbf{y}_{1:T}] &\approx \frac{1}{R} \sum_{r=1}^R E[y_j(\mathbf{s}_r, t) \mid \mathbf{y}_{1:T}], \quad \mathbf{s}_r \in \mathcal{A}_i, \\ &= \frac{1}{R} \sum_{r=1}^R \hat{y}_j(\mathbf{s}_r, t). \end{aligned} \quad (5.22)$$

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

$$E[y_j(\mathcal{A}_i, t) \mid \mathbf{y}_{1:T}] \approx E[y_j(\mathbf{s}_r, t) \mid \mathbf{y}_{1:T}], \quad (5.23)$$

where $\mathbf{s}_r$ 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 one of 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. We adopt a model-based scenario analysis rather than a formal counterfactual framework. Counterfactual analysis is a causal concept grounded in causal inference theory and typically relies on potential outcomes or structural equation models to identify causal effects. In contrast, the “what-if” scenario analysis considered here is a broader, primarily predictive exercise, which may or may not admit a causal interpretation. Specifically, the analysis addresses the question of how model-based predictions change when inputs are modified, without invoking the assumptions required for causal identification.

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 reducing 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 $PM_{2.5}$ concentrations 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 instead 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 interested in comparing the predicted $\text{PM}_{2.5}$, $\hat{y}(\mathbf{s}, t)$, here based on the observed NH_3 emissions with a prediction, $\hat{y}^r(\mathbf{s}, t)$, where the NH_3 emissions are reduced by a factor of r . The daily $\text{PM}_{2.5}$ change $\Delta_y(\mathbf{s}, t) = y(\mathbf{s}, t) - y^r(\mathbf{s}, t)$ and its variance are estimated by

$$\Delta_{\hat{y}}(\mathbf{s}, t) = \hat{y}(\mathbf{s}, t) - \hat{y}^r(\mathbf{s}, t)$$

$$\text{Var}(\Delta_{\hat{y}}(\mathbf{s}, t) - \Delta_y(\mathbf{s}, t)) = \Delta'_x \Sigma_{\hat{\beta}} \Delta_x + \text{Var}(\epsilon_t - \epsilon_t^r),$$

where $\Delta_x = \mathbf{x}(\mathbf{s}, t) - \mathbf{x}^r(\mathbf{s}, t)$; $\Sigma_{\hat{\beta}}$ is the estimate of the $\hat{\beta}$ covariance matrix, and 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 $\Delta_{\hat{y}}(\mathbf{s}, t)$ and its variance can be extended over the block $\mathcal{A}_i \subset \mathcal{D}$ and time interval $\mathcal{I} \subset \{1, \dots, T\}$ by considering the block average in Eq. (5.21) and its discrete Monte Carlo computation in Eq. (5.22). Specifically, for this problem, the block average and its variance are given by

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

$$\text{Var}[\bar{\Delta}_{\hat{y}}(\mathcal{A}_i, \mathcal{I}) - \bar{\Delta}_y(\mathcal{A}_i, \mathcal{I})] = \bar{\Delta}_x(\mathcal{N}_i, \mathcal{I})' \Sigma_{\hat{\beta}} \bar{\Delta}_x(\mathcal{N}_i, \mathcal{I}) + \frac{2}{|\mathcal{I}|^2 |\mathcal{N}_i|^2} \sum_{t \in \mathcal{I}} \hat{\sigma}_{\epsilon, t}^2 \quad (5.25)$$

where $\mathcal{N}_i$ is the set of support points used to compute the Monte Carlo integration, and $|\cdot|$ denotes the cardinality. $\hat{\sigma}_{\epsilon, t}^2$ is the estimate variance of the measurement error, and $\bar{\Delta}_x(\mathcal{N}_i, \mathcal{I})$ is the average of the design matrix. Eq. (5.25) provides the standard deviation (std) used to assess the estimate change uncertainty in Session 5.4.3.

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 in Eq. (5.7). Furthermore, the comparison conducted in Section 5.2 provided detailed insights into the impact of weather conditions on $\text{PM}_{2.5}$ formation in Lombardy. Fassò (2023) explored the seasonality of pollutant concentration, revealing that the sensitivity of $\text{PM}_{2.5}$ 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:

$$\begin{aligned} \text{PM}_{2.5} \sim & (\text{Intercept}) + \text{Windspeed} + \text{Temperature} + \text{RH} + \text{Rain} + \text{BLH} + \text{Urban} + \text{HVI} \\ & + \text{SO}_2 + \text{NO}_x + \text{NH}_3 + (\text{NO}_x + \text{NH}_3) : \text{Season}, \end{aligned} \quad (5.26)$$

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

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 23 summarises the estimated coefficients β 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 $\text{PM}_{2.5}$. The estimated parameters of the latent variable and the uncertainty associated with them are $\hat{\theta} = 2.26^\circ$ ($std < 0.06^\circ$), $\hat{f} = 0.79$ ($std < 0.01$), and $\hat{a} = 0.18$ ($std < 0.01$). Note that the latent process is stationary and that the magnitude of g indicates that $z(s, t)$ changes smoothly over time. Finally, Figure 26 shows the daily $\hat{\sigma}_{\epsilon, t}^2$, 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 LOSOCV scheme. 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 RMSE is $3.85\mu\text{g}/\text{m}^3$, while the CV-RMSE is $5.91\mu\text{g}/\text{m}^3$, which improves the chemical transport model's performance (Veratti et al., 2023).

Name	β	std	$ t $	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 20: Estimate of the fixed effect coefficients of the model (5.7). $|t|$ is the absolute value of the t -statistic.

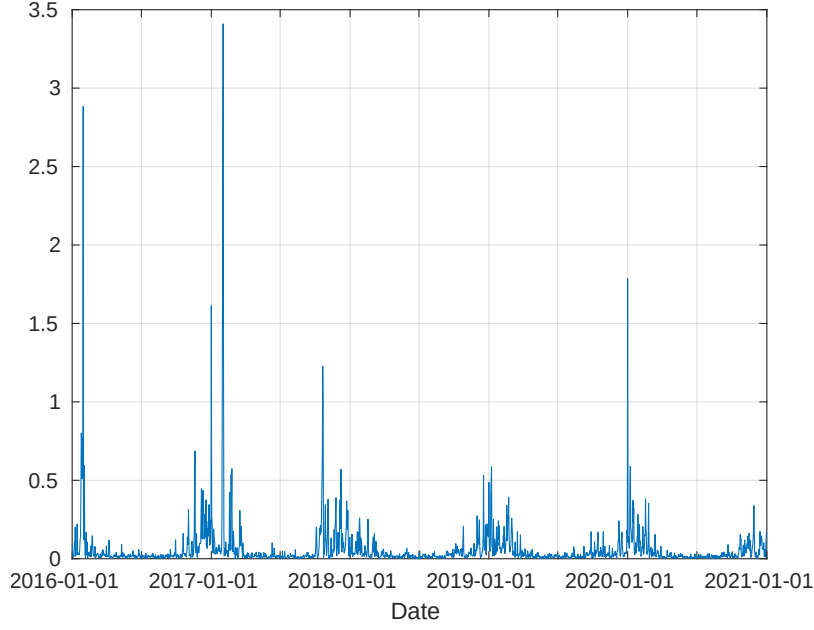


Figure 26: Estimated error variance, $\hat{\sigma}_{\epsilon,t}^2$.

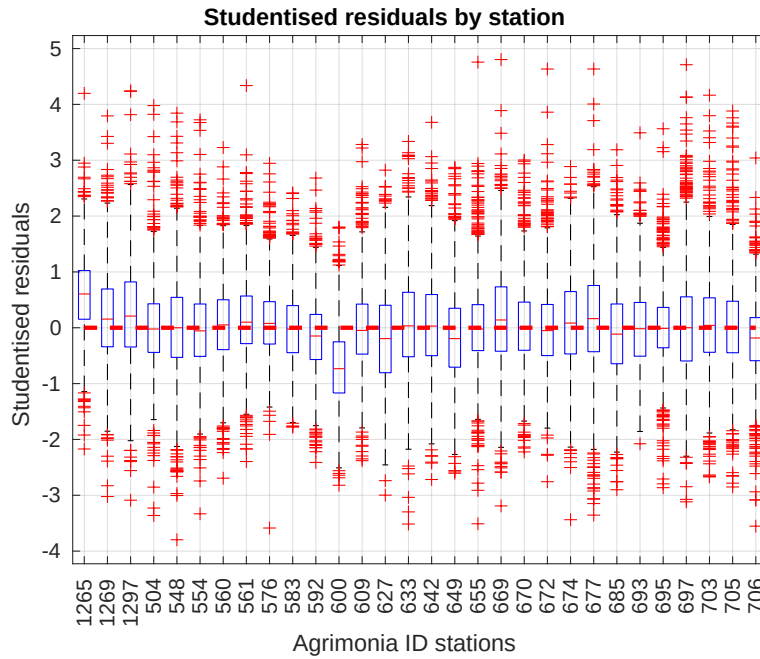


Figure 27: Studentised residuals grouped by station. The outliers (red cross) are identified by the Whisker length based on the first and third quartiles.

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 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.

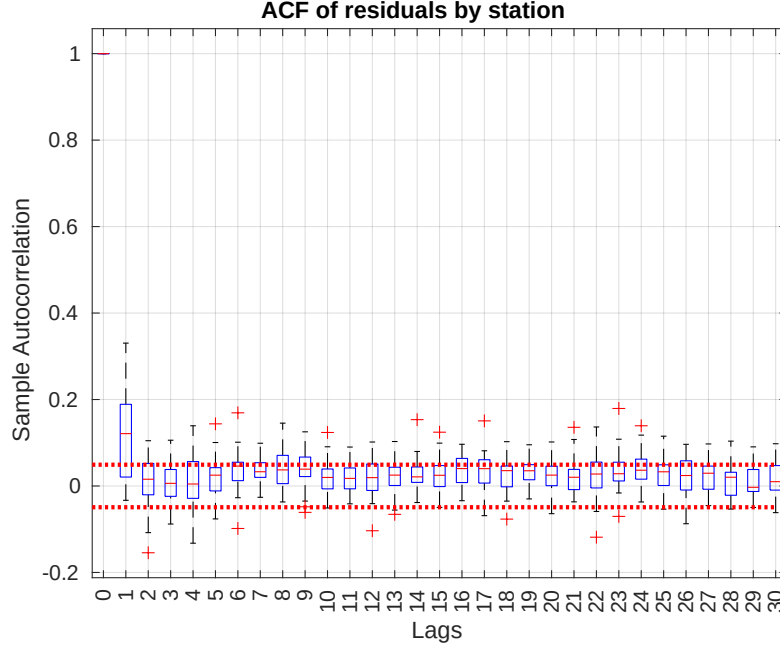


Figure 28: ACF computed by station residuals for the first 30 lags (days) and summarised through boxplots. The outliers (red cross) are identified by the Whisker length based on the first and third quartiles.

5.4.3 Scenario analysis

According to Veratti et al. (2023), we propose two different scenarios for NH_3 emission reduction: (i) a scenario named PRIA that accounts for a reduction of 26% in NH_3 emission, and (ii) a 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 ($\mu\text{g}/\text{m}^3$).

For each scenario, thanks to the discretisation in Eq. (5.22) and the approximation in Eq. (5.23), we map the average concentration and the uncertainty of the $\text{PM}_{2.5}$ changes due to NH_3 reductions on a regular grid of $0.1^\circ \times 0.1^\circ$ over the studied area. We focus on winter (452 daily predictions for each pixel), which is characterised by high $\text{PM}_{2.5}$ concentrations (see Table 14) and the largest effect of NH_3 (see Table 23). For winter, Figure 29 depicts the average $\text{PM}_{2.5}$ 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 $\text{PM}_{2.5}$ concentrations is close to $6 \mu\text{g}/\text{m}^3$. Considering the province’s plain average, Brescia and

Cremona, which are located in the southeast 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 NH₃ emissions, do not benefit from the reduction of ammonia emissions. The more marked reduction of NH₃ emissions (Strong scenario) leads to more significant improvements in air quality, achieving local PM_{2.5} reductions average close to 12 $\mu\text{g}/\text{m}^3$, reducing the risks associated with the population's health.

Furthermore, Figures 31 and 32 show the distribution of winter's daily PM_{2.5} reduction aggregated by province and land use for the two scenarios considered. As expected, the largest reductions are obtained in the 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 PM_{2.5} reduction.

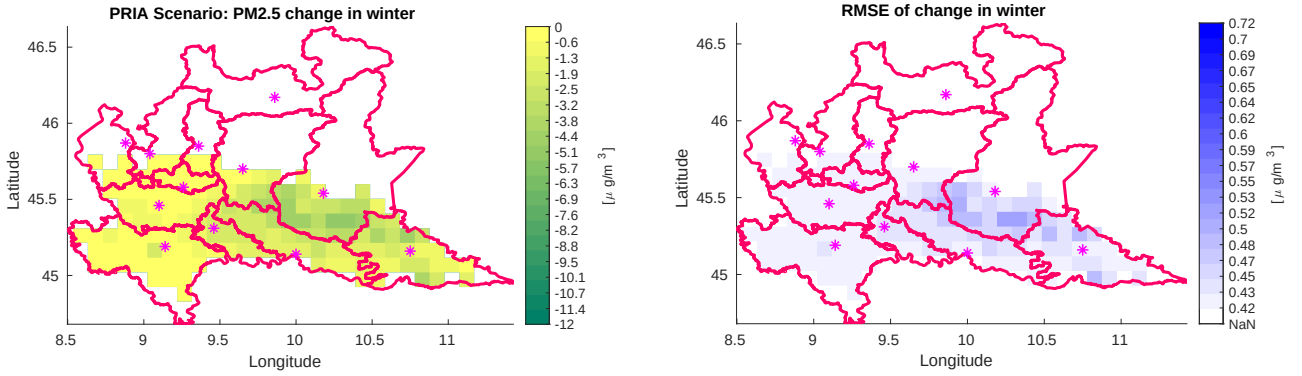


Figure 29: PRIA Scenario (-26%). Left: average PM_{2.5} reduction in winter, $\bar{\Delta}_{\hat{y}}$ (452 daily observations for each pixel). Right: uncertainty associated with the average reduction (std($\bar{\Delta}_{\hat{y}}$)). The pink stars depict the provincial capitals listed in Table 15. The model is only evaluated in non-forested areas below 640 m altitude.

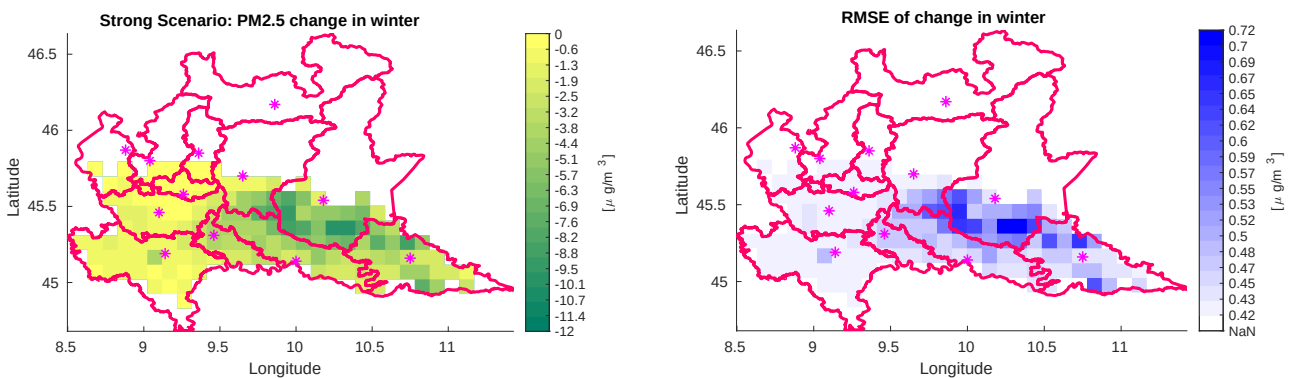


Figure 30: Strong Scenario (-50%). Left: winter $\bar{\Delta}_{\hat{y}}$ (452 daily observations for each pixel). Right: uncertainty associated with the average reduction (std($\bar{\Delta}_{\hat{y}}$)). The pink stars depict the provincial capitals listed in Table 15. The model is only evaluated in non-forested areas below 640 m altitude.

Table 21 summarises the average of PM_{2.5} reductions aggregated by province in $\mu\text{g}/\text{m}^3$. Note that, because of the presence of the Prealps and Alps, the scenario analysis covers a limited surface of

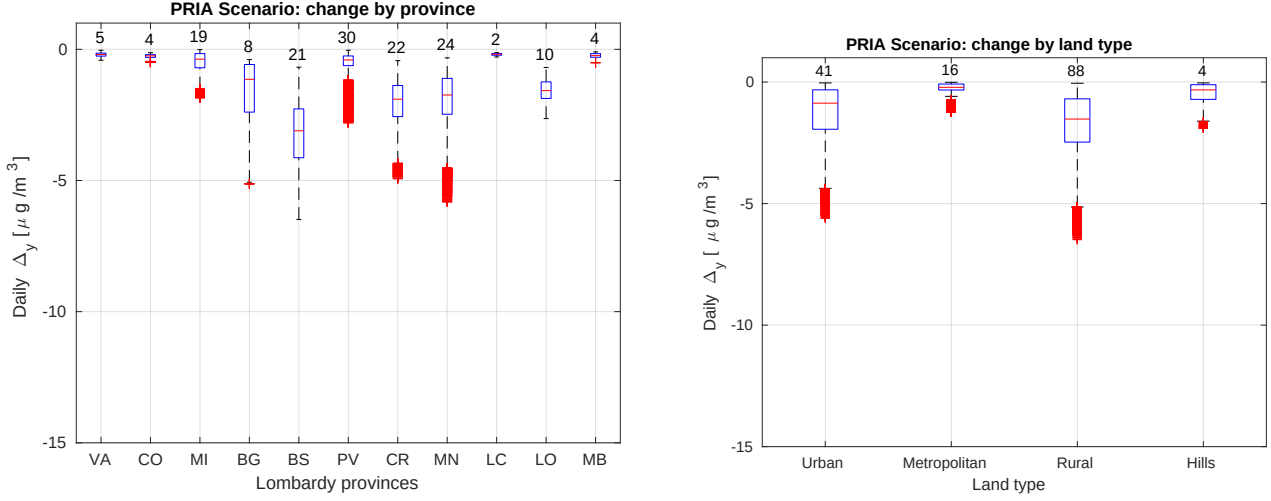


Figure 31: Scenario PRIA (-26%). Boxplot of daily Δ_y in winter (452 daily observations for each pixel). Left: grouped by province; Right: grouped by land type. The number of pixels of each category is reported over the corresponding boxplot. The model is evaluated only in the non-forested area under 640 m altitude.

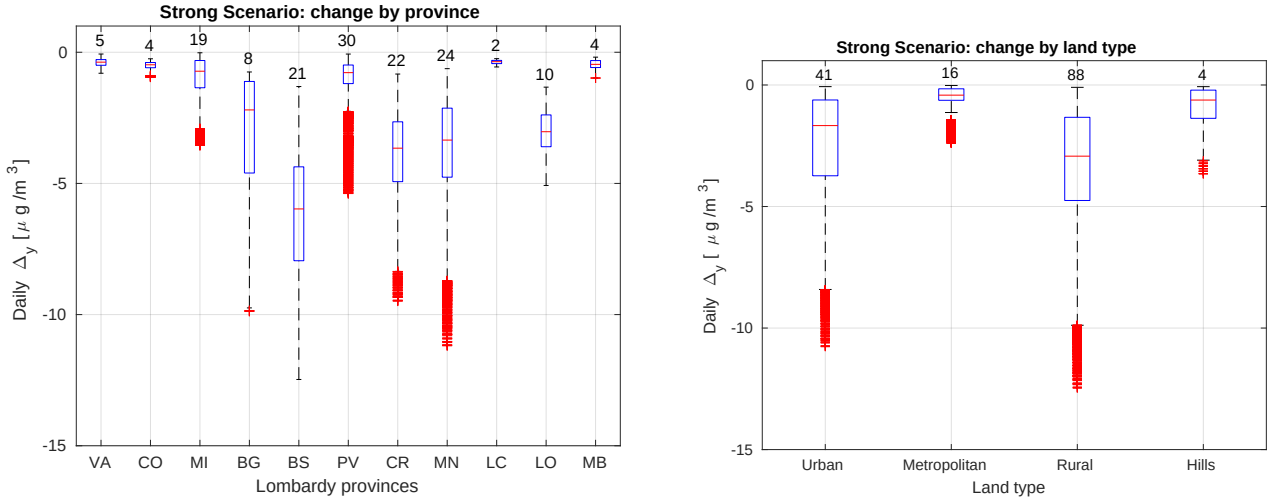


Figure 32: Scenario Strong (-50%). Boxplot of daily Δ_y in winter (452 daily observations for each pixel). Left: grouped by province; Right: grouped by land type. The number of pixels of each category is reported over the corresponding boxplot. The model is evaluated only in the non-forested area under 640 m altitude.

the Lecco, Varese, and Como provinces. Finally, Table 22 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 $\text{PM}_{2.5}$. We observe that, in spring, there is 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) and for the Strong scenario is 2.76 (std = 0.16). These results can be compared with the $\text{PM}_{2.5}$ estimated average of $39.49 \mu\text{g}/\text{m}^3$. Hence, considering the PRIA scenario, the $\text{PM}_{2.5}$ overall reduction is close to 3.5%, while considering the Strong scenario, the $\text{PM}_{2.5}$ overall reduction is close to 7%.

Province	$\mathcal{D}^*$	$\bar{y}$ $\mu\text{g}/\text{m}^3$	PRIA $\mu\text{g}/\text{m}^3$	Strong $\mu\text{g}/\text{m}^3$
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 21: 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 $\text{PM}_{2.5}$ 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	$\bar{y}$ $\mu\text{g}/\text{m}^3$	PRIA $\mu\text{g}/\text{m}^3$	$\%$	Strong $\mu\text{g}/\text{m}^3$	$\%$
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 22: $\text{PM}_{2.5}$ average change, $\bar{\Delta}_{\bar{y}}(s, t)$, by season and scenario. The number of pixels involved in the computation is $\mathcal{D}^* = 149$; $\bar{y}$ is the estimated average of the $\text{PM}_{2.5}$ 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.

