UNIVERSITY OF BARI ALDO MORO

Department of Economics and Finance

PH.D. IN ECONOMICS AND FINANCE OF PUBLIC ADMINISTRATIONS
CYCLE XXXVIII
Academic Discipline:
STAT - 01 - Statistics

THESIS TITLE

Statistical Learning for Integrating Environmental and Economic Dimensions of Coastal and Marine Resilience

---

Ph.D. Student: Dr. Lorena Ricciotti

16/01/2026 15:36:40
LORENA RICCIOTTI

Lorena Ricciotti

Supervisor: Prof. Alessio Pollice

Co-supervisor: Prof. Antonello Maruotti

Coordinator: Prof. Alessio Pollice

FINAL EXAM 2026

Funded by the European Union – Next Generation EU, Mission 4 Component 1 CUP H91I22000750007 Investment 4.1 Ministerial Decree no. 351 of 9 April 2022 – National Recovery and Resilience Plan. Project: "Statistical Learning for the Evaluation of Coastal and Marine Ecosystem Services" (approved by the MUR and present on the CINECA platform and approved by the Academic Board).

Contents

Introduction

The twenty-first century will likely be remembered as the time when the global community became fully aware of the profound risks associated with climate change, but failed to respond with the urgency and determination that such risks demanded. One of the most sensitive ecosystems to environmental pressures is the one concerning the marine and coastal habitats. The latter constitutes a wide range of services such as protection against storm surges and flooding, fisheries and aquacultures that sustain local communities, and recreational services. All of these are fundamental for human well-being and economic development. However, their complexity and variability and high vulnerability to climate change present substantial challenges for monitoring, evaluation, and policy design. Statistical learning offers a powerful set of tools to address these challenges. Modern data-driven methods can extract patterns from heterogeneous environmental datasets, deal with non-standard features such as zero-inflation, heavy tails, or complex dependence structures, and integrate information across spatial and temporal scales. By combining methodological innovation with applied case studies, statistical learning can provide actionable knowledge for environmental management and decision-making.

This dissertation contributes to the field of *statistical learning for the evaluation of coastal and marine ecosystem services* through a collection of studies that address different but interconnected aspects of marine and coastal systems. The unifying goal is to develop and apply advanced statistical models that capture the complexity of marine processes and coastal protection investments, thereby supporting both scientific understanding and policy interventions. The first part of this research focuses on marine physical processes, where accurate classification and modelling of sea states and hydrodynamic conditions are essential for risk assessment. In particular, bivariate finite mixture models of generalized additive models are employed to jointly model tides and wave heights, enabling the classification of sea conditions into meaningful clusters. This approach not only provides insight into marine dynamics but also delivers a framework for early-warning systems and adaptive coastal management. The dynamics of the Venice lagoon, a fragile and unique ecosystem under the threat of subsidence and sea-level rise, are then explored using zero-inflated hidden semi-Markov models. By incorporating covariate-dependent sojourn times, this methodology allows a nuanced understanding of lagoon dynamics, accounting for periods of calmness punctuated by extreme events. Such models offer valuable tools for local environmental management and climate adaptation strategies. Further, the dissertation explores the ecological dimension of fisheries, focusing on the variability of sardine landings in the Mediterranean. A Bayesian hierarchical approach, implemented through Integrated Nested Laplace Approximation (INLA), captures spatial and temporal heterogeneity in fishery yields. This application illustrates the relevance of advanced statistical methods for understanding ecosystem services that directly sustain livelihoods and food security.

The second part of the thesis shifts the focus from the physical and ecological characterisation of marine systems to the socio-economic and policy dimensions of coastal resilience, with particular attention to soil and coastal defence. Whereas the first part investigates the dynamics of sea conditions, meteorological tides, and biological productivity, this section examines how society responds to these environmental challenges through the design and financing of protective infrastructure. Understanding the interaction between environmental pressures and financial decisions is essential for developing an integrated framework for climate adaptation, where prevention and mitigation strategies are guided by both scientific evidence and efficient resource allocation.