5.4.4 Discussion of the results

We assess the $\text{PM}_{2.5}$ sensitivity to the NH_3 emissions through a geostatistical approach by using scenario analysis. The analysis is framed as a model-based “what-if” scenario rather than a formal counterfactual analysis, as causal identification assumptions are not explicitly enforced and the results should therefore be interpreted in terms of how model-based predictions change when the model inputs are modified, rather than as causal effects. We consider the heteroskedastic HDGM in Eq. (5.7); under the PRIA scenario, the overall $\text{PM}_{2.5}$ reduction is approximately 3.5%, while considering the Strong scenario, the overall $\text{PM}_{2.5}$ reduction is approximately 7%. The results are generally consistent with those obtained using 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 in 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 on where NH_3 emissions are very high. For this reason, we consider our results with caution and believe 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 $\text{PM}_{2.5}$ (number of available sensors, 45) and PM_{10} (number of available sensors, 101). Indeed the empirical cross-correlation between the $\text{PM}_{2.5}$ 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 LCM for the multivariate covariance of the innovation process η_t ; that is, $\Sigma_\eta = \mathbf{V}\rho(\|s - s'\|; \theta)$, where $\mathbf{V}$ is the $q \times q$ matrix of the coregionalisation, and q is the state process dimension, while the function $\rho(\cdot; \theta) : \mathbb{R} \rightarrow \mathbb{R}$ is a valid univariate correlation function.

The vector of the basis spline $\Psi_i(t)$ 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 23 summarises the estimated coefficients β 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

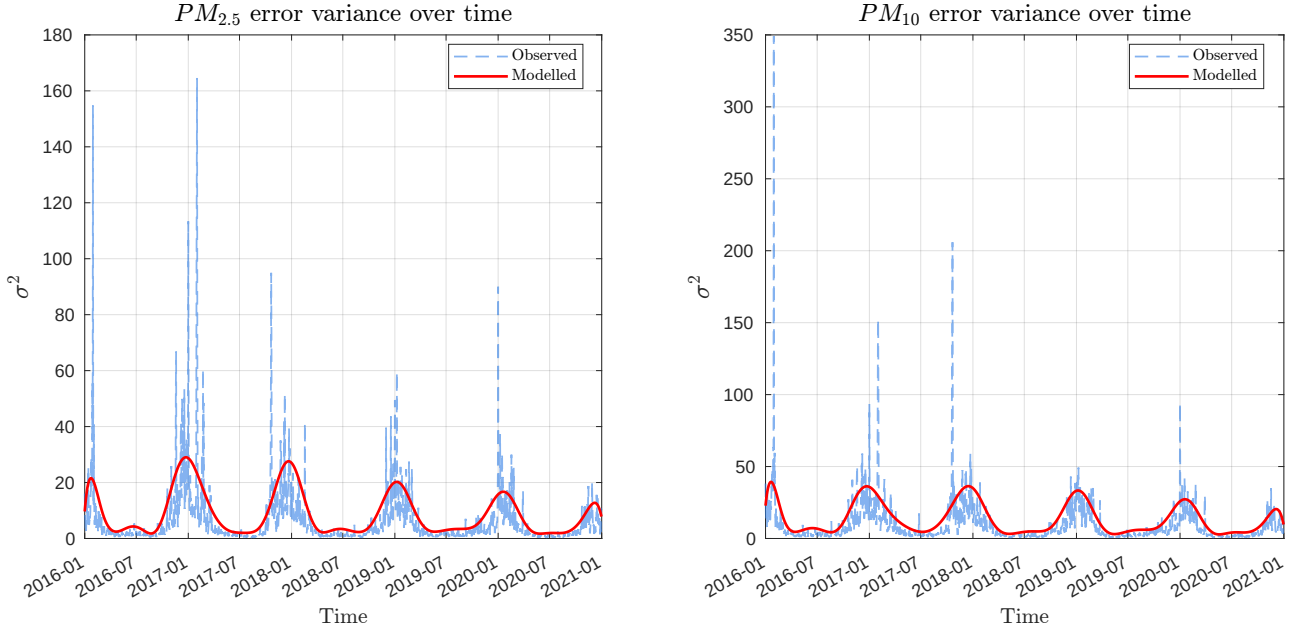


Figure 33: Left: $\hat{\sigma}_\epsilon(t)^2$ for the $PM_{2.5}$ (blue curve) compared with the empirical daily variance of the residual (orange line). Right: $\hat{\sigma}_\epsilon(t)^2$ for the PM_{10} (blue curve) compared with the empirical daily variance of the residual (orange line).

latent variable are $\hat{\theta} = 0.6^\circ$, $\hat{f}_1 = 0.77$, and $\hat{f}_2 = 0.78$, 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 23: Estimate of the fixed effect coefficients β of the bi-variate model.

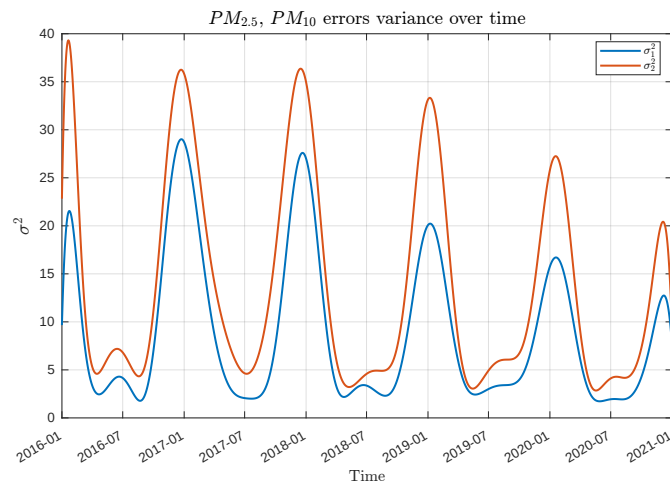


Figure 34: $\hat{\sigma}_\epsilon(t)^2$ for PM_{2.5} (blue curve) and PM₁₀ (orange curve).

Chapter 6

Multivariate low-rank state-space model with an 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 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 GMRFs. 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, this chapter will introduce a novel multivariate low-rank state-space model (LR-SSM) specifically 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 KF 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 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 briefly recalls the formulation of the Gaussian regression problem using the SSM framework. Section 6.4 presents 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 in Eq. (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 exhibit good performance and competitiveness in explaining daily $PM_{2.5}$ concentrations. See, for instance, Table 17 in Chapter 5, where prediction performance indices among three models are compared in cross-validation. Furthermore, a heteroskedastic SSM was proposed in Section 5.3 to assess the sensitivity of $PM_{2.5}$ to NH_3 .

To illustrate the computational advantage of using the KF on the SSM model in Eqs. (2.23)–(2.24) to address spatio-temporal regression, we compare the computational complexity of the univariate SSM to that of an evaluation of $E(y_t(s_0) \mid y_1(s), \dots, y_T(s))$ based on a generic joint Gaussian distribution of $(y_t(s_0), y_1(s), \dots, y_T(s))$. 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 $\mathbf{y}_{1:T} = (\mathbf{y}_1, \dots, \mathbf{y}_T)$ where $\mathbf{y}_t = (y(s_1, t), \dots, y(s_m, t))'$ is a realisation

of the process $y(\mathbf{s}, t)$ at time t and location set $\mathcal{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_m\}$, the calculation of $E(y_t(\mathbf{s}_0) \mid \mathbf{y}_{1:T})$, under the Gaussianity assumption, is given by the conditional expected values of Gaussian distribution which requires inverting the covariance matrix $\text{Cov}(\mathbf{y}_1^\top, \dots, \mathbf{y}_T^\top)$ of size $mT \times mT$ (see 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 $y(\mathbf{s}, t)$ simplifies to an iterative procedure for $t = \{1, \dots, T\}$. 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 the continuous-space 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 many approximation techniques, and a few of these are introduced in this section. While inference on the SSM using the KF 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 obser-

vation conditions only on a subset of others. Using $\pi(\mathbf{y}; \Pi)$ to indicate a generic density (possibly conditional), suppose that $\mathbf{y} = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_m))$ are observation of the spatial process $y(\mathbf{s})$ and has a joint density $\pi(\mathbf{y}; \Pi)$, where Π is a set of unknown parameters. Partition $\mathbf{y}$ into subvectors $\mathbf{y}_i \subset \mathbf{y}$ of possibly different lengths for $i = 1, \dots, K$ and define $\mathbf{y}_{(i)} = (\mathbf{y}_1, \dots, \mathbf{y}_i)$. We always have

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

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

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

The approximation in Eq. (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 $\mathbf{s}_{(j-1)}$ to be made up of the m observations that are nearest to $\mathbf{y}_j$. 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) processes is another likelihood-based method that offers a computationally efficient likelihood approximation for spatial statistical analysis. Considering the SRE model in Eq. (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

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

and the tapered covariance matrix $\tilde{\Sigma} = \Sigma(h; \vartheta) \odot \mathbf{K}_{tap}(h)$, where $\odot$ denotes the Hadamard (element-wise) product and $\mathbf{K}_{tap}$ is a sparse, positive definite tapering matrix. A popular choice for taper

functions is 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 $\tilde{\Sigma}$ 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 tapering 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 MSE 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 $\tilde{\Sigma}$ increases as γ decreases, 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 particular 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 $z(\mathbf{s})$ can be expressed, or at least approximated, through some finite basis expansion:

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

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 in Eq. (6.4), Example 6 considers the MLE of the SRE model in Eq. (2.16). Specifically, the computational cost for evaluating the log-likelihood $\log L(\mathbf{y} \mid \mathbf{X}; \Pi)$ 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).

Popular methods that fit into the low-rank approximations include the FRK (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, such as the predictive process method by Banerjee et al. (2008).

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

$$2 \log L(\mathbf{y} \mid \mathbf{X}; \Pi) = -n \log(2\pi) - \log |\Sigma(\vartheta) + \sigma_\epsilon^2 I_n| - (\mathbf{y} - \mathbf{X}'\boldsymbol{\beta})'(\Sigma(\vartheta) + \sigma_\epsilon^2 I_n)^{-1}(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}). \quad (6.5)$$

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

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

where $M = \boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}' + \sigma_\epsilon^2 I_n$. Using the Sherman–Morrison–Woodbury formula, $(M)^{-1}$ can be written as $M^{-1} = (\sigma_\epsilon^2)^{-1} I_n + (\sigma_\epsilon^2)^{-1} \boldsymbol{\Phi}(\mathbf{Q}^{-1} + (\sigma_\epsilon^2)^{-1} \boldsymbol{\Phi}'\boldsymbol{\Phi})((\sigma_\epsilon^2)^{-1} \boldsymbol{\Phi})'$, 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 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 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 process η_i , which corresponds to a GMRF representation of the original GP. Finn Lindgren, Håvard Rue, and Lindström (2011) gave the explicit connection between the GMRF and the Matérn GP; we refer the reader to Finn 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 LR-SSM designed for heterotopic observations.

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

We also assume that each $\eta_i(\mathbf{s}, t)$ 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 that controls the spatial range of dependence by $\theta_i = \sqrt{(8\nu)}/\kappa_i$. 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 $\boldsymbol{\epsilon}(\mathbf{s}, t)$, 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:

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

As we discussed in the introduction, however, the KF requires the inversion of an $m \times m$ matrix at every time $t = 1, \dots, T$, which results in a cost of $O(m^3T)$. While linear in time, KF does not scale properly in terms of the number of spatial locations m . For this reason, we present in Section 6.4 a low-rank SSM based on the SPDE approach (Finn Lindgren, Håvard Rue, and Lindström, 2011) that aims to approximate the data-generating process in Eqs. (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 (Finn Lindgren, Håvard Rue, and Lindström, 2011), designed to approximate the data-generating process in Eqs. (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 Eq. (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. The application of SPDE to SSM has been the subject of several previous studies, such as Ferkingstad, Held, and Håvard Rue, 2017. In addition, multivariate spatio-temporal coregionalization models closely related to ours have been studied, for example, in Gaedke-Merzhäuser et al., 2025 and Krainski et al., 2018. The model we propose is also connected to the spatio-temporal model introduced in Cameletti, Finn Lindgren, et al., 2013, although their formulation is strictly univariate. All these works are estimated with a Bayesian approach and rely on the Integrated nested Laplace approximations (INLA) (Håvard Rue, Martino, and Chopin, 2009) for estimation. Our proposal presents an SPDE-based SSM that leverages the EM estimation, including handling missing values, multivariate modelling, and an efficient Kalman filter implementation. This approach makes the multivariate SSM computationally efficient and suitable for high-dimensional datasets in both space and time.

To simplify the presentation, we present the resulting low-rank representation in this section, and defer the technical details to Appendix B.3.

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 $\mathcal{V}_R = \{\mathbf{r}_1, \dots, \mathbf{r}_R\}$ 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 $\boldsymbol{\psi}_R(\mathbf{s}) = (\psi_1(\mathbf{s}), \dots, \psi_R(\mathbf{s}))'$, where each $\psi_i : \mathbb{R}^2 \rightarrow \mathbb{R}$ is selected from the class of piecewise linear functions with compact support. These functions are constructed such that $\psi_i(\mathbf{r}_k) = 1$ if $i = k$, and $\psi_i(\mathbf{r}_k) = 0$ 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 $\eta_i(\mathbf{s}, t)$ exhibits the desired asymptotic properties, discussed in the sequel. A similar approach is presented in Finn 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:

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

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

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

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 $j = 1, \dots, q$), 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 $\mathbf{z}_j(\mathcal{V}_R, t)$ for $j = 1, \dots, q$ be the vector $\mathbf{z}_j(\mathcal{V}_R, t) = (z_j(\mathbf{r}_1, t), \dots, z_j(\mathbf{r}_R, t))'$ and $\mathbf{z}(\mathcal{V}_R, t)$ be the vector $\mathbf{z}(\mathcal{V}_R, t) = (\mathbf{z}_1(\mathcal{V}_R, t)', \dots, \mathbf{z}_q(\mathcal{V}_R, t)')'$.

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 Eq. (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:

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

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

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

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

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

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

Here, we summarise the advantages of using the model defined in Eqs (6.9)–(6.10) to make inference on Eqs. (2.23)–(2.24). First, the low-rank approximation reduces the computational complexity

of the KF from $O(m^3T)$ to $O(mR^2T)$, gaining efficiency for $R \ll m$. Therefore, the practitioner can 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 $z(\mathcal{V}_R, t)$ 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 $i = 1, \dots, q$; 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 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. For clarity, we first study the LR-SSM (6.9)–(6.10) for a fixed mesh, which provides a finite-dimensional representation of the model. This allows us to establish point identification in a distributional sense. We then investigate how the LR-SSM behaves as the mesh is refined: we show that the low-rank approximation of the latent process converges weakly to the continuous Matérn field and derive a uniform bound on the discrepancy between the outputs of the LR-SSM and those of the full continuous SSM. Together, these results quantify the approximation error and establish that it vanishes as the mesh resolution increases.

Assumption 1. *For a subset $\mathcal{S} \subset \mathcal{D}$, the limit $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \int_{\mathcal{S}} \mathbf{X}(s, t) ds = L_{\mathbf{X}}$ exists and is finite, and the matrix $L_{\mathbf{X}} L'_{\mathbf{X}}$ 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 β .

Assumption 2. (i) $f_1 > f_2 > \dots > f_q$; (ii) $\mathbf{W}$ has full column rank, with each column scaled so that its first non-zero element is positive; (iii) $\Psi_R(s) = \mathbf{I}_p \otimes \psi_R(s)'$ is known and has full row rank; (iv) The map $k \mapsto \mathbf{Q}_k$ is injective.

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 in Eqs. (6.9)–(6.10), with parameter set Π . Assume that the process $\{\mathbf{y}(\mathbf{s}, t) : (\mathbf{s}, t) \in \mathcal{S} \times \mathbb{N}\}$ 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 Π^* 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 Finn Lindgren, Håvard Rue, and Lindström (2011), we say $x^R \xrightarrow{D\{L^2(\mathcal{D})\}} x$ if for all $f, g \in L^2(\mathcal{D})$,

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

As $\mathcal{H}_R^1(\mathcal{D})$ is dense in $\mathcal{H}^1(\mathcal{D})$ as $R \rightarrow \infty$, Theorem 3 of Finn Lindgren, Håvard Rue, and Lindström (2011) ensures that for a fixed $t \in \{1, \dots, T\}$, it holds $\Psi(\mathbf{s})\boldsymbol{\eta}(\mathcal{V}_R, t) \xrightarrow{D\{L^2(\mathcal{D}^q)\}} \boldsymbol{\eta}(\mathbf{s}, t)$. Thanks to the Markovian structure of $\mathbf{z}(\mathbf{s}, t)$, 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. The following result is a direct corollary of Theorem 5 Finn Lindgren, Håvard Rue, and Lindström, 2011, where the convergence of each component implies convergence of the finite collection as a whole.

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

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

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

$$\mathbb{E}[\Psi_R(\mathbf{s})\mathbf{z}(\mathcal{V}_R, t)] \rightarrow \mathbb{E}[\mathbf{z}(\mathbf{s}, t)],$$

$$\text{Cov}(\Psi_R(\mathbf{s})\mathbf{z}(\mathcal{V}_R, t), \Psi_R(\tilde{\mathbf{s}})\mathbf{z}(\mathcal{V}_R, \tilde{t})) \rightarrow \text{Cov}(\mathbf{z}(\mathbf{s}, t), \mathbf{z}(\tilde{\mathbf{s}}, \tilde{t})).$$

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

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

where C is a constant that depends on $\|\mathbf{W}\|$ 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 3 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 $t \rightarrow \infty$. 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 $\|\mathbf{W}\|$ and q , since the matrix $\mathbf{W}$ modulates how the approximation error introduced by replacing $\mathbf{z}$ with $\mathbf{z}^R$ propagates to the observable output $\mathbf{y}$.

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

Theorem 3. *Under Assumptions 1–2, for a fixed R , the bias in the estimation of the vector parameters $\boldsymbol{\pi} = (\boldsymbol{\beta}', \boldsymbol{\sigma}', \text{vec}(\mathbf{W})', \mathbf{f}', k_1, \dots, k_q)'$ 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,*

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

where $\boldsymbol{\pi}_R^\star$ is the pseudo-true parameter vector minimising the Kullback-Leibler divergence (D_{KL}) between the true marginal law $\mathcal{P}_\pi$ of $\mathbf{y}(\mathbf{s}, t)$ with parameter vector $\boldsymbol{\pi}$ and the misspecified LR-SSM

marginal law $\mathcal{P}_{\vartheta}^{(R)}$ of $\mathbf{y}^R(s, t)$ with vector of parameter ϑ , i.e. $\pi_R^* = \operatorname{argmin}_{\vartheta} D_{KL}(\mathcal{P}_{\vartheta}^{(R)} || \mathcal{P}_{\pi})$. Furthermore, the parameter β has no bias.

The proof is reported in Appendix B.4.

Together, Theorems 1–3 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.

Discussion on the theoretical results

While we do not provide a formal proof of consistency for the ML estimator under the LR-SSM approximation, Theorems 1–3 offer an intuitive justification for why consistency is plausible. Theorem 1 shows that, as the mesh is refined ($h \rightarrow 0$), the LR-SSM reproduces the joint distribution of the true SSM, while Theorem 3 establishes that the pseudo-true parameter, the target of the ML estimator under misspecification, converges to the true parameter at rate $O(h)$. Taken together, these results suggest that, if the approximation error vanishes and the amount of spatial and/or temporal information grows sufficiently, the ML estimator should converge to the true parameter vector. However, establishing such a consistency result rigorously requires dealing with the combined challenges of misspecified likelihoods, the state-space structure of the model, and assumptions on the sampling regime, and is therefore beyond the scope of this work.

The SPDE-based approximation of the Matérn GP relies on several assumptions that are largely technical rather than fundamental. Following Finn Lindgren, Håvard Rue, and Lindström (2011) and Finn Lindgren, Bolin, and Håvard Rue (2022), the approach assumes that the latent field is a weak solution to a linear elliptic SPDE defined on a sufficiently regular domain, typically a bounded subset of $\mathbb{R}^d$ with appropriate boundary conditions, ensuring existence and uniqueness of the solution. The Matérn covariance structure arises from a differential operator of the form $(\kappa^2 - \Delta)^{\alpha/2}$, whose properties guarantee Gaussianity and computational tractability and result in a (non-stationary) Matérn field with smoothness parameter $\nu = \alpha + d/2$. However, more general formulations are possible. In particular, both the range parameter κ^2 and the driving noise $\mathcal{W}$ may be allowed to vary spatially as functions of the coordinate s , leading to non-stationary SPDE models (Finn Lindgren, Håvard Rue, and Lindström, 2011).

While such models induce globally non-stationary covariance structures, the resulting dependence remains locally isotropic. This restriction can be relaxed by extending the class of SPDEs to include anisotropic differential operators, for example, by replacing the Laplacian with a non-isotropic operator

or by incorporating directional derivative terms (Finn Lindgren, Bolin, and Håvard Rue, 2022). These extensions establish a direct link with the deformation approach to non-stationary covariance modelling introduced by Sampson and Guttorp (1992), in which the spatial domain is mapped into a latent space where the field is stationary. Within the SPDE framework, this construction can be interpreted as defining a non-stationary SPDE on the original domain, while retaining the computational advantages of the GMRF representation.

The finite-dimensional approximation (3.5) is obtained using locally supported, piecewise linear finite element basis functions defined on a conforming triangulation, leading to sparse precision matrices. While homogeneous Neumann or Dirichlet boundary conditions are commonly imposed for convenience, alternative formulations are possible. Many of these assumptions can be relaxed in future developments, including adaptive or higher-order finite elements, and domains such as manifolds or metric graphs, where the Laplace operator is replaced by the Laplace–Beltrami operator or graph Laplacian (Finn Lindgren, Bolin, and Håvard Rue, 2022). These extensions preserve the core advantages of the SPDE framework while broadening its applicability.

6.5 Parameters estimation

We estimate the parameter set Π using the 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 $y_j(\mathbf{s}, t)$ at spatial locations $\mathcal{S}_{j,t} = \{\mathbf{s}_{i,t}\}_{i=1}^{m_{j,t}}$.

We assume that the mesh $\mathcal{G}_R$, which defines the spatial domain for the latent process within the LR-SSM described in Eqs. (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, and the matrix formulation of the model are presented in Appendix B.5. We denote by $\mathbf{Q}_{\kappa_i}^{-1}$ 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 $\mathbf{Q}_{\kappa_i}$ is constructed at a linear cost using the FEM (Finn Lindgren, Håvard Rue, and Lindström, 2011). Thus, the evaluation of the objective function in the maximization step, which involves computing the log-determinant of $\mathbf{Q}_{\kappa_i}$, benefits on the one hand from the sparsity of the matrix and, on the other, from the low-rank approximation that limits its effective dimension, making the overall optimization step computationally fast.

Note that given the MLE of the parameter set Π , the plug-in predictions at the new site $\mathbf{s}_0 \in \mathcal{D}$

and time $t = 1, \dots, T$ are given by

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

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

where $\hat{\mathbf{z}}^T(\mathcal{V}_R, t)$ denotes the smoothed state process $\mathbf{z}(\mathcal{V}_R, t)$, with associated variance $\text{Var}[\hat{\mathbf{z}}^T(\mathcal{V}_R, t)]$, estimated by the KS variance $\mathbf{P}_t^T$ (see the KS derivation in Appendix B.5). The vectors $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{W}}$ are the MLE of $\boldsymbol{\beta}$ and $\mathbf{W}$, respectively. Equation (6.11) results in the BLUP procedure, and it relies only on linear algebra operations with a linear computational cost of $\mathcal{O}(R)$. Figure 35 offers a visual overview of the main methodological components and summarises the full workflow of the proposed method.

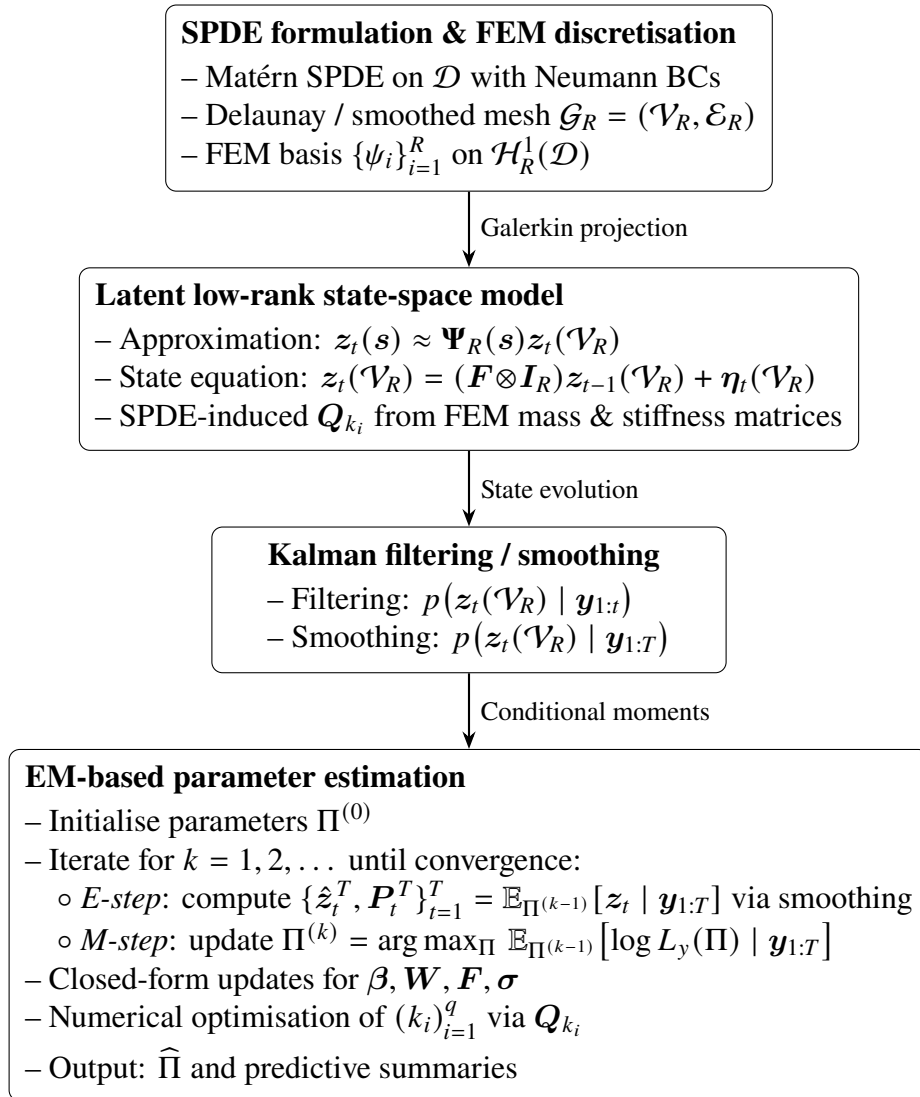


Figure 35: Schematic workflow of the proposed low-rank SPDE state-space model. The SPDE is discretised via FEM to construct a low-rank latent state-space representation, temporal dependence is handled through Kalman filtering and smoothing, and parameters are estimated via an EM algorithm.

6.5.1 Choice of the mesh

Theorem 1 establishes the weak convergence $\Psi_R(s)z(\mathcal{V}_R, t) \rightarrow z(s, t)$. 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 $\geq \theta_{\min}$

```

1 while any triangle angle  $< \theta_{\min}$  do
2   foreach interior vertex  $r \in \mathcal{V}_R$  do
3     Let  $\{r_1, \dots, r_k\}$  be the neighbours of  $r$ ;
4     Compute  $r^* \leftarrow \frac{1}{k} \sum_{i=1}^k r_i$ ;
5     if  $r^*$  preserves Delaunay and connectivity then
6       Update position:  $r \leftarrow r^*$ ;
7     else
8       Remove  $r$ ; locally retriangulate;
9       Insert vertex at  $r^*$ ;

```

6.5.2 Approximation of the precision matrix

In this section, we briefly recall how the approximate precision matrix $\mathbf{Q}$ is obtained within the SPDE framework of Finn 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 $\mathbf{C}$ and $\mathbf{G}$ denote the mass matrix and stiffness matrix, respectively, defined as

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

We then define the matrix

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

Using the Neumann boundary conditions, the finite-dimensional representation of the solution to Eq. (B.25) has precision:

$$\mathbf{Q}_{k_i} = \mathbf{K}_{\kappa_i^2} \mathbf{C}^{-1} \mathbf{K}_{\kappa_i^2}. \quad (6.13)$$

The matrix $\mathbf{C}^{-1}$ is dense, which leads to a dense precision matrix in the resulting model. To address this issue, Finn Lindgren, Håvard Rue, and Lindström (2011, Appendix C.5) showed that $\mathbf{C}$ can be approximated by a diagonal matrix $\tilde{\mathbf{C}}$, where $\tilde{C}_{ii} = \langle \psi_i, 1 \rangle$. 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, $\tilde{\mathbf{C}}$, $\mathbf{G}$, and $\mathbf{K}_{\kappa_i^2}$

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 $\mathbf{Q} = \text{blockdiag}(\mathbf{Q}_{k_1}, \dots, \mathbf{Q}_{k_q})$.

The result in Eq. (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 SPDE induces a folding effect in the resulting covariance function. This behaviour, analysed in detail by Finn Lindgren, Håvard Rue, and Lindström (2011), implies that the covariance of the SPDE solution on a bounded domain closely matches the stationary Matérn covariance whenever locations lie at least two Matérn ranges away from the boundary. To exploit this property, we enlarge the computational domain by adding a set of auxiliary vertices outside the region of interest. Specifically, we construct an augmented mesh

$$\mathcal{V}_R^* = \mathcal{V}_R \cup \mathcal{V}_R^{\text{aux}},$$

where the auxiliary set $\mathcal{V}_R^{\text{aux}}$ is chosen so that every vertex in $\mathcal{V}_R$ lies at least twice the Matérn range parameter away from the outer boundary of $\mathcal{V}_R^*$.

On this extended mesh we construct the SPDE-based precision matrix $\mathbf{Q}_{k_i}^*$ as in (6.13), imposing Neumann boundary conditions on the outer boundary of $\mathcal{V}_R^*$. We then partition $\mathbf{Q}_{k_i}^*$ according to the vertices in $\mathcal{V}_R$ and the auxiliary vertices in $\mathcal{V}_R^{\text{aux}}$,

$$\mathbf{Q}_{k_i}^* = \begin{pmatrix} \mathbf{Q}_{\mathcal{V}_R, \mathcal{V}_R}^* & \mathbf{Q}_{\mathcal{V}_R, \mathcal{V}_R^{\text{aux}}}^* \\ \mathbf{Q}_{\mathcal{V}_R^{\text{aux}}, \mathcal{V}_R}^* & \mathbf{Q}_{\mathcal{V}_R^{\text{aux}}, \mathcal{V}_R^{\text{aux}}}^* \end{pmatrix}.$$

The precision matrix corresponding to the SPDE solution restricted to the interior vertices $\mathcal{V}_R$ is obtained by marginalising out the auxiliary vertices, resulting in the Schur complement

$$\mathbf{Q}_{k_i} = \mathbf{Q}_{\mathcal{V}_R, \mathcal{V}_R}^* - \mathbf{Q}_{\mathcal{V}_R, \mathcal{V}_R^{\text{aux}}}^* (\mathbf{Q}_{\mathcal{V}_R^{\text{aux}}, \mathcal{V}_R^{\text{aux}}}^*)^{-1} \mathbf{Q}_{\mathcal{V}_R^{\text{aux}}, \mathcal{V}_R}^*.$$

This matrix is the exact marginal precision of the SPDE solution restricted to the domain of interest. Although $\mathbf{Q}_{k_i}$ is dense, its dimension is only $R \ll m$ thanks to the low-rank construction, so the resulting computational cost remains negligible within the overall LR-SSM framework.

6.5.3 Asymptotic distribution and Wald confidence intervals

For inference on the estimated parameters, we assume that the true data-generating process follows the SSM in Eqs. (2.23)–(2.24), while the proposed low-rank formulation in Eqs (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) described a simple bootstrap sampling algorithm for parameter estimates in general SSM that is appropriate for the spatio-temporal setting discussed here. In addition, Wall and Stoffer (2002) described 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 MLE $\hat{\Pi}$ is asymptotically normal. For precise assumptions in the context of SSM and EM-based estimation, see Jensen and Petersen (1999) and McLachlan and Krishnan (2008), respectively. Then:

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

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

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

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

$$\hat{\mathbf{V}} = \hat{\mathcal{I}}^{-1}. \quad (6.15)$$

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

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

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 the SSM in Eqs. (2.23)–(2.24) using the parameter vector denoted by Π^0 . The inference is carried out using the proposed LR-SSM in Eqs. (6.9)–(6.10). Indeed, Theorem 2 shows that the approximation error when using the LR-SSM for inference on a data-generating process from the SSM 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 LR-SSM defined by Eqs. (6.9)–(6.10), parametrised by Π^0 , synthetic data are generated using the SSM defined by Eqs. (2.23)–(2.24). Let $\mathcal{L}_\delta = \{(i\delta, j\delta) \in \mathbb{R}^2 : i, j \in \{0, 1, \dots, 24\}\}$ 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 $i \in \{1, \dots, p\}$, 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 for each variable. These choices ensure that the simulated sampling regimes are comparable to those encountered in the real-data application, where the heterotopic structure leads to an effective number of observations per component that is similar in magnitude to the simulated values of m . The LR is set to 100%, 75%, 50%, 25%, and 15%. As an example, Figure 36 shows the generated finite element mesh for the simulations when $m = 100$. Except for the case where $LR = 100\%$, 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.

In summary, for each rank R , the simulation campaign is based on six different simulation setups.

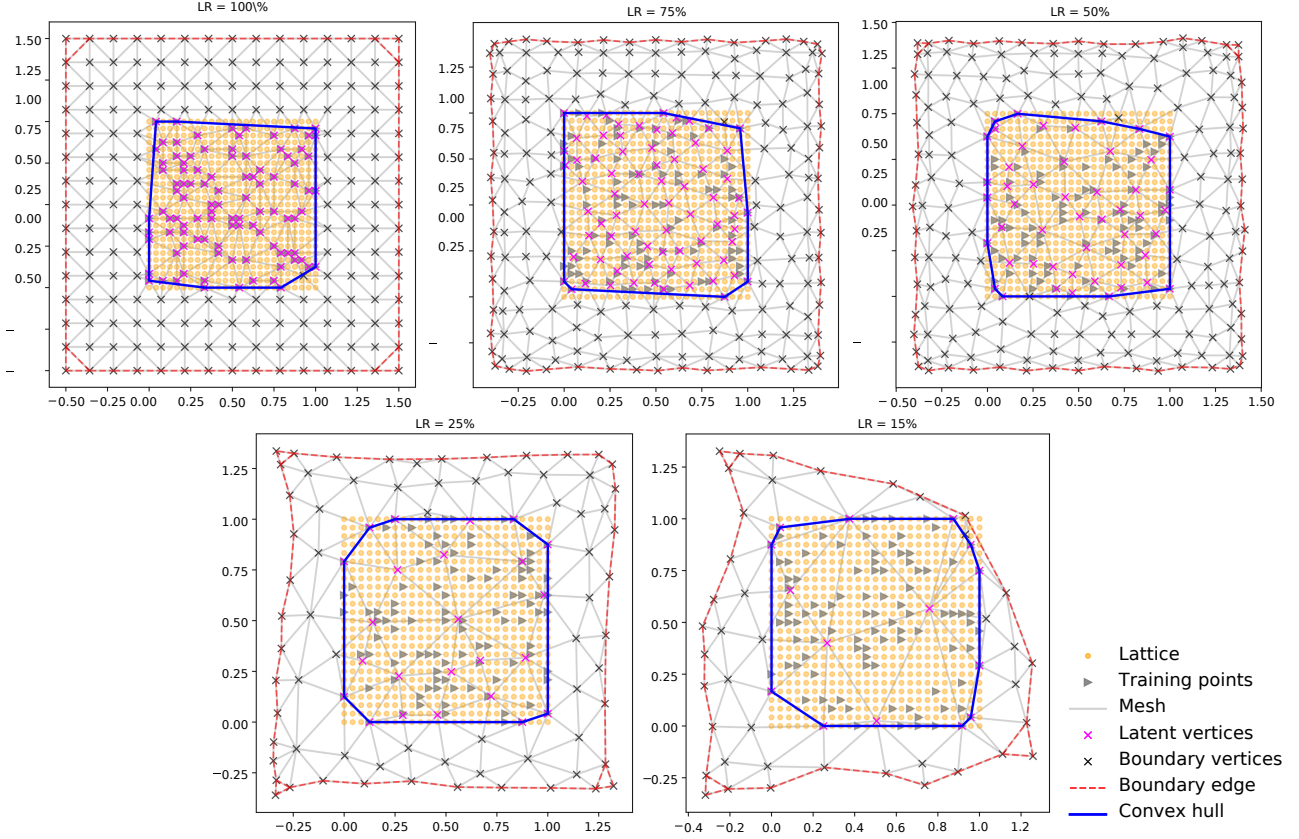


Figure 36: Simulated $m = 100$ spatial points location on the domain $\mathcal{D} = [0, 1]^2$, along with the resulting finite element mesh for $LR = 100\%, 75\%, 50\%, 25\%$, and 15% (from left to right, top to bottom). Laplacian smoothing is performed, providing a refined vertices (magenta crosses) 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.

For each simulation setup, $M = 100$ Monte Carlo replications are performed. The data-generating process is specified by $\Pi^0 = \{\beta^0, \sigma^{2,0}, \mathbf{f}^0, \mathbf{W}^0, k_1^0, \dots, k_q^0\}$ with the following parameter values: $\beta^0 = (1, 2, -1)'$ and

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

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

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

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

6.6.2 Estimation procedure

Once a single dataset j , for $j = 1, \dots, M$ 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 β , a random point in $(0.2, 0.8) \times (-0.2, -0.8)$ for f , and each element of 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 $|\log L_{k+1} - \log L_k| / |\log L_k| < 0.0001$, where $\log L_k$ denotes the log-likelihood at iteration k .

The implementation of the proposed algorithm is carried out in Python and relies on the NumPy, SciPy, and JAX packages for efficient numerical computation. Finite element calculations are performed using the MFEM library Andrej et al., 2024, which provides a flexible, high-performance framework for assembling variational forms. Through MFEM's abstractions, standard differential operators, such as the gradient, divergence, and Laplacian, are assembled efficiently, enabling the construction of stiffness and mass matrices in a unified and scalable manner. Mesh quality diagnostics, such as the distribution of minimum internal angles, are essential for evaluating the robustness of our meshing procedure and, consequently, the stability of the finite element discretisation. In the current implementation, mesh quality is monitored by imposing a lower bound on the minimum internal angle of each triangle in the mesh. The initial triangulation is generated using the SciPy Delaunay routine, while the Laplacian smoothing step has been implemented manually to improve mesh regularity.

6.6.3 Simulation results

We evaluate the models by comparing their parameter estimates using the bias and the RMSE where, for each $\theta \in \Pi$, the RMSE is computed as $\sqrt{N^{-1} \sum_{i=1}^M \|\hat{\theta}^i - \theta_0\|^2}$, where $\hat{\theta}^i$ corresponds to the estimation of the parameter θ in the i th Monte Carlo simulation, $\|\cdot\|$ is the euclidean norm, and θ_0 is the true parameter value. In addition, we assess predictive performance on the validation set S_i^* for $i = 1, \dots, p$, by comparing the observed process $y_i(s, t)$ with its estimate $\hat{y}_i(s, t)$. Prediction errors are defined as $e_i(s, t) = y_i(s, t) - \hat{y}_i(s, t)$, for $s \in S_i^*$. The different LR-SSM versions are then

compared in terms of the RMSE computed over these validation errors.

Tables 24, 25, and 26 present the bias and RMSE of the estimated parameters under each combination of m and T for $LR = 100\%$, 75% , and 15% , respectively. 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 β has an RMSE close to zero, improving with increasing observed data both in T and m . The noise variances σ^2 show more pronounced improvement with increasing m . Spatial range parameters κ are more challenging to estimate accurately; although the 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). The ML estimator of f has small RMSE, and the accuracy improves with increasing time length T . Regarding the accuracy of the estimate of W , we observe a small bias that is consistent with the issues in estimating the parameter κ , which suffers from low identifiability. Indeed, the loading matrix tends to compensate for the effect of the range parameter. Despite the bias affecting these parameters, the training RMSE improves, and the model also shows better performance in terms of test RMSE.

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 can 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 where $LR = 100\%$. This demonstrates that the proposed approximation procedure substantially reduces computation time while maintaining consistent parameter estimates and accurate spatial mapping performance. The computation time relates to a machine equipped with an Intel Xeon Platinum 8460Y+ CPU and 1007 GB RAM using Python 3.10.15.

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 PM with aerodynamic diameter less than $10\ \mu\text{m}$ (PM_{10}) and

Parameter	T = 50						T = 100					
	m = 100		m = 225		m = 400		m = 100		m = 225		m = 400	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
β_1	0.00		0.00		0.00		0.00		0.00		0.00	
β_2	0.00	0.02	0.00	0.02	0.00	0.01	0.00	0.02	0.00	0.01	0.00	0.01
β_3	-0.01		0.00		0.00		0.00		0.00		0.00	
σ_1^2	0.13		0.06		0.02		0.12		0.06		0.02	
σ_2^2	0.28	0.20	0.16	0.11	0.09	0.06	0.28	0.19	0.15	0.10	0.08	0.05
σ_3^2	0.12		0.06		0.02		0.11		0.05		0.02	
k_1	-0.77		-2.10		-3.50		-0.71		-2.42		-3.21	
k_2	-0.21	2.27	-0.63	3.44	-0.81	3.05	-0.33	2.23	-0.51	2.49	-0.79	2.94
W_1	0.05		0.03		0.05		0.09		0.06		0.06	
W_2	-0.03		0.00		0.03		0.00		0.02		0.07	
W_3	-0.01	0.12	-0.05	0.15	-0.10	0.21	0.03	0.12	-0.02	0.15	-0.07	0.21
W_4	-0.01		0.04		0.14		0.00		0.06		0.15	
W_5	0.05		0.06		0.11		0.07		0.08		0.10	
W_6	-0.04		-0.04		-0.06		-0.02		-0.02		-0.02	
f_1	-0.04	0.10	-0.05	0.09	-0.09	0.14	-0.05	0.08	-0.06	0.08	-0.08	0.12
f_2	-0.02		-0.02		-0.01		0.00		0.00		0.00	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
$RMSE_{train}$	1.02	0.02	0.99	0.01	0.97	0.01	1.02	0.01	0.98	0.01	0.96	0.01
$RMSE_{test}$	1.22	0.01	1.17	0.01	1.13	0.01	1.21	0.01	1.15	0.01	1.12	0.01
Min angle	9.40	1.25	10.51	0.82	10.86	0.26	9.27	1.20	10.53	0.75	10.85	0.28
Time	88.47	25.64	276.32	84.47	818.33	378.96	99.70	37.58	366.46	120.32	1,242.62	574.37

Table 24: $LR = 100\%$. Bias and RMSE of the estimated parameters are reported for different numbers of time steps (T) and spatial locations (m). $RMSE_{train}$ and $RMSE_{test}$ 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. *Avg* and *Std* denote the average and standard deviation, respectively, across replications. Min angle indicates the average of the minimum internal angles across all the Monte Carlo simulations, measured in degrees, while Time indicates the average computation time over replications in seconds.

fine PM with diameter less than $2.5 \mu\text{m}$ ($PM_{2.5}$) in Italy. These pollutants are typically observed at partially different monitoring stations (heterotopic) and are known to be highly correlated. Fassò, Finazzi, and Ndong (2016) demonstrated 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.