This part of the work emphasises the need to link public investments with environmental pressures, recognising that resilience cannot be achieved without an informed and equitable distribution of financial resources. Despite the relevance of soil-defence and coastal-protection funding in Italy, the lack of harmonised, spatially explicit data has long hindered systematic analysis. To address this gap, a comprehensive and georeferenced dataset of Italian public investments for soil defence is constructed and analysed, providing for the first time a harmonised view of spatial and temporal funding patterns across the national territory. The integrated database enables the exploration of how public expenditure aligns with hydrogeological vulnerability, erosion intensity, coastal morphology, and patterns of land consumption along the shore. In this way, the work contributes not only to statistical modelling but also to the broader discourse on environmental governance and territorial equity. Building upon this dataset, a dedicated R package, *PublicWorksFinanceIT*, is developed to streamline data retrieval, integration, and visualisation. The package automates access to multiple open administrative data sources, links them through a unique project identifier, and enriches them with detailed spatial attributes at the regional, provincial, and municipal levels. By enabling reproducible workflows and transparent analyses, it provides a publicly available tool for researchers and policymakers, fostering data-driven approaches to environmental planning and financial accountability. Moving beyond descriptive assessment, the thesis then introduces advanced spatio-temporal statistical models to quantify the relationship between coastal-defence funds and environmental drivers. These models incorporate spatial heterogeneity and temporal evolution, allowing for the evaluation of how funding responds to risk exposure, climatic stressors, and socio-economic conditions. The modelling framework provides a robust statistical basis for assessing the efficiency and effectiveness of public investments and for identifying potential delays or imbalances in resource allocation. Taken together, the development of the dataset, the R package, and the spatio-temporal models illustrate how methodological innovation can be directly linked to policy evaluation. By integrating statistical learning with environmental and economic data, the thesis demonstrates a comprehensive approach to understanding and governing coastal systems—one that connects the dynamics of the natural environment with the institutional and financial mechanisms that shape its protection and resilience.

Overall, this research contributes to both methodology and application. Methodologically, it advances the use of mixture models, hidden semi-Markov frameworks, and Bayesian hierarchical models for complex marine and environmental data. From an applied perspective, it provides insights into sea state classification, lagoon dynamics, coastal protection investments, and fisheries variability. Together, these studies demonstrate the potential of statistical learning to enhance our understanding of coastal and marine ecosystem services and to inform effective environmental policies.

The thesis is organized as follows: Chapter 1 provides research context and general motivation; Chapters 2-7 present the six research papers that constitute the core of the dissertation; Chapter 8 offers a synthesis of findings and a discussion of methodological and applied implications and concludes with future research perspectives.

Chapter 1

Research Context and Motivation

1.1 Marine and Environmental analysis: data challenges and novel modelling approaches

Marine and environmental data are inherently complex and often present multiple statistical challenges stemming from the natural variability, heterogeneity, and dependence structures that characterize the underlying processes. These data are typically generated by systems that are dynamic, non-linear, and influenced by numerous interacting factors, such as meteorological conditions, oceanographic circulation, and human activities [Marini, 2018, Giorgi, 2006, US EPA, 2016]. As a result, the statistical modelling of such data requires flexible methods capable of capturing these sources of variability and uncertainty.

A first major challenge arises from the focus on *extreme events*, which are of primary importance in environmental applications, such as storm tides, flooding, or anomalous sea states [Canale et al., 2020, Marsooli and Lin, 2018, Foster and Brown, 2015]. These phenomena often display heavy-tailed or skewed distributions and temporal dependence, making their statistical characterization particularly demanding. Moreover, when dealing with environmental count data, analysts frequently encounter excesses of zero counts—a condition known as *zero-inflation* [Feng, 2021]. For instance, in time series representing the daily duration of flooding events in the Venice lagoon, zero counts correspond to days without flooding. In such cases, the standard Poisson model tends to underestimate the frequency of zeros, necessitating more flexible model formulations, such as zero-inflated or hurdle models, or their extensions within hidden or semi-Markov frameworks.

Another pervasive issue concerns *heavy tails and non-normality* in environmental data. Real-world marine observations often include outliers, spurious measurements, or observational noise—sometimes referred to as "bad points." Ignoring these data irregularities can bias parameter estimates and compromise inference. To account for these features, flexible distributions such as the t or skew-t families are frequently adopted [Rigby and Stasinopoulos, 2005, Stasinopoulos et al., 2018, 2017]. The Generalized Additive Models for Location, Scale and Shape (GAMLSS) framework provides a robust approach for this purpose, allowing distributional parameters to depend on covariates while accommodating non-Gaussian behaviour [Stasinopoulos and Rigby, 2007].