6.7.1 Data description

We used 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 1 July 2018 to 30 December 2023. The AQCLIM-GRINS dataset extends the AGRIMONIA

Parameter	T = 50						T = 100					
	m = 100		m = 225		m = 400		m = 100		m = 225		m = 400	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
β_1	0.00		0.00		0.00		0.00		0.00		0.00	
β_2	0.00	0.03	0.00	0.02	0.00	0.01	0.00	0.02	0.00	0.01	0.00	0.01
β_3	0.00		0.00		0.00		0.00		0.00		0.00	
σ_1^2	0.42		0.27		0.21		0.41		0.26		0.19	
σ_2^2	0.32	0.32	0.22	0.21	0.16	0.16	0.34	0.32	0.23	0.20	0.15	0.14
σ_3^2	0.15		0.09		0.05		0.15		0.08		0.05	
k_1	2.63	4.94	-0.11	3.48	-4.01	4.50	3.02	3.95	0.93	4.70	-3.43	3.52
k_2	0.65		-0.55		-0.65		0.19		-0.60		-0.59	
W_1	0.15		0.08		0.09		0.16		0.07		0.10	
W_2	-0.01		-0.01		0.01		0.03		0.05		0.03	
W_3	0.11	0.15	-0.02	0.18	-0.08	0.22	0.13	0.15	0.03	0.16	-0.04	0.22
W_4	0.00		0.07		0.15		0.01		0.06		0.13	
W_5	0.11		0.12		0.15		0.11		0.09		0.14	
W_6	-0.02		-0.05		-0.08		0.01		0.02		-0.05	
f_1	-0.06	0.11	-0.06	0.11	-0.09	0.12	-0.04	0.08	-0.05	0.09	-0.09	0.13
f_2	-0.01		-0.02		-0.04		0.01		-0.01		-0.02	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
RMSE _{train}	1.09	0.01	1.05	0.01	1.02	0.01	1.08	0.01	1.04	0.01	1.02	0.01
RMSE _{test}	1.22	0.01	1.17	0.01	1.13	0.01	1.22	0.01	1.16	0.00	1.12	0.01
Min angle	11.40	3.50	10.67	1.12	10.68	0.67	11.74	3.48	10.61	0.98	10.74	0.56
Time	36.42	8.61	101.50	19.31	273.86	76.99	33.63	8.23	130.17	35.38	373.08	156.46

Table 25: $LR = 50\%$. 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. *Avg* and *Std* denote the average and standard deviation, respectively, across replications. Min angle indicates the average of the minimum internal angles across all the Monte Carlo simulations, measured in degrees, while Time indicates the average computation time over replications in seconds.

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 PM data consists of daily average concentrations (in $\mu\text{g}/\text{m}^3$) collected at 661 ground-level monitoring stations irregularly distributed across Italy. Moreover, we included the following covariates: altitude (in metres above sea level), temperature at 2 metres (in $^\circ\text{C}$), relative humidity (RH; in %), and wind speed (in m/s), along with an intercept term. Table 27 summarises the variables used in this study and the associated main statistics.

It can be noticed that the resulting measurement network in Figure 37 is partially heterotopic, meaning that 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 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.

Parameter	T = 50						T = 100					
	m = 100		m = 225		m = 400		m = 100		m = 225		m = 400	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
β_1	0.00		0.00		0.00		0.00		0.00		0.00	
β_2	0.01	0.03	0.00	0.02	0.00	0.01	0.00	0.02	0.00	0.01	0.00	0.01
β_3	0.00		0.00		0.00		0.00		0.00		0.00	
σ_1^2	0.76		0.54		0.41		0.75		0.52		0.39	
σ_2^2	0.52	0.56	0.39	0.40	0.31	0.31	0.54	0.56	0.40	0.39	0.30	0.29
σ_3^2	0.27		0.17		0.13		0.26		0.16		0.11	
k_1	-0.27		4.16		4.80		0.23		4.05		5.55	
k_2	0.86	3.12	0.68	5.50	0.42	7.49	0.72	2.35	0.34	5.63	0.42	6.92
W_1	0.77		0.57		0.45		0.86		0.59		0.43	
W_2	0.08		0.13		0.15		0.10		0.21		0.19	
W_3	0.34		0.29		0.19		0.39		0.31		0.21	
W_4	0.05	0.47	0.12	0.38	0.19	0.37	0.05	0.51	0.16	0.41	0.19	0.35
W_5	0.57		0.42		0.35		0.64		0.42		0.32	
W_6	0.06		0.07		0.05		0.07		0.12		0.09	
f_1	-0.37		-0.22		-0.18		-0.41		-0.21		-0.17	
f_2	0.08	0.31	0.06	0.19	0.01	0.18	0.08	0.33	0.07	0.20	0.03	0.17
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
RMSE _{train}	1.20	0.01	1.14	0.01	1.10	0.01	1.20	0.01	1.14	0.01	1.10	0.01
RMSE _{test}	1.27	0.01	1.21	0.01	1.16	0.01	1.27	0.01	1.20	0.01	1.16	0.01
Min angle	19.34	3.88	15.70	4.47	13.98	4.56	19.57	3.61	15.79	4.39	13.75	4.22
Time	14.07	5.55	23.27	6.98	59.05	22.03	11.46	4.02	28.12	9.03	81.01	31.13

Table 26: $LR = 15\%$. 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. *Avg* and *Std* denote the average and standard deviation, respectively, across replications. Min angle indicates the average of the minimum internal angles across all the Monte Carlo simulations, measured in degrees, while Time indicates the average computation time over replications in seconds.

6.7.2 Preliminary analysis

Figure 37 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 PM variables. This is confirmed by Figure 38 reporting the weighted least squares Matérn variograms and covariograms 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 $2.88 \mu\text{g}/\text{m}^3$ for PM_{10} and $2.24 \mu\text{g}/\text{m}^3$ for $PM_{2.5}$. Accordingly, the model of the next section includes a measurement error component as specified in Eq. (6.10). The estimated marginal variances of the spatial process are approximately 15.4 and $6.8 (\mu\text{g}/\text{m}^3)^2$. 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

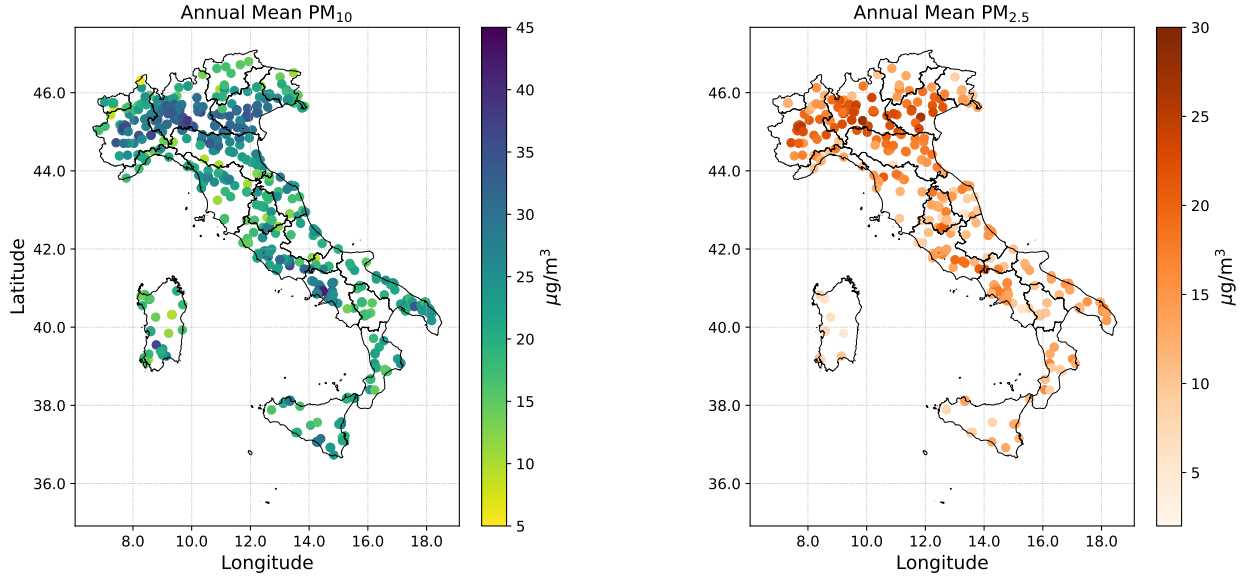


Figure 37: Monitoring network and overall averages concentration over Italy of PM_{10} on the left and $PM_{2.5}$ on the right. A total of $S = 661$ stations are available: $S_1 = 574$ stations measuring PM_{10} , and $S_2 = 313$ measuring $PM_{2.5}$, with 304 stations providing co-located measurements.

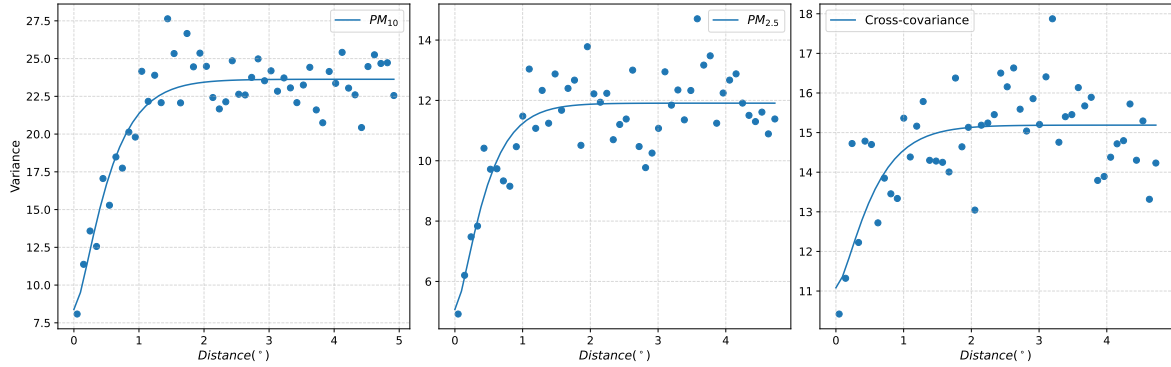


Figure 38: Spatial variograms and covariogram computed from the average concentration at each spatial location in $(\mu\text{g}/\text{m}^3)^2$. The plots show the variogram of PM_{10} on the left, the variogram of $PM_{2.5}$ on the centre, and the cross-covariogram between PM_{10} and $PM_{2.5}$ on the right. The fitted Matérn models are estimated with weighted least squares (Cressie, 1985).

particularly high (see Figure 38, 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 39 (left panel). Not surprisingly, there is a clear seasonality with higher concentrations in winter. Moreover, the empirical ACF is shown in Figure 39. The first lag is close to 0.7 for both pollutants, indicating the presence of a temporal dependence.

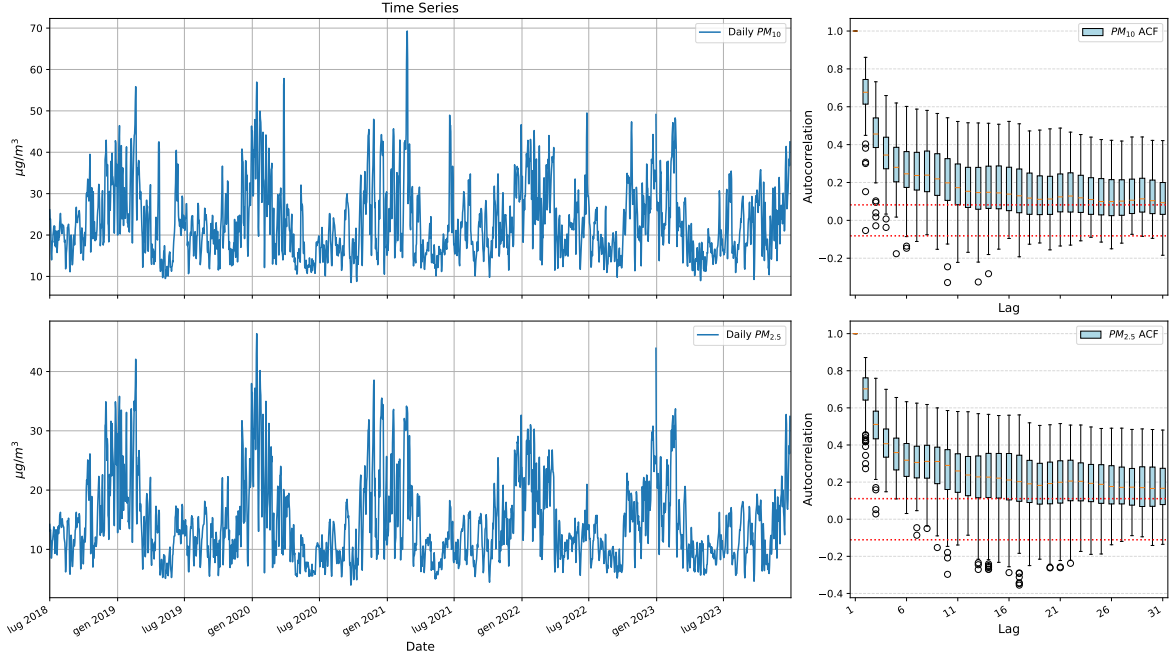


Figure 39: Temporal structure of particulate matter concentrations. The left column shows daily averages of PM_{10} on the top and $PM_{2.5}$ on the bottom; the right column displays corresponding ACFs over the first 30 lags, summarised across stations using boxplots.

Statistics	PM_{10}	$PM_{2.5}$	Altitude	Temperature	RH%	Wind Speed
Unit of measurement	$\mu\text{g}/\text{m}^3$	$\mu\text{g}/\text{m}^3$	m	$^{\circ}\text{C}$		m/s
<i>Percentiles</i>						
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
<i>Descriptive statistics</i>						
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

Table 27: Summary statistics of variables from the AQCLIM-GRINS dataset across all locations.

6.7.3 Application of the LR-SSM

To apply the proposed LR-SSM in Eqs. (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° . Additional details on the triangulation and smoothing procedures are given in Section 6.5.1. The resulting discretised domain is shown in Figure 40. 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).

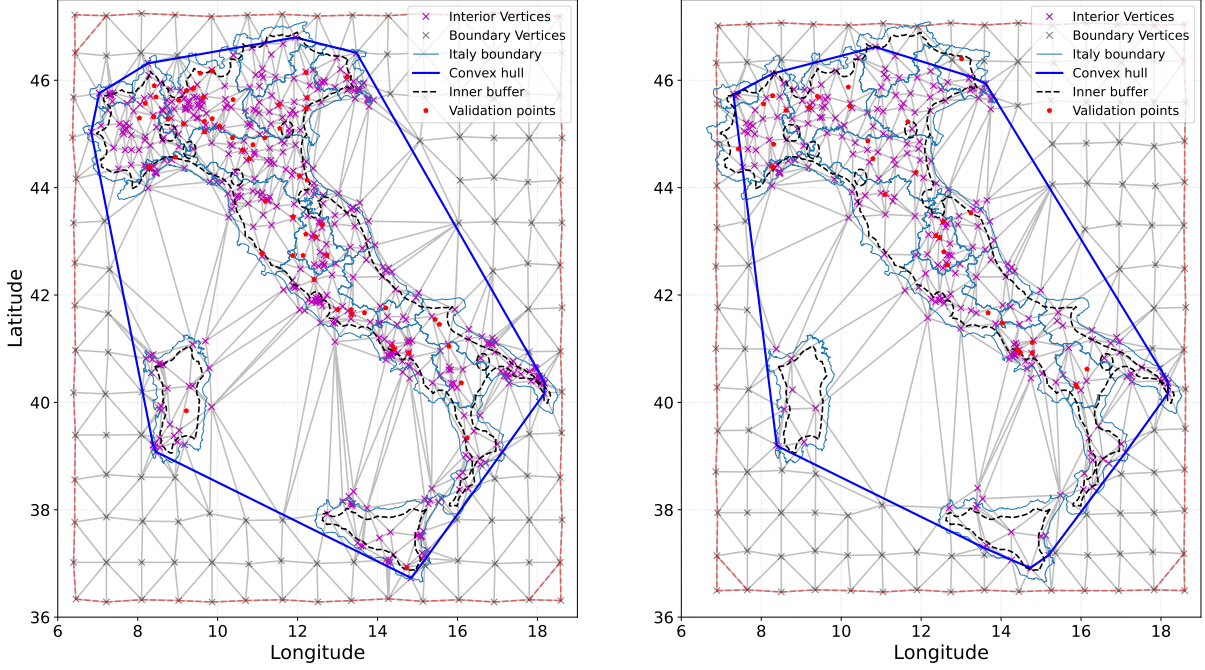


Figure 40: Spatial meshes generated via Laplacian smoothing. Left: mesh constructed for the PM_{10} dataset; right: mesh constructed for the $PM_{2.5}$ dataset. Interior vertices (magenta), boundary vertices (green), convex hulls (blue), inner buffer (black), and validation locations (red) are shown.

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 selecting only stations located at least 0.15° away from the boundary (see the inner buffer in Figure 40). 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 MSE optimality is of interest. Hence, considering $\hat{y}_i(\mathbf{s}, t)$ as in Eq. (6.11), errors $e_i(\mathbf{s}, t) = y_i(\mathbf{s}, t) - \hat{y}_i(\mathbf{s}, t)$ with $\mathbf{s} \in \mathcal{D}$ are computed, and the LR-SSM candidates are compared in terms of the RMSE and R^2 statistic, where, for each pollutant i , $R_i^2 = 1 - \sum_{t=1}^T \sum_{\mathbf{s} \in \mathcal{S}_{i,t}} (y_i(\mathbf{s}, t) - \hat{y}_i(\mathbf{s}, t))^2 / \sum_{t=1}^T \sum_{\mathbf{s} \in \mathcal{S}_{i,t}} (y_i(\mathbf{s}, t) - \bar{y}_i)^2$, where $\bar{y}_i = T^{-1} \sum_{t=1}^T |\mathcal{S}_{i,t}|^{-1} \sum_{\mathbf{s} \in \mathcal{S}_{i,t}} y_i(\mathbf{s}, t)$.

We consider low-rank approximations where LR is set to 100%, 75%, 50%, and 25%. Table 28 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

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		

Table 28: 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.

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 3 h 37 m to just 16 m. 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 Ndonga (2016).

Table 29 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 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 41, 42, 43, and 44 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 $\hat{W}H(s)\hat{z}_t$ and its standard deviation, as defined in the predictive Eqs. (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 41, 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

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

Table 29: 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..

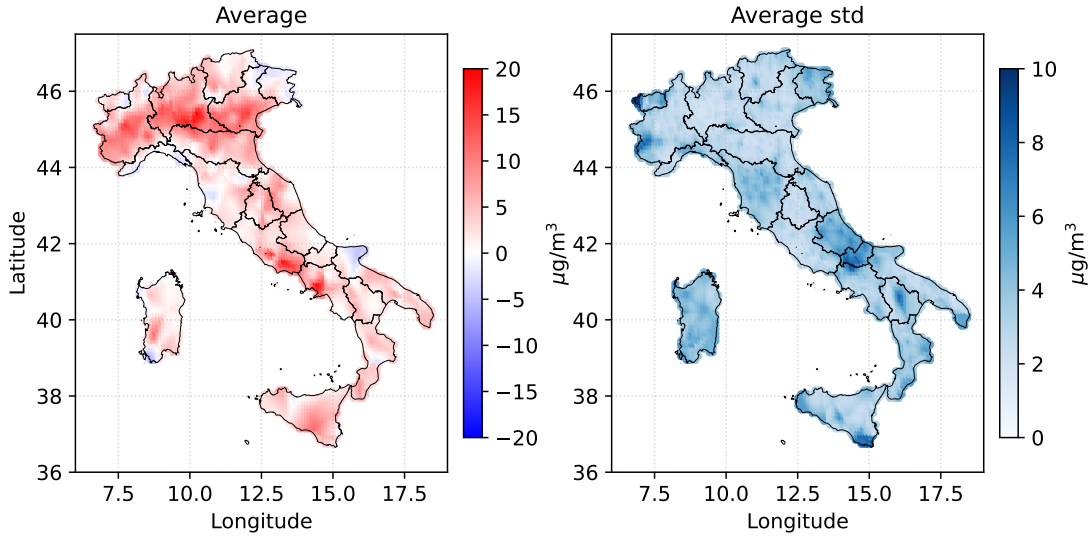


Figure 41: Estimated average latent effect component on PM_{10} on the left, and the associated average standard deviation on the right.

(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 where $LR = 25\%$ in Figure 44, 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 $6 \mu\text{g}/\text{m}^3$. Notably, no boundary effects are observed, and the uncertainty remains stable even near the edges of the domain.

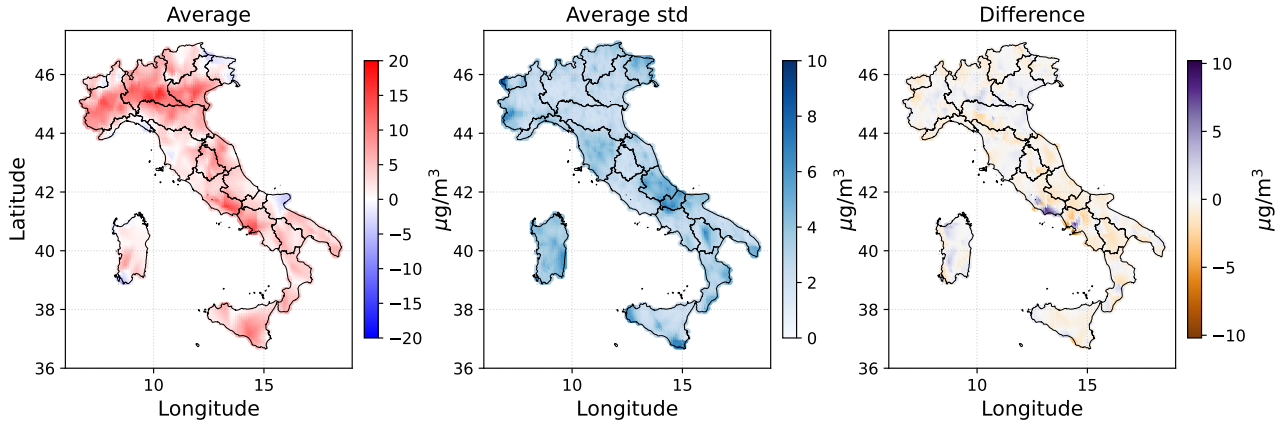


Figure 42: $LR = 75\%$. Estimated average latent effect component on PM_{10} on the left, associated with the average standard deviation on the centre and the average difference between $LR = 100\%$ and $LR = 75\%$ estimations on the right.

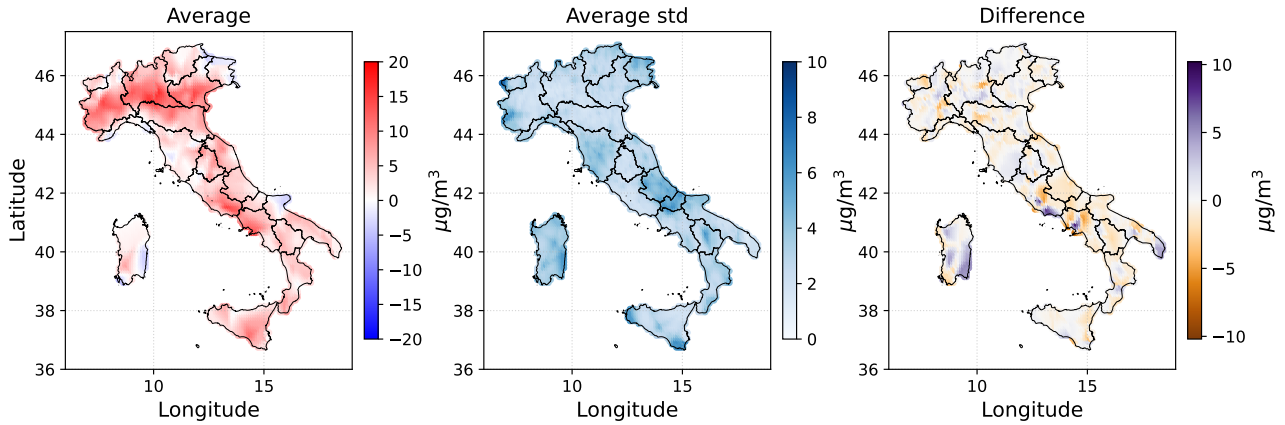


Figure 43: $LR = 50\%$. Estimated average latent effect component on PM_{10} on the left, associated with the average standard deviation on the centre and the average difference between $LR = 100\%$ and $LR = 50\%$ estimations on the right.

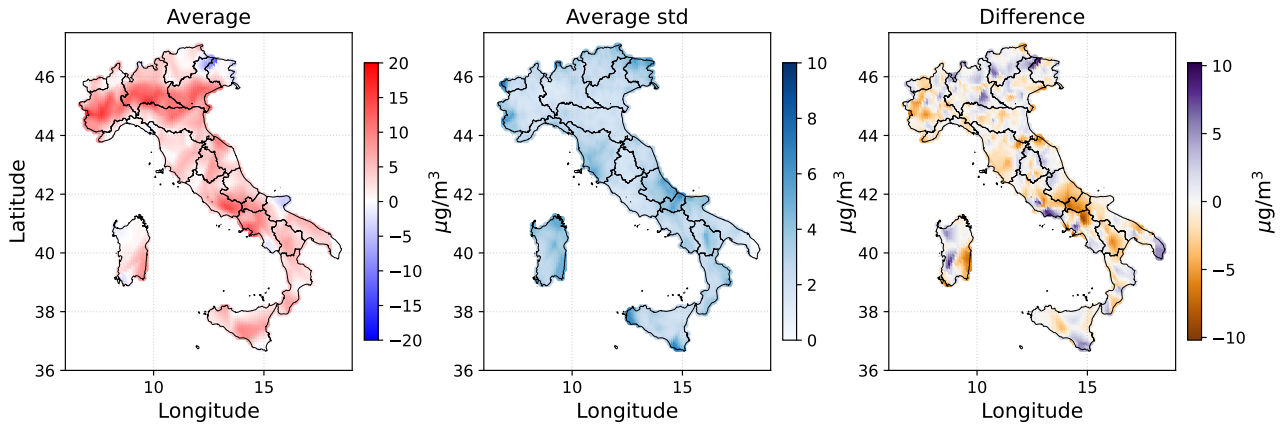


Figure 44: $LR = 25\%$. Estimated average latent effect component on PM_{10} on the left, associated with the average standard deviation on the centre and the average difference between $LR = 100\%$ and $LR = 25\%$ estimations on the right.

6.8 Comparison with the INLA framework

In this section, we compare the proposed approach against competing models within the Integrated Nested Laplace Approximation (INLA) framework (Håvard Rue, Martino, and Chopin, 2009), a widely used tool for Bayesian spatio-temporal modelling. In particular, we are mainly interested in prediction accuracy and estimation time. For the comparison, we consider a bivariate setting in which PM_{10} and $\text{PM}_{2.5}$ are jointly modelled. We adopt the measurement equation in Eq. (2.23), assuming that the bivariate random effect $z(s, t)$ has independent components, while allowing the dynamic specifications to differ across models.

First, we consider a first-order Markovian state-space model, in which z satisfies Eq. (2.24). This specification corresponds to the INLA implementation of our proposed model and serves as a benchmark for comparing the EM-based estimation procedure with the Bayesian approach. We denote this implementation by M_{INLA} . Second, we model $z(s, t)$ using the diffusion-based spatio-temporal SPDE framework proposed by F. Lindgren et al. (2024). Specifically, we consider model 102 in the taxonomy introduced in that work, as alternative specifications were found to exhibit similar predictive performance at the expense of substantially higher computational cost. This model defines $z(s, t)$ as a non-separable spatio-temporal Gaussian random field obtained as the solution to a stochastic partial differential equation combining a Matérn-type spatial operator with a first-order temporal diffusion process, and it is implemented within the INLA framework through a sparse precision matrix representation. We refer to this specification as $M_{\text{INLA},102}$.

Although both M_{INLA} and the $M_{\text{INLA},102}$ induce first-order Markovian dynamics in time, they differ substantially in their construction. The former is formulated as a discrete-time state-space model, whereas the latter defines a continuous spatio-temporal Gaussian random field through a stochastic partial differential equation, leading to different interpretations of the spatial operator and of the space-time interaction. These Bayesian models are compared with our proposed approach using low-rank approximations of $LR = 25\%$, 50% , 75% , and 100% .

All computations are performed on the same hardware described in Section 6.6.3. The model M_{INLA} is estimated using the INLA R package (version 25.12.16), while the model $M_{\text{INLA},102}$ is implemented through the INLAspacetime R package (version 0.1.12). The proposed EM-based estimation procedure, evaluated for $LR = 25\%$, 50% , 75% , and 100% , is implemented in Python. To ensure a fair comparison of computational efficiency, we explicitly control the number of threads available to the INLA internal solver, thereby enforcing consistent resource allocation across all model estimations.

Regarding prior specification for M_{INLA} , we consider the marginal variance of $\mathbf{z}(s, t)$ fixed to 1 to ensure correspondence with our proposal. The range parameter is assigned a penalised complexity (PC) weakly informative prior such that $\Pr(\rho < 0.5) = 0.5$. The autoregressive coefficients are given the default INLA prior, while the regression coefficients β , and the loading matrix $\mathbf{W}$ are assigned independent weakly informative Gaussian priors $N(0, 10^3)$. For the diffusive model $M_{INLA,102}$ we adopt PC priors for the spatial range r_s , the temporal range r_t , and the field standard deviation σ such that $\Pr(r_s < 0.5) = 0.5$, $\Pr(r_t < 5) = 0.5$, and $\Pr(\sigma > 1) = 0.5$. This specification implies a soft lower bound on the spatial correlation range and a soft upper bound on the temporal range, while allowing for data-driven flexibility in the variance. The SPDE approximation across all the INLA-based models is constructed on a mesh with around 250 vertices, depending on the selected subset of the original data, yielding a problem size comparable to that of the full-rank version of our method.

For the comparison, we used a subset of the GRINS dataset (see Section 6.7) with a varying time length $T \in \{15, 30, 60\}$ and retaining the same fixed partition of training and validation locations as defined in the empirical application. This choice allows us to assess scalability with respect to the time dimension. For each T , we run 20 repetitions by randomly selecting the start day in 2023. Averaging the RMSE across the repetitions effectively marginalises over temporal process variability and provides a robust assessment of predictive performance. The comparative analysis presented in Table 30 reveals significant differences in both computational efficiency and predictive behaviour between the proposed framework and the Bayesian benchmarks.

The results show that M_{INLA} requires significantly more time than the full-rank $M_{LR,100\%}$, the cost of $M_{INLA,102}$ is even more pronounced. Crucially, as the temporal dimension increases, the computational demands of the Bayesian inference escalate rapidly, whereas the proposed method maintains a much more favourable scaling profile. This scalability makes the low-rank approach embedded within the SSM particularly advantageous for analysing long time series where standard Bayesian approximations become computationally prohibitive.

Regarding predictive accuracy, the M_{INLA} and $M_{INLA,102}$ generally achieve the lowest out-of-sample RMSE on average. In particular, the $M_{INLA,102}$ frequently yields the most accurate predictions for $PM_{2.5}$, though at the highest computational cost. The proposed method at full rank $M_{LR,100\%}$ delivers competitive performance. Specifically, the model achieves very low training RMSE but higher test RMSE. This discrepancy suggests that the full-rank parameterisation is prone to overfitting and does not generalise to unobserved locations.

A key insight from the results is the efficacy of the low-rank approximations ($LR = 25\%, 50\%, 75\%$) as a regularisation strategy. By constraining the rank of the spatial covariances,

these variants enforce a smoother latent structure. While this naturally results in higher training errors compared to the full-rank model, the degradation in test performance is negligible compared to the gains in computational speed. This behaviour indicates that the low-rank constraint effectively mitigates the overfitting observed in the 100% model. As noted by Otto, Fusta Moro, et al. (2024), mixed-effects models often struggle with generalisation due to their flexibility; here, rank reduction acts as a structural regularizer, offering an attractive trade-off that preserves stable predictive capabilities while reducing estimation time.

Time	Model	RMSE Train		RMSE Test		Computation	
		$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}	Time (s)	Time (%)
$T = 15$	$M_{LR,25\%}$	3.95 (1.25)	5.34 (1.27)	5.22 (1.79)	6.17 (1.56)	20 (4)	11%
	$M_{LR,50\%}$	3.28 (0.92)	4.35 (0.72)	6.26 (2.79)	6.81 (2.22)	54 (10)	29%
	$M_{LR,75\%}$	2.79 (0.68)	3.61 (0.54)	6.29 (2.60)	6.55 (1.52)	109 (23)	59%
	$M_{LR,100\%}$	2.25 (0.72)	2.75 (0.58)	4.90 (1.09)	6.86 (1.44)	185 (29)	100%
	M_{INLA}	3.93 (1.40)	5.62 (1.45)	4.11 (1.61)	5.92 (1.43)	328 (86)	177%
	$M_{INLA,102}$	3.73 (1.22)	5.37 (1.22)	3.98 (1.49)	5.77 (1.30)	512 (133)	277%
$T = 30$	$M_{LR,25\%}$	4.19 (1.28)	5.68 (1.34)	5.55 (1.67)	6.51 (1.66)	22 (4)	10%
	$M_{LR,50\%}$	3.52 (0.95)	4.60 (0.84)	6.01 (2.25)	6.83 (1.67)	64 (12)	30%
	$M_{LR,75\%}$	3.05 (0.74)	3.93 (0.60)	6.33 (2.50)	6.88 (1.54)	126 (33)	59%
	$M_{LR,100\%}$	2.44 (0.71)	3.11 (0.53)	5.13 (1.23)	7.09 (1.42)	215 (57)	100%
	M_{INLA}	4.02 (1.33)	5.81 (1.40)	4.35 (1.58)	6.32 (1.43)	1077 (366)	501%
	$M_{INLA,102}$	3.88 (1.21)	5.61 (1.22)	4.25 (1.48)	6.17 (1.27)	1845 (393)	858%
$T = 60$	$M_{LR,25\%}$	4.26 (1.19)	5.71 (1.33)	5.56 (1.57)	6.49 (1.54)	27 (7)	15%
	$M_{LR,50\%}$	3.69 (0.88)	4.70 (0.84)	5.77 (1.82)	6.55 (1.47)	90 (21)	49%
	$M_{LR,75\%}$	3.17 (0.64)	4.04 (0.61)	6.06 (2.06)	6.57 (1.40)	181 (53)	99%
	$M_{LR,100\%}$	2.34 (0.11)	3.35 (0.29)	4.77 (0.35)	5.89 (0.20)	182 (53)	100%
	M_{INLA}	4.02 (1.21)	5.69 (1.24)	4.27 (1.39)	6.17 (1.18)	3078 (1048)	1691%
	$M_{INLA,102}$	3.89 (1.06)	5.55 (1.09)	4.16 (1.26)	6.03 (1.02)	5692 (1725)	3128%

Table 30: Comparative performance analysis on the GRINS dataset subset. The table reports average Training and Test RMSE (mg/m^3) and computational cost for the proposed Low-Rank (LR) approach (at varying percentages), the benchmark INLA state-space implementation (M_{INLA}), and the diffusion-based SPDE model ($M_{INLA,102}$). Results are stratified by time series length $T \in \{15, 30, 60\}$. Computational time is expressed in seconds and as a percentage relative to $M_{LR,100\%}$. Values represent averages over 20 random repetitions, with standard deviations provided in parentheses.

6.9 Discussion

This chapter develops a scalable framework for high-dimensional multivariate spatio-temporal data, building on state-space models (SSMs) that leverage the SPDE–GMRf approach for computationally efficient inference. The core contribution lies in the combination of these components, which confers several advantages over existing alternatives. First, the SPDE-based basis provides a model-based,

interpretable low-rank representation with controlled approximation error, thereby avoiding the ad hoc basis selection commonly required in other low-rank methods. Second, the resulting sparse structure is embedded within the Kalman filtering framework, preserving linear-time filtering and enabling likelihood-based parameter estimation via an EM algorithm with largely closed-form updates.

Our proposal naturally supports multivariate datasets with both isotopic and heterogeneous observed data, while the EM-based estimation procedure efficiently accommodates extensive missing data. We further demonstrated weak convergence of the approximated latent field to the target Matérn field as the domain discretisation is refined. For prediction, we established a uniform, time-independent bound on the resulting prediction error, which rigorously quantifies the quality of the SPDE approximation to respect the time horizon.

We showed that the computational cost can be controlled through the size of the mesh used to discretise the spatial domain, allowing users to balance the trade-off between accuracy and run time according to available computational resources. Numerical experiments indicate that an increasing approximation level yields prediction errors comparable to those of the full-rank SSM, while substantially lowering computational demands. Notably, a higher rank does not necessarily imply better performance, as SSMs are flexible models that can easily overfit the training data. Consequently, the selection of the rank remains an open issue, and cross-validation may be required to identify an appropriate value. In a real-case application to bivariate air-quality modelling, the proposed method achieves an approximate 93% reduction in computation time with only a slight increase in prediction error.

It is worth noting that the choice of the mesh in the estimation procedure plays a crucial role. Indeed, the satisfaction of topological constraints, such as the boundness of the internal minimum angle, is essential to obtain correct inference. However, this is a common requirement in the SPDE approach, and a vast literature can be found detailing algorithms designed to generate high-quality meshes that satisfy these regularity conditions.

To illustrate the computational capabilities of our proposed low-rank method, we compare the EM-based estimation approach with Bayesian inference via INLA. The results demonstrate that the EM-based approach achieves substantial computational gains relative to INLA, particularly as the time horizon increases, while maintaining comparable predictive accuracy. This comparison should be interpreted as illustrative rather than definitive; its primary purpose is to show that the proposed MLE-based approach yields stable and statistically well-behaved estimates while remaining computationally competitive.

The proposed multivariate modelling assumption is based on the linear model of coregionalisation, in which the latent components are assumed to be mutually independent and are combined through a linear weighting matrix. This represents a standard modelling strategy for multivariate datasets and it has been shown to provide good predictive performance in practice. In addition, we restrict the latent independent fields to have a smoothness parameter $\nu = 1$. Future work may relax this assumption. Furthermore, the proposed SSM (6.9)-(6.10), relies on a linear relationship between the outcome and the latent variables. More complex systems and real case applications, however, may exhibit nonlinear behaviours. Nonlinearity may arise in the measurement equation, the state equation, or both. The optimality of the Kalman Filter (KF) is limited to linear systems. Future work could extend the proposed low-rank approach to nonlinear SSM by considering extensions of the KF, such as the Extended KF and the Unscented KF (Wan and Van Der Merwe, 2000).

The architectural choice of using the SPDE–GMRF approach offers versatility and potential for extension. Because the methodology relies on a mesh and a sparse precision structure, the entire framework can be straightforwardly transferred from planar regions to curved two-dimensional manifolds, such as the sphere, or adapted to discrete networks, such as road systems, power grids, or social graphs (Krainski et al., 2018), by simply modifying the underlying geometry or basis functions. In all these extensions, the preservation of the sparse precision matrix ensures that the core linear-time Kalman–EM machinery remains intact.

6.10 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 programmes, where traditional methods quickly become computationally infeasible.

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 addressed 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 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 Markov chain Monte Carlo 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 (e.g. Yaqiong Wang, Finazzi, and Fassò (2021)). 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 perfor-

mance 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 in our GitHub repository: https://github.com/jacopoRodeschini/Low_Rank_State_Space_Model.

Chapter 7

Conclusions and future work

7.1 Main contributions of the thesis

This thesis develops methodological contributions within the SSM framework suitable for environmental datasets. We proposed 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 developed a multivariate SSM suitable for handling heteroskedastic spatio-temporal processes by allowing for a time-varying error variance. We applied 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 developed 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, enables 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 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 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 the 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 is the development of a multivariate heteroskedastic spatio-temporal model. In particular, we proposed modelling heteroskedasticity in two main frameworks: (i): time-varying unstructured flexible error variance and (ii) stochastic heteroskedastic functions characterised by a lower number of degrees of freedom. Building on the model comparison, we implemented 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 EM algorithm. Using the Agrimonia dataset, we applied the proposed heteroskedastic HDGM to assess the sensitivity of $PM_{2.5}$ to ammonia emissions through a ‘what-if’ scenario analysis. Our findings suggest that considering a 26% reduction in ammonia emissions, the overall $PM_{2.5}$ reduction is approximately 3.5%. In contrast, considering a 50% reduction in ammonia emissions, the overall $PM_{2.5}$ reduction is approximately 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 a moderate m .

4. The fourth main contribution of this thesis is the development of a multivariate low-rank SSM (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 KF, this reduces the computational cost of the KF 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 proved 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. The air quality application showed the method’s ability to handle large, multivariate heterotopic datasets with substantial missingness. In this application, we reduced computation time by approximately 93%, with only a 15% increase in 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 allocating 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.

Extension of the SSM framework on manifold and network-based domains

Leveraging the flexibility of the SSM framework, we outline here potential extensions that enable inference for spatial and spatio-temporal data observed on a manifold and, in particular, on network-based domains. This direction is motivated by the growing relevance of stochastic processes defined on networks in spatial and spatio-temporal statistics, as reflected in an expanding range of applications (Baddeley et al., 2017; Anderes, Møller, and J. G. Rasmussen, 2020).

In Chapter 3, we introduced the GMRF representation of a Matérn GP as the solution of an SPDE driven by Gaussian white noise. The resulting approach does not involve having to construct explicit formulae for the covariance functions, which are instead only defined implicitly through the SPDE specifications. In particular, the SPDE approach extends to Matérn fields on manifolds, network-based domains, non-stationary fields, and fields with oscillating covariance functions. Considering the manifold extension, the main idea is to formulate the SPDE directly on the manifold. The Gaussian white noise driving the SPDE can be defined on the manifold as a zero-mean GP $\mathcal{W}(\cdot)$ such that, for any measurable subsets A and B of the manifold, the covariance between $\mathcal{W}(A)$ and $\mathcal{W}(B)$ is proportional to the surface measure of $A \cap B$. The theory to solve the SPDE over a general manifold is developed in Appendices B–D of (Finn Lindgren, Håvard Rue, and Lindström, 2011). Building on the GMRF construction of Matérn fields on a manifold, the SSM framework can be extended to enable spatio-temporal inference on manifold domains. Following the approach adopted in Chapter 6, the state equation can be defined directly on the manifold by exploiting the GMRF representation of the innovation term.

The construction of GPs on linear networks and metric graphs is challenging due to the difficulty of specifying flexible, positive definite covariance functions compatible with network topology. Within the SSM framework, this challenge translates into the proper definition of the covariance matrix for the innovation term $\eta(\mathbf{s}, t)$. Recently, Porcu, White, and Genton (2023) introduced new classes of stationary, non-separable space-time covariance functions suitable for generalised networks, including Euclidean trees and linear networks. These constructions were extended by Filosi, Agostinelli, and Porcu (2023) to settings with temporally evolving topological structures. An alternative approach characterises GPs through SPDEs defined directly on the network (Bolin, Simas, and Wallin, 2024). Moreover, a subclass of the processes introduced by Bolin, Simas, and Wallin (2024) exhibits Markov properties, enabling exact evaluation of finite-dimensional distributions. This, in turn, allows for exact

likelihood-based inference and spatial prediction with high computational efficiency. Both approaches can be considered for defining the innovation process $\eta(s, t)$.

7.2 Summary

This thesis 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 a 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 the 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 batch inference algorithms (using all available data) within the Kalman–EM machinery.

Appendices

Appendix A

Appendix for Chapter 4

A.1 Data availability, abundance and standards

The Agrimonia dataset originates from a multi-pollutant monitoring network, where not every measurement station was equipped with all sensors. In addition to this heterogeneity, maintenance of the AQ monitoring network and temporal and/or geographic constraints of the CAMS, BDN and SIARL datasets lead to varying data availability. For convenience, the dataset is published in tabular format with size $309072 \text{ rows} \times 41 \text{ columns}$, and all variables are included for each station and time point. If a sensor/measurement was not available, the values are replaced with (pseudo-)‘NaN’ values. Even though this tabular format increases the required storage capacity, modern data analysis software can efficiently handle the large number of ‘NaN’ values associated with it.

In particular, for AQ, the number of stations per pollutant is reported in Table 3 with an average of 73, about one-half of the 141 stations. The number of pseudo-random missing data per year of active sensors is reported in Table 7. As a result, about 54% of the AQ cells are set to ‘NaN’ in the Agrimonia dataset. For EM, data are available only up to 2020. For LI, the three stations in Switzerland have missing data for bovines and swine because the BDN census data is available only in Italy. Similarly, for LA soil use, SIARL data is available only in Lombardy up to 2019. Hence, the geographic buffer of 48 stations and the last two years are set at ‘NaN’ in our dataset. Considering the above-mentioned aims of the AgrImOnIA project, the missing values outside Lombardy for LA and LI predictors are not a concern due to the focus on the region and the auxiliary role of the buffer. Instead, the lack of SIARL data after 2019 limits the possibility of using the cultivation classification for the entire time series.