Marine and coastal data also exhibit pronounced *heterogeneity and latent structure*, often reflecting the coexistence of different sea regimes or environmental conditions. Finite mixture models have proven effective in capturing such heterogeneity by assuming that observations arise from multiple underlying subpopulations [McLachlan et al., 2019, Huang and Dong, 2020, 2019]. In the marine context, mixture and hidden Markov approaches have been successfully applied to classify sea states, identify wave regimes, and analyse multivariate time series with circular and skewed components [Bulla et al., 2012, Lagona and Picone, 2012]. Such models enable a more accurate representation of environmental complexity, supporting the identification of latent clusters corresponding to distinct physical or meteorological conditions [Morucci et al., 2016, APAT, 2004].

Finally, marine and environmental processes are intrinsically *spatio-temporal*. The variability of environmental indicators—such as wave height, sediment transport, or public investments for coastal protection—typically spans multiple spatial and temporal scales. Analyses of spatially distributed processes, like the allocation of soil-defence funds or the distribution of fish landings, must jointly model structured and unstructured spatial components as well as temporal dynamics [Maruotti and Alaimo Di Loro, 2023, Guillou, 2017, Booij et al., 1999]. Spatio-temporal models thus play a crucial role in understanding how environmental patterns evolve over time and across regions, integrating ecological, socioeconomic, and climatic factors.

In summary, marine and environmental data challenge traditional statistical assumptions and require advanced modelling frameworks—such as mixture models, GAMLSS, hidden Markov processes, and spatio-temporal models—to appropriately capture the complexity, dependence, and heterogeneity of the natural systems they describe.

To address these complex data challenges, several advanced statistical frameworks have been developed and utilized in environmental and marine studies. The General Additive Model for Location, Scale, and Shape (GAMLSS, Rigby and Stasinopoulos [2005]) is a flexible regression approach that relaxes the assumption of the exponential family (common in classical linear regression). In the GAMLSS framework, the systematic part of the model is expanded to allow all parameters of the conditional distribution of the outcome—not just the mean (location) but also the variance (scale) and other shape parameters (μ_{ikj}, σ_{ikj}, ν_{ikj})—to be modelled as functions of independent or latent variables. This allows researchers to jointly investigate how independent variables affect both the expected value and the variability of the data. Finite mixtures of GAMLSSs extend this concept, incorporating a latent structure to model heterogeneity in the outcomes' distributions, enabling the approximation of any data shape by increasing the number of mixture components. The approach assumes that outcomes are conditionally independent given the latent variable (clusters). By defining cluster-specific regression models, this method helps to identify and characterize distinct sea regimes.

Hidden Semi-Markov Models (HSMMs, Yu [2010]) are a class of latent-variable models useful for time series analysis. They approximate the data distribution through a mixture of densities contingent on the evolution of a hidden semi-Markov chain. A key advantage of the HSMM over the standard Hidden Markov Model (HMM) is its ability to explicitly structure the sojourn times (the state-specific duration parameters) of the hidden states. HMMs implicitly assume a geometric distribution for the sojourn time, which may be unrealistic [Zucchini et al., 2017]. By explicitly modelling sojourn times, HSMMs gain enhanced adaptability and precision in capturing real-world complexities and temporal structures. This framework can be applied to handle zero-inflation through a zero-inflated hidden semi-Markov model, allowing for a mix of sampling zeros and structural zeros. HSMMs can be further generalized to non-homogeneous HSMMs by introducing covariate-dependent parameters. This extension enables the use of concomitant-variable models, where covariates (such as a binary variable indicating the activation of the MOSE barrier system) are included in the linear predictors for the observation process parameters and also influence the sojourn time distribution and transition probabilities of the latent states. Parameters for HSMMs are typically estimated using maximum-likelihood estimation, often involving the direct numerical maximization of the log-likelihood function.