The Agrimonia dataset is consistent with the following reference systems: World Geodetic System 1984 (WGS84) Defense Mapping Agency Washington DC, 1991 for georeferentiation, Coordinated

Universal Time (UTC) for time referentiation and the International System of Units (SI) for the metric system, except for the temperature expressed in Celsius degrees ($^{\circ}\text{C}$). The values in the dataset are represented in scientific notation with four significant digits here, considered sufficient for statistical modelling purposes. The coordinates (latitude, longitude) have a fixed point representation with nine significant digits to identify the stations' locations correctly.

A.2 Validation of AQ variables

The last column of Table 3 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 $\text{PM}_{2.5}$, 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 $t \in \{1, \dots, 24T\}$, 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:

$$\begin{cases} x_t = Ax_{t-1} + Bu_t \\ y_t = x_t + \epsilon_t \end{cases} \quad (\text{A.1})$$

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 $\sigma_d = \sqrt{1/24^2 \sum_{t \in d} \text{Var}(x_t | y_1, \dots, y_{24T})}$ where $\text{Var}(x_t | y_1, \dots, y_{24T})$ is the variance of the smoothed states x_t for the hour $t \in \{1, \dots, 24T\}$.

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 45. 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 45 shows the two daily time series: the blue line depicts our method, while the orange diamonds are the ARPA Lombardy daily data. From Figure 45, 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 46 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 $\text{PM}_{2.5}$ 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 Misericordia’. In fact, the data on $\text{PM}_{2.5}$ concentrations are available from May 2021, are validated and have a bi-hourly frequency. Figure 47 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 48 draws the imputed uncertainty of daily average concentrations computed from imputed hourly time series and shows that it is small.

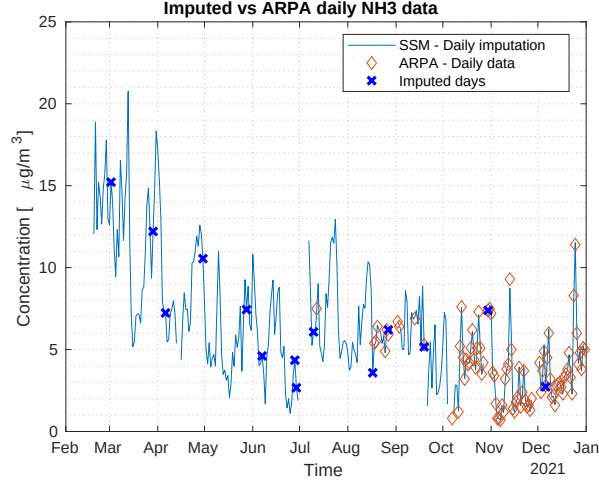


Figure 45: Hourly NH_3 data. Impact of Kalman smoother on daily data for the monitoring station named ‘Bergamo Via Meucci’. Daily time series obtained with our method (blue line) with highlighted imputed days (blue crosses) and daily raw data (orange diamonds), RMSE = 0.1710.

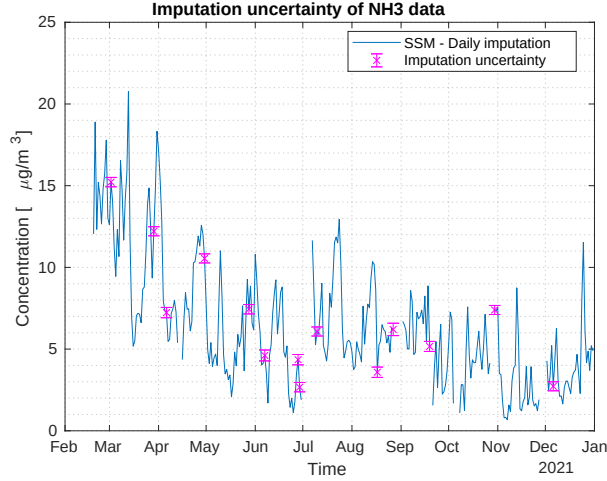


Figure 46: Hourly NH_3 data. Impact of Kalman smoother on daily data for the monitoring station named ‘Bergamo Via Meucci’. Imputed daily data (blue line) with associated imputed uncertainty ($\pm 2\sigma_d$, magenta bar).

We examined the dataset for the presence of anomalous values that are clearly outliers after the daily time series construction. Table 31 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.

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 31: Extremely large values for the AQ variables [$\mu\text{g}/\text{m}^3$] identified in the dataset. The values are replaced with the ‘NaN’ value.

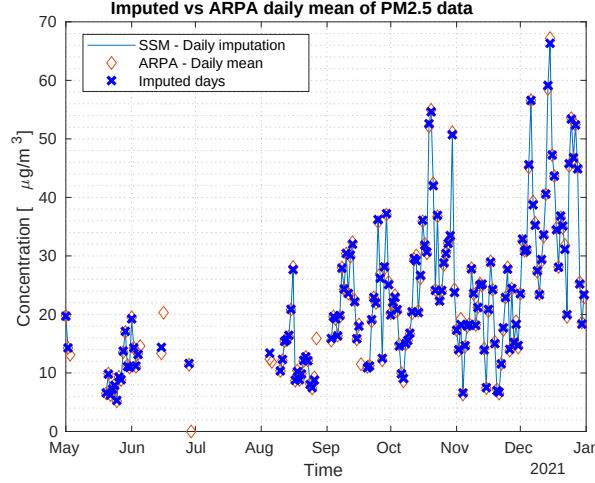


Figure 47: Bi-hourly $PM_{2.5}$ data. Impact of Kalman smoother on daily data for the monitoring station named ‘Parona Via della Misericordia’. Daily time series obtained with our method (blue line) with highlighted imputed days (blue crosses) and daily mean (orange diamonds) computed without considering missing values, RMSE = 0.3078.

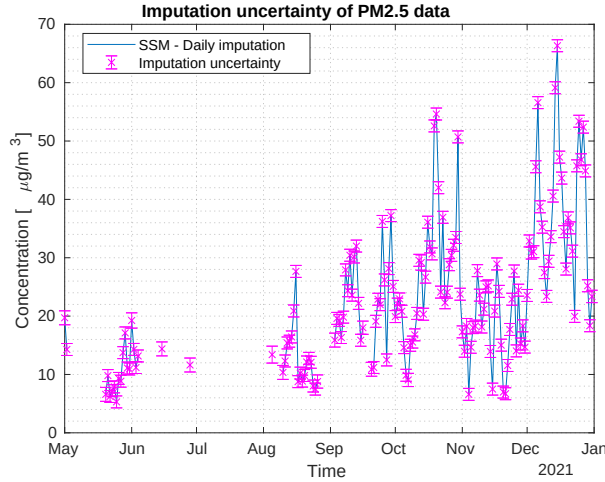


Figure 48: Bi-hourly $PM_{2.5}$ data. Impact of Kalman smoother on daily data for the monitoring station named ‘Parona Via della Misericordia’. Imputed daily data (blue line) with associated imputed uncertainty ($\pm 2\sigma_d$, magenta bar).

A.2.1 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 $0.25^\circ \times 0.25^\circ$. 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.3 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 continuous 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 49 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.

A.3.1 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

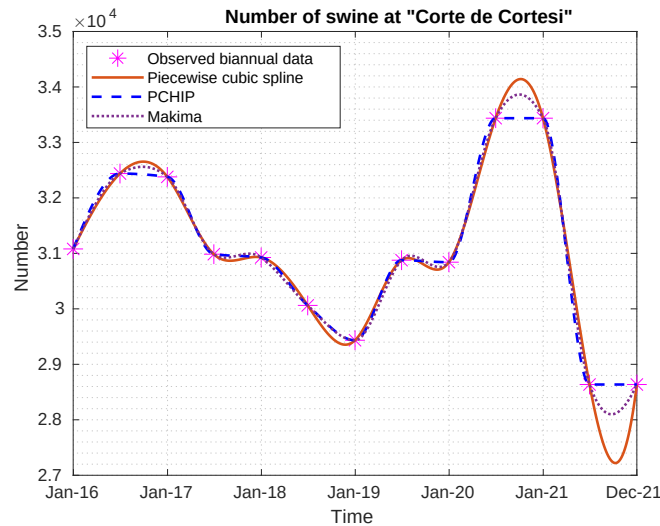


Figure 49: Piecewise cubic spline, PCHIP and Makima interpolation methods applied to swine time series for the monitoring station named ‘Corte de Cortesi’.

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 50 shows the soil use change in 2019 provided by the SIARL dataset for the station named ‘Corte dei Cortesi’.

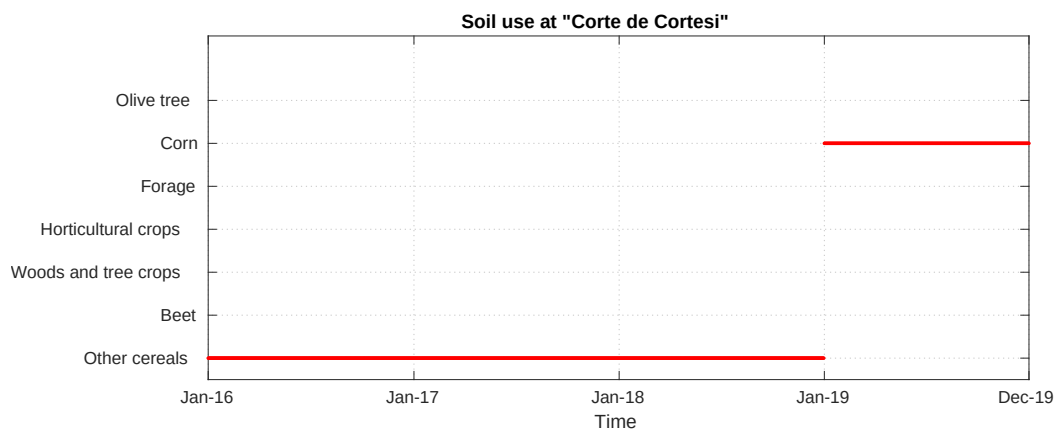


Figure 50: Piecewise constant function for soil use provided by SIARL dataset for the station named ‘Corte de Cortesi’. Note that the SIARL dataset covers data up to 2019 only.

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 I_n the identity matrix of dimension n .

B.2 Proof of Theorem 1

Proof. Let

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

be two parameter sets that yield the same joint distribution for the observable field $\{y(s, t) : (s, t) \in S \times \mathbb{N}\}$ under the LR-SSM (6.9)–(6.10) for a fixed R .

Consider

$$y(s, t) = X'(s, t)\beta + W\tilde{\Psi}_R(s)z_t + \varepsilon(s, t), \quad z_t = (\text{diag } f \otimes I_R)z_{t-1} + \eta_t,$$

where $\{z_t\}$ is strictly stationary and ergodic with $\mathbb{E}z_t = 0$ and $\{\varepsilon(s, t)\}_{t \geq 1}$ is i.i.d. with $\mathbb{E}\varepsilon(s, t) = 0$ and $\sup_{s \in S} \mathbb{E}\|\varepsilon(s, t)\|^2 < \infty$. Let

$$L_X := \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \int_S X(s, t) ds,$$

which exists and is finite by Assumption 1. Suppose $\lambda(S) < \infty$ and $\sup_s \|\tilde{\Psi}_R(s)\| < \infty$. By the Birkhoff ergodic theorem $T^{-1} \sum_{t=1}^T z_t \rightarrow 0$ almost surely; hence, by dominated convergence and Fubini,

$$\frac{1}{T} \sum_{t=1}^T \int_S W\tilde{\Psi}_R(s)z_t ds \xrightarrow{\text{a.s.}} 0.$$

Similarly, by the Strong Law of Large Numbers $T^{-1} \sum_{t=1}^T \varepsilon(s, t) \rightarrow 0$ almost surely for each s , and again by dominated convergence,

$$\frac{1}{T} \sum_{t=1}^T \int_S \varepsilon(s, t) ds \xrightarrow{\text{a.s.}} 0.$$

Therefore,

$$\frac{1}{T} \sum_{t=1}^T \int_S y(s, t) ds = \left(\frac{1}{T} \sum_{t=1}^T \int_S X(s, t) ds \right)' \beta + o_{\text{a.s.}}(1) \xrightarrow{\text{a.s.}} L'_X \beta.$$

Since $L_X L'_X$ has full rank by Assumption 1,

$$\beta = (L_X L'_X)^{-1} L'_X \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \int_S y(s, t) ds,$$

and thus β is uniquely identified.

Next, define the demeaned observations as

$$r(s, t) = y(s, t) - X'(s, t)\beta = W\tilde{\Psi}_R(s)z_t + \varepsilon(s, t),$$

where $\mathbf{z}_t = (\text{diag}(\mathbf{f}) \otimes \mathbf{I}_R) \mathbf{z}_{t-1} + \boldsymbol{\eta}_t$ and $\tilde{\boldsymbol{\Psi}}_R(\mathbf{s}) = \mathbf{I}_q \otimes \boldsymbol{\psi}_R(\mathbf{s})'$ ensures dimension compatibility. The lag-one covariance function is

$$\boldsymbol{\Gamma}_1(\mathbf{s}, \mathbf{s}') = \text{Cov}(\mathbf{r}(\mathbf{s}, t), \mathbf{r}(\mathbf{s}', t-1)) = \mathbf{W} \tilde{\boldsymbol{\Psi}}_R(\mathbf{s}) \mathbf{A} \mathbf{V}_z \tilde{\boldsymbol{\Psi}}_R(\mathbf{s}')' \mathbf{W}',$$

where $\mathbf{A} = \text{diag}(\mathbf{f}) \otimes \mathbf{I}_R$ and $\mathbf{V}_z = \oplus_{i=1}^q (1 - f_i^2)^{-1} \mathbf{Q}_{\kappa_i}^{-1}$ is the stationary covariance of $\mathbf{z}_t$. Let $\mathbf{W}^+$ be a left inverse of $\mathbf{W}$ such that $\mathbf{W}^+ \mathbf{W} = \mathbf{I}_q$, and define

$$\mathbf{S}_\ell = \int_S \int_S \tilde{\boldsymbol{\Psi}}_R(\mathbf{s})' \mathbf{W}^+ \boldsymbol{\Gamma}_\ell(\mathbf{s}, \mathbf{s}') \mathbf{W}^+ \tilde{\boldsymbol{\Psi}}_R(\mathbf{s}') d\mathbf{s} d\mathbf{s}', \quad \ell = 0, 1,$$

where $\boldsymbol{\Gamma}_0(\mathbf{s}, \mathbf{s}') = \text{Cov}(\mathbf{r}(\mathbf{s}, t), \mathbf{r}(\mathbf{s}', t))$. Because ε is temporally white, it contributes to $\boldsymbol{\Gamma}_0$ but not to $\boldsymbol{\Gamma}_1$. Hence, in the population limit,

$$\mathbf{S}_1 = \mathbf{A} \mathbf{G}, \quad \mathbf{S}_0 = \mathbf{G}, \quad \mathbf{G} = \int_S \int_S \tilde{\boldsymbol{\Psi}}_R(\mathbf{s})' \mathbf{V}_z \tilde{\boldsymbol{\Psi}}_R(\mathbf{s}') d\mathbf{s} d\mathbf{s}'.$$

Since $\mathbf{G}$ is positive definite by Assumption 2(iii),

$$\mathbf{H} = \mathbf{S}_0^{-1} \mathbf{S}_1 = \mathbf{A} = \text{diag}(\mathbf{f}) \otimes \mathbf{I}_R.$$

The eigenvalues of $\mathbf{H}$ are $\{f_1, \dots, f_q\}$, each with multiplicity R . Equality in distribution between Π and Π^* implies $\mathbf{H} = \mathbf{H}^*$, and Assumption 2(i), which imposes the strict ordering $f_1 > \dots > f_q$, ensures that $\mathbf{f} = \mathbf{f}^*$.

Stacking a dense set of spatial locations into $\mathbf{r}_t$ gives

$$\text{Cov}(\mathbf{r}_t) = (\mathbf{I}_m \otimes \mathbf{W})(\mathbf{P} \mathbf{V}_z \mathbf{P}')(\mathbf{I}_m \otimes \mathbf{W})' + \mathbf{I}_m \otimes \boldsymbol{\Sigma},$$

where $\mathbf{P}$ collects the evaluations of $\tilde{\boldsymbol{\Psi}}_R(\mathbf{s})$ over $\mathcal{S}$. Since $\mathbf{P}$ and $\mathbf{W}$ have full column rank by Assumption 2(ii)–(iii), and $\mathbf{V}_z$ is block-diagonal with distinct blocks (functions of κ_i), standard factor-analysis identification arguments (Bai and P. Wang, 2015) imply that $\mathbf{W} = \mathbf{W}^*$ and $\mathbf{V}_z = \mathbf{V}_z^*$, up to sign changes already fixed by the column-scaling convention. Because the SPDE mapping $k_i \mapsto \mathbf{Q}_{\kappa_i}$ is injective (see Lemma B.2), it follows that $k_i = k_i^*$ for all i .

Finally, from the diagonal elements of

$$\text{Var}(\mathbf{r}(\mathbf{s}, t)) = \mathbf{W} \tilde{\boldsymbol{\Psi}}_R(\mathbf{s}) \mathbf{V}_z \tilde{\boldsymbol{\Psi}}_R(\mathbf{s})' \mathbf{W}' + \boldsymbol{\Sigma},$$

each σ_i^2 is uniquely determined, implying $\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^{2*}$. Therefore, all parameters coincide, $\Pi = \Pi^*$, and the parameter set of the LR-SSM is point-identified in distribution. $\square$

Let $\mathbf{C}$ and $\mathbf{G}$ be the mass and stiffness matrices associated with a fixed finite element mesh, where $\mathbf{C}$ is symmetric positive definite and $\mathbf{G}$ is symmetric positive semidefinite. For $k \geq 0$ define

$$\mathbf{Q}_k = (\mathbf{k}^2 \mathbf{C} + \mathbf{G}) \mathbf{C}^{-1} (\mathbf{k}^2 \mathbf{C} + \mathbf{G}).$$

Then the map $k \mapsto \mathbf{Q}_k$ is injective: if $\mathbf{Q}_{k_1} = \mathbf{Q}_{k_2}$ for $k_1, k_2 \geq 0$, then $k_1 = k_2$. The same statement holds if $\mathbf{C}$ is replaced by a diagonal mass-matrix approximation $\tilde{\mathbf{C}}$ that is positive definite.

Proof. Write $t = k^2 \geq 0$ and define $\mathbf{Q}(t) = \mathbf{Q}_k$. Expanding,

$$\mathbf{Q}(t) = \mathbf{C} t^2 + 2\mathbf{G} t + \mathbf{G} \mathbf{C}^{-1} \mathbf{G}.$$

Assume $\mathbf{Q}(t_1) = \mathbf{Q}(t_2)$ for some $t_1, t_2 \geq 0$. Subtracting gives

$$\mathbf{C} (t_1^2 - t_2^2) + 2\mathbf{G} (t_1 - t_2) = \mathbf{0},$$

hence

$$(t_1 - t_2) ((t_1 + t_2)\mathbf{C} + 2\mathbf{G}) = \mathbf{0}.$$

If $t_1 \neq t_2$, then $(t_1 + t_2)\mathbf{C} + 2\mathbf{G} = \mathbf{0}$. But $(t_1 + t_2) \geq 0$, $\mathbf{C} \succ \mathbf{0}$, and $\mathbf{G} \succeq \mathbf{0}$, so $(t_1 + t_2)\mathbf{C} + 2\mathbf{G} \succeq (t_1 + t_2)\mathbf{C}$. If $t_1 + t_2 > 0$, this matrix is positive definite and cannot be the zero matrix, a contradiction. The remaining possibility $t_1 + t_2 = 0$ forces $t_1 = t_2 = 0$, contradicting $t_1 \neq t_2$. Therefore $t_1 = t_2$, and hence $k_1^2 = k_2^2$ with $k_1, k_2 \geq 0$ implies $k_1 = k_2$.

If $\mathbf{C}$ is replaced by a diagonal approximation $\tilde{\mathbf{C}} \succ \mathbf{0}$, the same algebra shows that

$$(t_1 - t_2) ((t_1 + t_2)\tilde{\mathbf{C}} + 2\mathbf{G}) = \mathbf{0},$$

and the identical positive-(semi)definiteness argument yields $t_1 = t_2$. □

B.3 Weak convergence of the low-rank approximation

In this section, we discuss the weak convergence of the process $\Psi_R(s)z(\mathcal{V}_R, t)$ to the continuous process $z(s, t)$. 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 Finn Lindgren, Håvard Rue, and Lindström (2011). The goal is to approximate a GF with covariance matrix Σ using a GMRF with precision matrix $\mathbf{Q}$ such that $\mathbf{Q}^{-1}$ 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 $x(\mathbf{s})$ with Matérn covariance characterised by parameters (σ^2, ν, κ) is a stationary solution of:

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

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 $x(\mathbf{s})$ 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:

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

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

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

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:

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

The weak solution satisfies:

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

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

$$x^R(\mathbf{s}) = \sum_{k=1}^R \psi_k(\mathbf{s})w_k = \psi_R' \mathbf{w}, \quad (\text{B.4})$$

where $\psi_R(\mathbf{s}) = (\psi_1(\mathbf{s}), \dots, \psi_R(\mathbf{s}))'$ are piecewise linear basis functions, and $\mathbf{w}$ 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:

$$\begin{aligned} C_{ij} &= \langle \psi_i, \psi_j \rangle, \\ G_{ij} &= \langle \nabla \psi_i, \nabla \psi_j \rangle, \\ K_{\kappa^2} &= \kappa^2 C + G. \end{aligned}$$

Under Neumann conditions, the precision matrix for w becomes:

$$Q_{\kappa^2} = K_{\kappa^2} C^{-1} K_{\kappa^2}. \quad (\text{B.5})$$

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

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

$$\sup_{f \in \mathcal{H}^1, \|f\|_{\mathcal{H}^1} \leq 1} \mathbb{E} [\langle f, x^R - x \rangle_{\mathcal{H}^1}^2] \leq ch^2, \quad (\text{B.6})$$

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 Finn 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 $f, g \in L^2(\mathcal{D})$,

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

Theorem 4. *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 $\mathcal{H}_R^1(\mathcal{D}, \kappa)$ using Gaussian white noise $\mathcal{W}$. Then,*

$$\begin{aligned} x^R &\xrightarrow{D\{L^2(\mathcal{D})\}} x, \\ \mathcal{L}x^R &\xrightarrow{D\{L^2(\mathcal{D})\}} \mathcal{L}x, \end{aligned}$$

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

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

B.3.1 Proof of Corollary 1

Proof. By Theorem 4, 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 $t = 0, \dots, T$. Due to independence, and the fact that $t = 1, \dots, T$ 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 $z_i^R(s, t) = \boldsymbol{\psi}_R(s) z_i(\mathcal{V}_R, t)$ where the Gaussian weights are $z_i(\mathcal{V}_R, t) = \sum_{j=0}^t f_i^{t-j} \boldsymbol{\eta}_i(\mathcal{V}_R, j)$. 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 $i = 1, \dots, q$, we obtain the aimed result:

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

This convergence result also implies that

$$\begin{aligned} \mathbb{E}[\boldsymbol{\Psi}_R(s) \mathbf{z}(\mathcal{V}_R, t)] &\rightarrow \mathbb{E}[\mathbf{z}(s, t)], \\ \text{Cov}(\boldsymbol{\Psi}_R(s) \mathbf{z}(\mathcal{V}_R, t), \boldsymbol{\Psi}_R(\tilde{s}) \mathbf{z}(\mathcal{V}_R, \tilde{t})) &\rightarrow \text{Cov}(\mathbf{z}(s, t), \mathbf{z}(\tilde{s}, \tilde{t})). \end{aligned}$$

□

□

B.3.2 Proof of Theorem 2

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

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

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

$$\mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_t^R\|_2^2] \leq C \cdot \mathbb{E}[\|\mathbf{z}_t - \mathbf{z}_t^R\|_2^2], \quad (\text{B.7})$$

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

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

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

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

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

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

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

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

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

$$\mathbb{E}[\|z_t^R - z_t\|_{L^2}^2] \leq \sum_{r=0}^t \varphi^{2r} q C_{\text{emb}}^2 c h^2 = \frac{1 - \varphi^{2(t+1)}}{1 - \varphi^2} q C_{\text{emb}}^2 c h^2 \quad (\text{B.8})$$

Equations (B.7) and (B.8) yield the result. $\square$

B.4 Proof of Theorem 3

Let $\mathbf{y}(s, t)$ 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 $(\boldsymbol{\beta}', \boldsymbol{\sigma}', \mathbf{f}', k_1, \dots, k_q)'$. Denote by

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

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

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

Define the covariance perturbation

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

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

$$\|\Delta_R(\boldsymbol{\pi})\| \leq C h, \quad (\text{B.9})$$

where 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

$$\boldsymbol{\pi}_R^* := \arg \min_{\boldsymbol{\vartheta} \in \Theta_R} \text{KL}(\mathcal{P}_{\boldsymbol{\pi}} \parallel \mathcal{P}_{\boldsymbol{\vartheta}}^{(R)}) \quad (\text{B.10})$$

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

$$\text{KL}(\pi, \vartheta) = \frac{1}{2} \left\{ \text{tr}[\boldsymbol{\Sigma}_y^{(R)}(\vartheta)^{-1} \boldsymbol{\Sigma}_y(\pi)] - \log \det[\boldsymbol{\Sigma}_y^{(R)}(\vartheta)^{-1} \boldsymbol{\Sigma}_y(\pi)] - p \right\}. \quad (\text{B.11})$$

Setting the gradient to zero yields, after vectorising (vec) the matrices,

$$0 = \mathbf{J}' \boldsymbol{\Sigma}_y^{-1}(\pi) \otimes \boldsymbol{\Sigma}_y^{-1}(\pi) \text{vec } \boldsymbol{\Delta}_R(\pi) + \mathbf{H}(\pi_R^* - \pi) + o(\|\boldsymbol{\Delta}_R(\pi)\|),$$

where $\mathbf{J} = \nabla_{\vartheta} \text{vec } \boldsymbol{\Sigma}_y^{(R)}(\pi) \big|_{\vartheta=\pi}$ and $\mathbf{H}$ is the Fisher-information matrix of the correct model, which is positive definite. Solving the equation, we obtain the exact first-order bias

$$\pi_R^* - \pi = -\mathbf{H}^{-1} \mathbf{J}' (\boldsymbol{\Sigma}_y^{-1}(\pi) \otimes \boldsymbol{\Sigma}_y^{-1}(\pi)) \text{vec } \boldsymbol{\Delta}_R(\pi) + o(\|\boldsymbol{\Delta}_R(\pi)\|) \quad (\text{B.12})$$

Combining (B.12) with the covariance bound (B.9) delivers the rate

$$\|\pi_R^* - \pi\| = O(h).$$

Note that the term (B.11) does not involve the mean of the processes, which implies the parameter $\boldsymbol{\beta}$ has no bias. $\square$

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 $\mathbf{y}_t$ the $(m_t = \sum_{j=1}^p m_{j,t})$ -dimensional vector

$$\mathbf{y}_t = (\mathbf{y}_1(\mathcal{S}_{1,t}, t)', \mathbf{y}_2(\mathcal{S}_{2,t}, t)', \dots, \mathbf{y}_p(\mathcal{S}_{p,t}, t)')',$$

where $\mathcal{S}_{j,t} = \{\mathbf{s}_{i,t}\}_{i=1}^{m_{j,t}}$ is the set of locations where the process $y_j(\cdot, t)$ has been observed, and

$$\mathbf{y}_{j,t} = \mathbf{y}_j(\mathcal{S}_{j,t}, t) = \left(y_j(\mathbf{s}_{1,t}, t), \dots, y_j(\mathbf{s}_{m_{j,t},t}, t) \right)'.$$

Let $\mathbf{y}_{1:T}$ denote the vector $(\mathbf{y}'_1, \dots, \mathbf{y}'_T)'$.

An analogous notation is used for $\mathbf{z}_{1:T} = (\mathbf{z}'_1, \dots, \mathbf{z}'_T)'$, where

$$\mathbf{z}_t = (\mathbf{z}_1(\mathcal{V}_R, t)', \dots, \mathbf{z}_q(\mathcal{V}_R, t)')', \quad \text{with} \quad \mathbf{z}_{j,t} = \mathbf{z}_j(\mathcal{V}_R, t) = (z_j(\mathbf{r}_1, t), \dots, z_j(\mathbf{r}_R, t))'.$$

Moreover, let $\mathbf{X}_{1:T} = [\mathbf{X}_1; \dots; \mathbf{X}_T]$, where each $\mathbf{X}_t$ is a $b \times m_t$ matrix, with $b = \sum_{j=1}^p b_j$, defined as

$$\mathbf{X}_t = \mathbf{X}_{1,t} \oplus \mathbf{X}_{2,t} \oplus \dots \oplus \mathbf{X}_{p,t}.$$

Each component matrix $\mathbf{X}_{j,t}$ is constructed as

$$\mathbf{X}_{j,t} = [\mathbf{X}_j(s_{1,t}, t), \dots, \mathbf{X}_j(s_{m_{j,t},t}, t)],$$

with $\mathbf{X}_{j,t}$ being a $b_j \times m_{j,t}$ matrix, and each column vector $\mathbf{X}_j(s, t) \in \mathbb{R}^{b_j}$ representing the covariate values associated with the j -th component of $\mathbf{y}(s, t)$ at location 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 $L(\Pi; \mathbf{y}_{1:T}, \mathbf{X}_{1:T}, \mathbf{z}_{1:T}) = p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T} | \mathbf{X}_{1:T}; \Pi)$ be the full likelihood function. The MLE of Π is obtained by maximising the marginal likelihood of the observed data likelihood

$$p(\mathbf{y}_{1:T} | \mathbf{X}_{1:T}; \Pi) = \int p(\mathbf{y}_{1:T} | \mathbf{X}_{1:T}, \mathbf{z}_{1:T}, \Pi) p(\mathbf{z}_{1:T} | \Pi) d\mathbf{z}_{1:T}. \quad (\text{B.13})$$

However, this quantity is intractable since $\mathbf{z}_{1:T}$ 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 $Q(\Pi | \Pi^{(k)})$ as the expected value of the log-likelihood function of Π , with respect to the current conditional distribution of $\mathbf{z}_{1:T}$ given $(\mathbf{y}_{1:T}, \mathbf{X}_{1:T})$ and the current estimates of the parameters $\Pi^{(k)}$:

$$Q(\Pi | \Pi^{(k)}) = \mathbb{E}_{\mathbf{z}_{1:T} \sim p(\cdot | \mathbf{y}_{1:T}, \mathbf{X}_{1:T}; \Pi^{(k)})} [\log p(\mathbf{y}_{1:T} | \mathbf{X}_{1:T}, \mathbf{z}_{1:T}; \Pi)].$$

M-step Obtain the next value of Π , namely $\Pi^{(k+1)}$ as the maximiser of $Q(\Pi | \Pi^{(k)})$:

$$\Pi^{(k+1)} = \arg \max_{\Pi \in \mathcal{P}} Q(\Pi | \Pi^{(k)}), \quad (\text{B.14})$$

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 $\mathbf{z}_{1:T}$ 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 $\{\beta, \sigma^2, f, w\}$ and express the objective function that we need to numerically minimise to obtain an estimate of the value of κ_i for $i = 1, \dots, q$. First, we write explicitly the full log-likelihood $\ell(\Pi; \mathbf{y}_{1:T}, \mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \log p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T} | \mathbf{x}_{1:T}; \Pi)$. Denote by w_{ij} the element at row i and column j of the matrix $\mathbf{W}$ in (6.9) and define the matrix Ψ_t^w as

$$\Psi_t^w = \begin{bmatrix} w_{11} \Psi_{11,t} & \dots & w_{1q} \Psi_{1q,t} \\ \vdots & \ddots & \vdots \\ w_{p1} \Psi_{p1,t} & \dots & w_{pq} \Psi_{pq,t} \end{bmatrix}, \quad (\text{B.15})$$

where $\Psi_{ij,t}$ is a known basis matrix that characterises the effect of the vector $\mathbf{z}_j(\mathcal{V}_j, t)$ on $\mathbf{y}_i(\mathcal{S}_{i,t}, t)$, that is $\Psi_{ij,t} = [\psi_R(s_{1,t})'; \dots; \psi_R(s_{m_{i,t},t})']$. Using the notation introduced in Section 6.5, we can write the SSM (6.9)-(6.10) in matrix form as

$$\begin{aligned} \mathbf{y}_t &= \mathbf{X}_t' \beta + \Psi_t^w \mathbf{z}_t + \varepsilon_t \\ \mathbf{z}_t &= \mathbf{F} \mathbf{z}_{t-1} + \eta_t, \end{aligned}$$

where $\varepsilon_t = (\varepsilon_1(\mathcal{S}_{1,t}, t)', \dots, \varepsilon_p(\mathcal{S}_{p,t}, t))'$ with $\varepsilon_i(\mathcal{S}_{i,t}, t) = (\varepsilon_i(s_{1,t}, t), \dots, \varepsilon_i(s_{m_{i,t},t}, t))'$. Thus, the full log-likelihood $\ell(\Pi; \mathbf{y}_{1:T}, \mathbf{X}_{1:T}, \mathbf{z}_{1:T})$ satisfies

$$\begin{aligned} -2\ell(\Pi; \mathbf{y}_{1:T}, \mathbf{X}_{1:T}, \mathbf{z}_{1:T}) &= \log |\Sigma_0| + (\mathbf{z}_0 - \mu_0)' \Sigma_0^{-1} (\mathbf{z}_0 - \mu_0) \\ &+ \sum_{t=1}^T \left(\log |\Sigma_t| + \mathbf{e}_t' \Sigma_t^{-1} \mathbf{e}_t \right) - T \log |\mathbf{Q}_\kappa| + \sum_{t=1}^T (\mathbf{z}_t - \mathbf{F} \mathbf{z}_{t-1})' \mathbf{Q}_\kappa (\mathbf{z}_t - \mathbf{F} \mathbf{z}_{t-1}), \end{aligned}$$

where $\mathbf{e}_t = (\mathbf{y}_t - \mathbf{X}_t' \beta - \Psi_t^w \mathbf{z}_t)$, and $\Sigma_t = (\sigma_1 \mathbf{I}_{m_{1,t}}) \oplus \dots \oplus (\sigma_p \mathbf{I}_{m_{p,t}})$ is the covariance matrix of ε_t , and $\mathbf{Q}_\kappa = \mathbf{Q}_{\kappa_1} \oplus \dots \oplus \mathbf{Q}_{\kappa_q}$ is the precision matrix of the innovation term $\mathbf{z}_t - \mathbf{F} \mathbf{z}_{t-1}$, the matrix $\mathbf{F} = (f_1 \mathbf{I}_{R_1}) \oplus \dots \oplus (f_q \mathbf{I}_{R_q})$ corresponds to the transition matrix and μ_0 and Σ_0 are the vector mean and covariance matrix of $\mathbf{z}_0$, 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

$$\mathcal{Q}(\Pi | \Pi^{(k)}) = \mathbb{E}_{\Pi^{(k)}} [-2\ell(\Pi; \mathbf{y}_{1:T}, \mathbf{x}_{1:T}, \mathbf{z}_{1:T}) | \mathbf{y}_{1:T}, \mathbf{X}_{1:T}] \quad (\text{B.16})$$

where $\mathbb{E}_{\Pi^{(k)}}$ stands for the expected value considering the posterior distribution of the latent variables $\mathbf{z}_{1:T}$ given the observed data $(\mathbf{y}_{1:T}, \mathbf{X}_{1:T})$ and the current parameter estimate $\Pi^{(k)}$.

Applying the conditional expectation, it follows that

$$\begin{aligned}
 Q(\Pi|\Pi^{(k)}) &= \log |\Sigma_0| + \text{Tr} \left[\Sigma_0^{-1} \left((z_0^T - \mu_0)(z_0^T - \mu_0)' + P_0^T \right) \right] \\
 &+ \sum_{t=1}^T \log |\Sigma_t| + \sum_{t=1}^T \text{Tr} \left[\Sigma_t^{-1} \left((y_t - X_t' \beta - \Psi_t^w z_t^T) (y_t - X_t' \beta - \Psi_t^w z_t^T)' + \Psi_t^w P_t^T \Psi_t^w \right) \right] \\
 &- T \log |Q_\kappa| + \text{Tr} \left[Q_\kappa \sum_{t=1}^T (z_t^T - F z_{t-1}^T)(z_t^T - F z_{t-1}^T)' \right] \\
 &+ \text{Tr} \left[Q_\kappa \sum_{t=1}^T (P_t^T + F P_{t-1}^T F' - F P_{t,t-1}^T - (P_{t,t-1}^T)' F') \right]
 \end{aligned} \tag{B.17}$$

where the quantities z_t^t and P_t^t for $t = 0, \dots, T$ are obtained sequentially using the Kalman Filter equations, whose forward pass for $t = 1, \dots, T$ consists of:

$$\begin{aligned}
 \textbf{Prediction:} \quad z_t^{t-1} &= F^{(k)} z_{t-1}^{t-1}, \\
 P_t^{t-1} &= F^{(k)} P_{t-1}^{t-1} F^{(k)'} + Q_{\kappa^{(k)}}^{-1}, \\
 \text{with} \quad z_0^0 &= \mu_0^{(k)}, \quad P_0^0 = \Sigma_0^{(k)}.
 \end{aligned}$$

$$\begin{aligned}
 \textbf{Update (Filter):} \quad \epsilon_t &= y_t - X_t' \beta^{(k)} - \Psi_t^w z_t^{t-1}, \\
 \Sigma_{\epsilon_t} &= \Psi_t^w P_t^{t-1} (\Psi_t^w)' + \Sigma_t, \\
 K_t &= P_t^{t-1} (\Psi_t^w)' (\Sigma_{\epsilon_t})^{-1}, \\
 z_t^t &= z_t^{t-1} + K_t \epsilon_t, \\
 P_t^t &= P_t^{t-1} - K_t \Psi_t^w P_t^{t-1}.
 \end{aligned}$$

and where the quantities z_t^T and P_t^T for $t = 0, \dots, T$ are obtained sequentially using the Kalman Smoothing equations, whose backwards pass for $t = T, \dots, 1$ 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 $t = 1, \dots, T$ is also obtain sequentially for $t = T, \dots, 2$ 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 $i = 1, \dots, q$.

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 $i = 1, \dots, p$, define $\Psi_{i,t} = \Psi_{i1,t} \oplus \dots \oplus \Psi_{iq,t}$ that is a $(qm_{i,t}) \times (qR)$ matrix.

We obtain:

$$\beta = \left(\sum_{t=1}^T \mathbf{X}'_t (\Sigma_t^{(k)})^{-1} \mathbf{X}_t \right)^{-1} \left(\sum_{t=1}^T \mathbf{X}'_t (\Sigma_t^{(k)})^{-1} (\mathbf{y}_t - \Psi_t^{w^{(k)}} \mathbf{z}_t^T) \right) \quad (\text{B.18})$$

For $i = 1, \dots, p$, update $(\sigma_i^2)^{(k+1)}$ as:

$$\begin{aligned} (\sigma_i^2)^{(k+1)} = & \left(\sum_{t=1}^T m_{i,t} \right)^{-1} \sum_{t=1}^T \text{Tr} \left\{ \left(\mathbf{y}_{i,t} - \mathbf{X}'_{i,t} \beta^{(k)} - \begin{bmatrix} w_{i1}^{(k)} \Psi_{i1}^t, \dots, w_{iq}^{(k)} \Psi_{iq}^t \end{bmatrix} \mathbf{z}_t^T \right) \right. \\ & \times \left(\mathbf{y}_{i,t} - \mathbf{X}'_{i,t} \beta^{(k)} - \begin{bmatrix} w_{i1}^{(k)} \Psi_{i1}^t, \dots, w_{iq}^{(k)} \Psi_{iq}^t \end{bmatrix} \mathbf{z}_t^T \right)' \\ & \left. + \begin{bmatrix} w_{i1}^{(k)} \Psi_{i1}^t, \dots, w_{iq}^{(k)} \Psi_{iq}^t \end{bmatrix} \mathbf{P}_t^T \begin{bmatrix} w_{i1}^{(k)} \Psi_{i1}^t, \dots, w_{iq}^{(k)} \Psi_{iq}^t \end{bmatrix}' \right\}. \end{aligned} \quad (\text{B.19})$$

Continuing, we have

$$f_i = \frac{\text{Tr} \left[\mathbf{Q}_{\kappa_i^{(k)}} \sum_{t=1}^T (e'_i \otimes \mathbf{I}_R) (\mathbf{P}_{t,t-1}^T + \mathbf{z}_{t-1}^T \mathbf{z}_t^{T'}) (e_i \otimes \mathbf{I}_R) \right]}{\text{Tr} \left[\mathbf{Q}_{\kappa_i^{(k)}} \sum_{t=1}^T (e'_i \otimes \mathbf{I}_R) (\mathbf{P}_{t-1}^T + \mathbf{z}_{t-1}^T \mathbf{z}_{t-1}^{T'}) (e_i \otimes \mathbf{I}_R) \right]}. \quad (\text{B.20})$$

Moreover,

$$\mu_0 = \mathbf{z}_0^T, \quad \text{and} \quad \Sigma_0 = \mathbf{P}_0^T.$$

Finally, to derive the closed-form expression for w , proceed as follows.