The Integrated Nested Laplace Approximation (INLA) is a method used for performing approximate Bayesian inference, particularly effective for latent Gaussian models [Rue et al., 2009, Bakka et al., 2018]. INLA provides significant computational advantages over traditional Markov Chain Monte Carlo (MCMC) methods, offering faster computations and ease of use, making it suitable for high-dimensional spatial and spatio-temporal modelling, which is common in ecological and environmental studies. The INLA framework is useful to model both spatial and temporal components. For spatial analysis of areal data, the spatial component is often modelled using a re-parametrization of the Besag-York-Molliè (BYM) model, which incorporates both a structured (spatially dependent) and an independent (unstructured) component [Riebler et al., 2016b]. For spatially continuous data, the Stochastic Partial Differential Equation (SPDE) approach is used within INLA to provide a computationally efficient approximation of spatially continuous Gaussian random fields [Lindgren et al., 2011]. Model performance and selection within the Bayesian framework are typically assessed using criteria such as the Deviance Information Criterion (DIC, Spiegelhalter et al. [2002], Gelman et al. [1995].) and the Watanabe–Akaike Information Criterion (WAIC, Watanabe [2013]), which balance model fit and complexity, alongside the marginal log-likelihood (m-lik, Gómez-Rubio [2020]).

1.2 Policy and economic background: soil and coastal defence funding in Italy

Public investments in soil and coastal defence represent a central pillar of Italy's environmental and territorial policy framework. These interventions, encompassing hydraulic engineering, slope stabilisation, flood protection, and coastal restoration, aim to safeguard both ecosystems and human settlements from the increasing frequency and intensity of hydrogeological hazards such as erosion, landslides, and flooding. The need for such measures is particularly urgent in Italy, where complex geomorphology, steep terrain, and a long, densely inhabited coastline converge with intensive land use and urbanisation patterns that often developed in the absence of comprehensive spatial planning [Buscemi et al., 2024]. The combination of natural fragility and anthropogenic pressure has left vast areas exposed to risk, prompting sustained public expenditure on soil-defence and coastal-protection works.

Coastal areas, in particular, concentrate a high proportion of Italy's population ($\sim 30\%$), infrastructure, and economic activity, yet are among the most vulnerable to the effects of climate change. Rising sea levels, and the growing frequency of storm surges threaten both natural habitats and socio-economic assets. The Italian National Adaptation Strategy (SNAC) and subsequent National Plan for Adaptation to Climate Change (PNACC) have underscored the urgency of coordinated coastal defence, calling for the integration of physical risk assessments with financial and planning instruments [Ministero dell'Ambiente e della Sicurezza Energetica (MASE), 2015]. In this context, investments in sea defences—such as breakwaters, dikes, beach nourishment, and dune reconstruction—constitute not only protective infrastructure but also essential components of territorial resilience and climate adaptation policy.

Despite the scale and importance of these interventions, analysing and monitoring soil-defence funding in Italy has long been hindered by the fragmentation of public data sources. Financial and administrative information on the same projects is distributed across distinct institutional platforms that differ in scope, format, and update frequency. The most relevant among them are: *OpenCoesione*, which publishes data on European and national cohesion-policy projects; *OpenBDAP*, maintained by the Ministry of Economy and Finance and focused on the budgetary cycle of all public investments; and *ReNDiS*, managed by ISPRA and specifically dedicated to hydrogeological-risk mitigation works. While these databases collectively provide a comprehensive overview of public spending, their heterogeneity in coding systems, spatial references, and financial classifications has limited their combined analytical use. To address these challenges, the *PublicWorksFinanceIT* R package was developed as part of this research effort [Ricciotti and Pollice, 2024a].

The tool automates the retrieval, harmonisation, and integration of data from the aforementioned sources, linking project records through the unique project identifier (*Codice Unico di Progetto*, CUP) and enriching them with detailed geographic information at the regional, provincial, and municipal levels. Each intervention is geo-referenced using official spatial boundaries and centroids, providing a consistent framework for mapping and spatial analysis. This process produces a unified, open, and reproducible database of public investments for soil and coastal defence in Italy—an essential foundation for the statistical learning analyses presented in the later chapters of this thesis. The harmonised dataset allows for an unprecedented level of spatial and temporal detail in evaluating the allocation and effectiveness of soil-defence funds. It supports the exploration of fundamental policy questions: are investments distributed proportionally to environmental risk and exposure? Do they respond to local hazard dynamics or reflect administrative and political factors? How do temporal patterns of funding align with environmental events or changing policy priorities? These questions are central to understanding how societies adapt to environmental pressures and how public finance mechanisms contribute to resilience in coastal and inland territories.