Let us consider the quantity in the likelihood involving the only $w_i = (w_{i1}, \dots, w_{iq})'$, that is:

$$\sum_{t=1}^T \mathbb{E} \left[\left\| \mathbf{y}(\mathcal{S}_{i,t}, t) - \mathbf{X}^i(\mathcal{S}_{i,t}, t) - \sum_{j=1}^q w_{ij} (e'_j \otimes \mathbf{I}_{m_{i,t}}) \Psi_{i,t} \mathbf{z}_t \right\|^2 \middle| \mathbf{y}_{1:T}, \mathbf{X}_{1:T} \right],$$

where $\|\cdot\|$ denotes the euclidean norm. Observe that

$$\sum_{j=1}^q w_{ij} (e'_j \otimes \mathbf{I}_{m_{i,t}}) \Psi_{i,t} \mathbf{z}_t = \begin{bmatrix} (e'_1 \otimes \mathbf{I}_{m_{i,t}}) \Psi_{i,t} \mathbf{z}_t, \dots, (e'_j \otimes \mathbf{I}_{m_{i,t}}) \Psi_{i,t} \mathbf{z}_t \end{bmatrix} w_i.$$

Therefore, we have to minimise in $\mathbf{w}_i$ the quantity

$$\begin{aligned}
 & -2 \sum_{t=1}^T \left(\mathbf{y}(\mathcal{S}_{i,t}, t) - \mathbf{X}^i(\mathcal{S}_{i,t}, t) \right)' \mathbb{E} \left[(\mathbf{e}'_1 \otimes \mathbf{I}_{m_{i,t}}) \boldsymbol{\Psi}_{i,t} \mathbf{z}_t, \dots, (\mathbf{e}'_j \otimes \mathbf{I}_{m_{i,t}}) \boldsymbol{\Psi}_{i,t} \mathbf{z}_t \middle| \mathbf{y}_{1:T}, \mathbf{X}_{1:T} \right] \mathbf{w}_i \\
 & + \mathbf{w}'_i \mathbb{E} \left[\left[(\mathbf{e}'_1 \otimes \mathbf{I}_{m_{i,t}}) \boldsymbol{\Psi}_{i,t} \mathbf{z}_t, \dots, (\mathbf{e}'_j \otimes \mathbf{I}_{m_{i,t}}) \boldsymbol{\Psi}_{i,t} \mathbf{z}_t \right]' \right. \\
 & \quad \left. \times \left[(\mathbf{e}'_1 \otimes \mathbf{I}_{m_{i,t}}) \boldsymbol{\Psi}_{i,t} \mathbf{z}_t, \dots, (\mathbf{e}'_j \otimes \mathbf{I}_{m_{i,t}}) \boldsymbol{\Psi}_{i,t} \mathbf{z}_t \right] \middle| \mathbf{y}_{1:T}, \mathbf{X}_{1:T} \right] \mathbf{w}_i.
 \end{aligned} \tag{B.21}$$

The first term in (B.21) is equivalent to $-2\mathbf{g}'_i \mathbf{w}_i$, where

$$\mathbf{g}'_i = \sum_{t=1}^T \left(\mathbf{y}(\mathcal{S}_{i,t}, t) - \mathbf{X}^i(\mathcal{S}_{i,t}, t) \boldsymbol{\beta}^{(k)} \right)' \left[(\mathbf{e}'_1 \otimes \mathbf{I}_{m_{i,t}}) \boldsymbol{\Psi}_{i,t} \mathbf{z}_t^T, \dots, (\mathbf{e}'_q \otimes \mathbf{I}_{m_{i,t}}) \boldsymbol{\Psi}_{i,t} \mathbf{z}_t^T \right]. \tag{B.22}$$

The second term in (B.21) is equivalent to $\mathbf{w}'_i \mathbf{R}_i \mathbf{w}_i$, where $\mathbf{R}_i = [r_{i,jk}]_{j,k=1}^q$ is a $q \times q$ matrix with

$$\begin{aligned}
 r_{i,jk} &= \sum_{t=1}^T \mathbb{E} \left[\mathbf{z}'_t \boldsymbol{\Psi}'_{i,t} (\mathbf{e}'_j \otimes \mathbf{I}_{m_{i,t}})' (\mathbf{e}'_k \otimes \mathbf{I}_{m_{i,t}}) \boldsymbol{\Psi}_{i,t} \mathbf{z}_t \middle| \mathbf{y}_{1:T}, \mathbf{X}_{1:T} \right] \\
 &= \sum_{t=1}^T \mathbf{z}_t^{T'} \boldsymbol{\Psi}'_{i,t} (\mathbf{e}'_j \otimes \mathbf{I}_{m_{i,t}})' (\mathbf{e}'_k \otimes \mathbf{I}_{m_{i,t}}) \boldsymbol{\Psi}_{i,t} \mathbf{z}_t^T + \text{Tr} \left(\boldsymbol{\Psi}'_{i,t} (\mathbf{e}'_j \otimes \mathbf{I}_{m_{i,t}})' (\mathbf{e}'_k \otimes \mathbf{I}_{m_{i,t}}) \boldsymbol{\Psi}_{i,t} \mathbf{P}_t^T \right) \\
 &= \sum_{t=1}^T \text{Tr} \left(\boldsymbol{\Psi}'_{i,t} ((\mathbf{e}_j \mathbf{e}'_k) \otimes \mathbf{I}_{m_{i,t}}) \boldsymbol{\Psi}_{i,t} \left(\mathbf{z}_t^T \mathbf{z}_t^{T'} + \mathbf{P}_t^T \right) \right).
 \end{aligned}$$

Therefore, we have to minimise

$$-2\mathbf{g}'_i \mathbf{w}_i + \mathbf{w}'_i \mathbf{R}_i \mathbf{w}_i, \tag{B.23}$$

and therefore

$$\mathbf{w}_i^{(k+1)} = \mathbf{R}_i^{-1} \mathbf{g}_i \tag{B.24}$$

For the updating the value of κ_i we numerically minimise the the function $g_i^{(k)} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ defined as

$$\begin{aligned}
 g_i(\kappa_i) &= -T \log |\mathbf{Q}_{\kappa_i}| + \text{Tr} \left[\mathbf{Q}_{\kappa_i} \sum_{t=1}^T (\mathbf{e}'_i \otimes \mathbf{I}_R) (\mathbf{z}_t^T - \mathbf{F}^{(k)} \mathbf{z}_{t-1}^T) (\mathbf{z}_t^T - \mathbf{F}^{(k)} \mathbf{z}_{t-1}^T)' (\mathbf{e}_i \otimes \mathbf{I}_R) \right] \\
 &+ \text{Tr} \left[\mathbf{Q}_{\kappa_i} \sum_{t=1}^T (\mathbf{e}'_i \otimes \mathbf{I}_R) (\mathbf{P}_t^T + \mathbf{F}^{(k)} \mathbf{P}_{t-1}^T \mathbf{F}^{(k)'} - \mathbf{F}^{(k)} \mathbf{P}_{t,t-1}^T - (\mathbf{P}_{t,t-1}^T)' \mathbf{F}^{(k)'}) (\mathbf{e}_i \otimes \mathbf{I}_R) \right],
 \end{aligned}$$

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 $\eta(s, t)$ as in Finn Lindgren, Håvard Rue, and Lindström (2011). In particular, consider the SPDE:

$$\begin{aligned} (\kappa^2 - \Delta)^{\alpha/2} x(\mathbf{u}) &= \mathcal{W}(\mathbf{u}), \quad \mathbf{u} \in \mathcal{D}, \\ \partial_n (\kappa^2 - \Delta)^j x(\mathbf{u}) &= 0, \quad \mathbf{u} \in \partial\mathcal{D}, \quad j = 0, 1, \dots, \left\lfloor \frac{\alpha - 1}{2} \right\rfloor, \end{aligned} \quad (\text{B.25})$$

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 5. *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*

$$\text{cov}\{x(u), x(v)\} = \sum_{k=-\infty}^{\infty} \{r_M(u, v - 2kL) + r_M(u, 2kL - v)\},$$

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 $\mathbf{Q}_{k_i}$ in Section 6.5.2.

List of Figures

- 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 $PM_{2.5}$ 7
- 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. 8
- 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. 8
- Figure 4: Matérn covariance in the left and the associated variogram functions in the right. $\nu = 0.5$ corresponds to the exponential covariance while larger values of ν yield smoother sample realisation. 17
- Figure 5: An example of the bivariate spatial process $\mathbf{z}(\mathbf{s}) = (z_1(\mathbf{s}), z_2(\mathbf{s}))$ defined on $\mathcal{D} = [0, 1]^2$, with an intrinsic multivariate covariance model $C(\mathbf{h}) = \mathbf{V}\rho(\mathbf{h}; \vartheta)$, 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 $\rho(\mathbf{h}; \vartheta)$, parametrised by $\vartheta = \sigma^2 = 2, \nu = 0.5, k = 0.2$ 20

- Figure 6: An example of the bivariate spatial process $\mathbf{z}(\mathbf{s}) = (z_1(\mathbf{s}), z_2(\mathbf{s}))$ defined on $\mathcal{D} = [0, 1]^2$, with an intrinsic multivariate covariance model $C(\mathbf{h}) = \mathbf{V}, \rho(\mathbf{h}; \boldsymbol{\theta})$, 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 $\rho(\mathbf{h}; \boldsymbol{\theta})$, parametrised by $\boldsymbol{\theta} = \sigma^2 = 2, \nu = 0.5, k = 0.2$ 21
- Figure 7: An example of the SRE data model (2.16). Left: the observed data $(y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))$ with $n = 100$ and $\mathcal{D} = [0, 100]^2$. Centre: the GP $z(\mathbf{s})$ with $\mathbf{s} \in \mathcal{D}$ and $c(\mathbf{h}; \boldsymbol{\theta})$ is the Matérn covariance function parametrised by $\boldsymbol{\theta} = \{\sigma^2 = 1, \nu = 1, k = 40\}$. Right: the measurement error $(\epsilon(\mathbf{s}_1), \dots, \epsilon(\mathbf{s}_n))$. . 29
- Figure 8: Simulation of the univariate state equation Eq. (2.24). Sample data $\{z(\mathbf{s}_1, t), \dots, z(\mathbf{s}_n, t)\}$ for $t \in \{0, \dots, 4\}$ are show from left to right. $z(\mathbf{s}, 0) \sim N(0, I_n)$, $f = 0.2$. The innovation covariance matrix $\mathbf{Q}$ is given by the stationary Matérn function parametrised by $\boldsymbol{\theta} = \{\sigma^2 = 1, \nu = 1, \theta = 0.5\}$ 36
- Figure 9: Simulation of the univariate state equation Eq. (2.24). Sample data $\{z(\mathbf{s}_1, t), \dots, z(\mathbf{s}_n, t)\}$ for $t \in \{0, \dots, 4\}$ are show from left to right. $z(\mathbf{s}, 0) \sim N(0, I_n)$, $f = 0.9$. The innovation covariance matrix $\mathbf{Q}$ is given by the stationary Matérn function parametrised by $\boldsymbol{\theta} = \{\sigma^2 = 1, \nu = 1, \theta = 0.5\}$ 36
- 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. 40
- 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. 41
- 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. 44
- 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. 51

List of Figures

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.	56
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.	58
Figure 16:	Livestock density over the augmented Lombardy region on 31 December 2021. Left: swine density, right: bovine density. 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 star). Based on this, stations in the same municipality share the same municipality centroid, so they have the same livestock time series.	61
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. The NH_3 air quality network is depicted by magenta crosses.	65
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).	66
Figure 19:	Spatio-temporal variogram computed on $\text{PM}_{2.5}$ daily observations at 45 stations from 2016 to 2020.	72
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 the i -th lag.	72
Figure 21:	In-sample residual diagnostics for the three models. Standard deviation of residuals grouped by months for each model.	80
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.	83
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 $\text{PM}_{2.5}$ concentrations and the corresponding regressors.	86

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.	88
Figure 25:	Variable importance factor (measured as IncMSE) for the 11 selected features in the large-scale component of the RFSTK.	89
Figure 26:	Estimated error variance, $\hat{\sigma}_{\epsilon,t}^2$	99
Figure 27:	Studentised residuals grouped by station. The outliers (red cross) are identified by the Whisker length based on the first and third quartiles.	99
Figure 28:	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.	100
Figure 29:	PRIA Scenario (-26%). Left: average $PM_{2.5}$ reduction in winter, $\bar{\Delta}_{\hat{y}}$ (452 daily observations for each pixel). Right: uncertainty associated with the average reduction ($\text{std}(\bar{\Delta}_{\hat{y}})$). The pink stars depict the provincial capitals listed in Table 15. The model is only evaluated in non-forested areas below 640 m altitude. . . .	101
Figure 30:	Strong Scenario (-50%). Left: winter $\bar{\Delta}_{\hat{y}}$ (452 daily observations for each pixel). Right: uncertainty associated with the average reduction ($\text{std}(\bar{\Delta}_{\hat{y}})$). The pink stars depict the provincial capitals listed in Table 15. The model is only evaluated in non-forested areas below 640 m altitude.	101
Figure 31:	Scenario PRIA (-26%). Boxplot of daily $\Delta_{\hat{y}}$ in winter (452 daily observations for each pixel). Left: grouped by province; Right: 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.	102
Figure 32:	Scenario Strong (-50%). Boxplot of daily $\Delta_{\hat{y}}$ in winter (452 daily observations for each pixel). Left: grouped by province; Right: 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.	102
Figure 33:	Left: $\hat{\sigma}_{\epsilon}(t)^2$ for the $PM_{2.5}$ (blue curve) compared with the empirical daily variance of the residual (orange line). Right: $\hat{\sigma}_{\epsilon}(t)^2$ for the PM_{10} (blue curve) compared with the empirical daily variance of the residual (orange line).	105
Figure 34:	$\hat{\sigma}_{\epsilon}(t)^2$ for $PM_{2.5}$ (blue curve) and PM_{10} (orange curve).	106

- Figure 35: Schematic workflow of the proposed low-rank SPDE state-space model. The SPDE is discretised via FEM to construct a low-rank latent state-space representation, temporal dependence is handled through Kalman filtering and smoothing, and parameters are estimated via an EM algorithm. 121
- Figure 36: 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. 127
- Figure 37: Monitoring network and overall averages concentration over Italy of PM_{10} on the left and $PM_{2.5}$ on the right. 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. 133
- Figure 38: 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} on the left, the variogram of $PM_{2.5}$ on the centre, and the cross-covariogram between PM_{10} and $PM_{2.5}$ on the right. The fitted Matérn models are estimated with weighted least squares (Cressie, 1985). 133
- Figure 39: Temporal structure of particulate matter concentrations. The left column shows daily averages of PM_{10} on the top and $PM_{2.5}$ on the bottom; the right column displays corresponding ACFs over the first 30 lags, summarised across stations using boxplots. 134
- Figure 40: Spatial meshes generated via Laplacian smoothing. Left: mesh constructed for the PM_{10} dataset; right: mesh constructed for the $PM_{2.5}$ dataset. Interior vertices (magenta), boundary vertices (green), convex hulls (blue), inner buffer (black), and validation locations (red) are shown. 135
- Figure 41: Estimated average latent effect component on PM_{10} on the left, and the associated average standard deviation on the right. 137
- Figure 42: $LR = 75\%$. Estimated average latent effect component on PM_{10} on the left, associated with the average standard deviation on the centre and the average difference between $LR = 100\%$ and $LR = 75\%$ estimations on the right. 138

Figure 43:	$LR = 50\%$. Estimated average latent effect component on PM_{10} on the left, associated with the average standard deviation on the centre and the average difference between $LR = 100\%$ and $LR = 50\%$ estimations on the right.	138
Figure 44:	$LR = 25\%$. Estimated average latent effect component on PM_{10} on the left, associated with the average standard deviation on the centre and the average difference between $LR = 100\%$ and $LR = 25\%$ estimations on the right.	138
Figure 45:	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), $RMSE = 0.1710$	158
Figure 46:	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).	158
Figure 47:	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, $RMSE = 0.3078$	159
Figure 48:	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).	159
Figure 49:	Piecewise cubic spline, PCHIP and Makima interpolation methods applied to swine time series for the monitoring station named ‘Corte de Cortesi’.	161
Figure 50:	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.	161

List of Tables

Table 1:	Main classes of covariance functions and construction principles used in spatial and spatio-temporal modelling. The categories describe complementary properties and are not mutually exclusive.	27
Table 2:	Input sources dataset used to build the Agrimonia dataset.	52
Table 3:	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.	53
Table 4:	WE variables selected from ERA5 datasets. <i>Categ.</i> stands for Categorical. . . .	54
Table 5:	EM variables selected from the CAMS-anthropogenic emissions dataset. Spatial resolution of (0.1×0.1) degrees. Unit is $kg/(m^2s)$	55
Table 6:	LA variables selected from the ERA5, CLC and SIARL datasets. <i>Categ.</i> stands for Categorical.	57
Table 7:	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 3 only. . .	59
Table 8:	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. <i>Categ.</i> stands for Categorical.	60

Table 9:	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.	61
Table 10:	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).	62
Table 11:	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). <i>Categ.</i> stands for Categorical.	62
Table 12:	Summary of validation procedures and methodological assumptions adopted for each component of the Agrimonia dataset.	63
Table 13:	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.	71
Table 14:	Seasonal average of daily Temperature [C°], NH_3 , NO_x and SO_2 emissions [mg/m^2] and $PM_{2.5}$ concentrations [$\mu g/m^3$].	73
Table 15:	Total NH_3 emissions by province in 2020.	73
Table 16:	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).	79
Table 17:	Prediction performance indices evaluated with the LOSOCV scheme [μgm^{-3}] . . .	81
Table 18:	Estimated coefficients of the large-scale component of the HDGM model	85
Table 19:	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).	87
Table 20:	Estimate of the fixed effect coefficients of the model (5.7). $ t $ is the absolute value of the t -statistic.	98

Table 21:	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 $\text{PM}_{2.5}$ concentrations; the columns headed PRIA and Strong show the average reduction and its uncertainty, in brackets, computed using Eq. (5.25) for both scenarios.	103
Table 22:	$\text{PM}_{2.5}$ 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 $\text{PM}_{2.5}$ 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.	103
Table 23:	Estimate of the fixed effect coefficients β of the bi-variate model.	105
Table 24:	$LR = 100\%$. 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. <i>Avg</i> and <i>Std</i> denote the average and standard deviation, respectively, across replications. <i>Min angle</i> indicates the average of the minimum internal angles across all the Monte Carlo simulations, measured in degrees, while <i>Time</i> indicates the average computation time over replications in seconds.	130
Table 25:	$LR = 50\%$. 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. <i>Avg</i> and <i>Std</i> denote the average and standard deviation, respectively, across replications. <i>Min angle</i> indicates the average of the minimum internal angles across all the Monte Carlo simulations, measured in degrees, while <i>Time</i> indicates the average computation time over replications in seconds.	131

Table 26:	$LR = 15\%$. 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. Avg and Std denote the average and standard deviation, respectively, across replications. $Min\ angle$ indicates the average of the minimum internal angles across all the Monte Carlo simulations, measured in degrees, while $Time$ indicates the average computation time over replications in seconds.	132
Table 27:	Summary statistics of variables from the AQCLIM-GRINS dataset across all locations.	134
Table 28:	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. .	136
Table 29:	ML estimates of regression coefficients and random effect parameters for the bi-variate 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.. . . .	137
Table 30:	Comparative performance analysis on the GRINS dataset subset. The table reports average Training and Test RMSE (mg/m^3) and computational cost for the proposed Low-Rank (LR) approach (at varying percentages), the benchmark INLA state-space implementation (M_{INLA}), and the diffusion-based SPDE model ($M_{\text{INLA},102}$). Results are stratified by time series length $T \in \{15, 30, 60\}$. Computational time is expressed in seconds and as a percentage relative to $M_{LR,100\%}$. Values represent averages over 20 random repetitions, with standard deviations provided in parentheses.	141
Table 31:	Extremely large values for the AQ variables [$\mu g/m^3$] identified in the dataset. The values are replaced with the ‘NaN’ value.	158

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](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.
- Anderes, Ethan, Jesper Møller, and Jakob G Rasmussen (2020). “Isotropic covariance functions on graphs and their edges”. In.
- Andrej, Julian et al. (2024). “High-performance finite elements with MFEM”. In: *The International Journal of High Performance Computing Applications* 38.5, pp. 447–467.
- 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>.
- Baddeley, Adrian et al. (2017). ““Stationary” point processes are uncommon on linear networks”. In: *Stat* 6.1, pp. 68–78.
- 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 Drăguț (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.
- Bolin, David, Alexandre B Simas, and Jonas Wallin (2024). “Gaussian Whittle–Matérn fields on metric graphs”. In: *Bernoulli* 30.2, pp. 1611–1639.
- 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 NH₃ 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.
- (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.
- Datta, Abhirup et al. (2016). “Hierarchical nearest-neighbor Gaussian process models for large geo-statistical 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.
- Defense Mapping Agency Washington DC (1991). “Department of defense world geodetic system 1984: its definition and relationships with local geodetic systems. Second edition.” In: *Department of Defense, NIMA USA*. URL: %5Curl%7Bhttps://apps.dtic.mil/sti/pdfs/ADA280358.pdf%7D.
- 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.
- EEA (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.
- (2021). *Data Air Quality e-Reporting (AQ e-Reporting)*. European Environmental Agency portal https://www.eea.europa.eu/ds_resolveuid/d4b3817c7a1640459e783f9342aac786.

- (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>.
- 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 (Nov. 2022). *Air quality in Europe*. Last Access: 18-03-2023.
- 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 Ndonga (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](https://doi.org/10.5281/zenodo.7956006). URL: <https://doi.org/10.5281/zenodo.7956006>.

- Ferkingstad, Egil, Leonhard Held, and Håvard Rue (2017). “Fast and accurate Bayesian model criticism and conflict diagnostics using R-INLA”. In: *Stat* 6.1, pp. 331–344. doi: <https://doi.org/10.1002/sta4.163>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/sta4.163>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.163>.
- Field, David A (1988). “Laplacian smoothing and Delaunay triangulations”. In: *Communications in applied numerical methods* 4.6, pp. 709–712.
- Filosi, Tobia, Claudio Agostinelli, and Emilio Porcu (2023). “Temporally-Evolving Generalised Networks and their Reproducing Kernels”. In: *arXiv preprint arXiv:2309.15855*.
- 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 PM₁₀ 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.

- Gaedke-Merzhäuser, Lisa et al. (2025). “Accelerated Spatio-Temporal Bayesian Modeling for Multivariate Gaussian Processes”. In: *arXiv preprint arXiv:2507.06938*.
- 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 (NH₃) 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/CB09781107049994>.
- 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.
- Hjelle, Øyvind and Morten Dæhlen (2006). *Triangulations and applications*. Springer Science & Business Media.
- 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](https://doi.org/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 Zootechnica 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](https://doi.org/10.3150/12-BEJSP17). URL: <https://projecteuclid.org/journals/bernoulli/volume-19/issue-4/On-statistics-computation-and-scalability/10.3150/12-BEJSP17.full>.
- Julier, Simon J and Jeffrey K Uhlmann (2004). “Unscented filtering and nonlinear estimation”. In: *Proceedings of the IEEE* 92.3, pp. 401–422.
- 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 PM_{2.5} 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.
- Krainski, Elias et al. (2018). *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC.
- 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, F. et al. (2024). “A diffusion-based spatio-temporal extension of Gaussian Matérn fields”. In: *SORT: Statistics and Operations Research Transactions*.
- Lindgren, Finn, David Bolin, and Håvard Rue (2022). “The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running”. In: *Spatial Statistics* 50, p. 100599.
- 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 NO₂: 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.
- Porcu, Emilio, Philip A White, and Marc G Genton (2023). “Stationary nonseparable space-time covariance functions on networks”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 85.5, pp. 1417–1440.
- 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.
- Ribeiro, Maria Isabel (2004). “Kalman and extended kalman filters: Concept, derivation and properties”. In: *Institute for Systems and Robotics* 43.46, pp. 3736–3741.
- Rodeschini, Jacopo, Alessandro Fassò, et al. (2024). “Scenario analysis of livestock-related PM_{2.5} 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.
- Rue, Håvard, Sara Martino, and Nicolas Chopin (2009). “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 71.2, pp. 319–392.
- Sampson, Paul D and Peter Guttorp (1992). “Nonparametric estimation of nonstationary spatial covariance structure”. In: *Journal of the American Statistical Association* 87.417, pp. 108–119.
- 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 PM_{2.5} 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 PM₁₀ 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 NH₃ 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 NH₃ 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.

- Wan, Eric A and Rudolph Van Der Merwe (2000). “The unscented Kalman filter for nonlinear estimation”. In: *Proceedings of the IEEE 2000 adaptive systems for signal processing, communications, and control symposium (Cat. No. 00EX373)*. Ieee, pp. 153–158.
- 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 PM₁₀), 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.

References

- 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.