Chapter 2

Modelling and Clustering Sea Conditions: Bivariate Finite Mixtures of Generalized Additive Models for Location, Shape, and Scale Applied to the Analysis of Meteorological Tides and Wave Heights[1]

[1]Ricciotti , L., Picone, M., Pollice, A., & Maruotti, A. (2024). Modelling and Clustering Sea Conditions: Bivariate Finite Mixtures of Generalized Additive Models for Location, Shape, and Scale Applied to the Analysis of Meteorological Tides and Wave Heights. Journal of Marine Science and Engineering, 12(5), 740. https://doi.org/10.3390/jmse12050740

2.1 Paper Summary

This first empirical chapter represents the methodological and thematic foundation of the thesis. Within the broader objective of statistical learning for the evaluation of coastal and marine ecosystems, this work focuses on the physical component of the marine environment—specifically, the dynamics of sea-state conditions and their underlying drivers. Understanding and characterising the variability of sea conditions is necessary for assessing coastal risks, supporting the design of marine infrastructure, and analysing subsequent socio-economic and ecological processes that depend on the marine environment.

Coastal and marine systems are complex, governed by interacting physical, meteorological, and oceanographic processes. The latter jointly determine short-term sea-level fluctuations, which are critical for coastal erosion, and flooding risk. Traditional deterministic models, such as those based on hydrodynamic simulations (e.g., SWAN or ROMS), can reproduce these processes but rely on detailed physical inputs and are computationally demanding. From a statistical-learning perspective, there is growing interest in developing data-driven probabilistic models that can capture both the joint behaviour of sea-state variables and the heterogeneity of the underlying mechanisms.

This chapter, based on Ricciotti et al. [2024], contributes to this effort by proposing a novel statistical framework that simultaneously models and classifies sea conditions through bivariate finite mixtures of Generalized Additive Models for Location, Scale, and Shape (GAMLSS). The central objective of the study is to develop a model capable of (i) describing the joint distribution of meteorological tide and significant wave height; (ii) accounting for the influence of key meteorological covariates; and (iii) revealing unobserved sea-state regimes with distinct characteristics. The empirical application focuses on the Gulf of Ancona along the Italian Adriatic coast, a region frequently affected by storm surges. The analysis is carried out on two different outcomes provided by the Italian Institute for Environmental Protection and Research (ISPRA): sea-level residuals, representing the meteorological tide, from the Rete Mareografica Nazionale (RMN); and significant wave height from the Rete Ondametrica Nazionale (RON). These measurements were coupled with meteorological information including atmospheric pressure, wind speed and direction, and a fetch variable derived from wind direction to quantify the extent of open water exposed to wind forcing.

The proposed unifying methodological framework—based on flexible, probabilistic, and data-driven models— recurs, in various forms, throughout the subsequent chapters. By addressing the physical dimension of marine dynamics, it sets the stage for the more applied and socio-ecological analyses that follow. Conceptually, the latent-regime structure developed here anticipates the use of hidden and state-space models in Chapter 3, where a zero-inflated hidden semi-Markov model is employed to analyse the temporal evolution of marine processes in the Venice Lagoon. Methodologically, the emphasis on covariate-dependent parameters and non-Gaussian distributions resonates with the Bayesian and hierarchical frameworks later used in Chapter 4, which investigates spatial and temporal variability in Mediterranean sardine landings. Furthermore, the insights obtained in this chapter regarding the physical drivers of sea-state variability—wind, pressure, and fetch—provide a natural link to the economic and policy-oriented Chapters 5, 6, and 7. Coastal defence planning, marine resource management, and public investment allocation all depend on an accurate characterisation of the physical pressures acting on the coastal system. Thus, this first study anchors the entire thesis in a physically informed statistical understanding of the marine environment.

Beyond its methodological novelty, this research contributes to the broader field of environmental and marine data science. It illustrates how statistical learning can enrich traditional marine analyses by offering insight into sea-state variability, rather than focusing only on mean relationships, it also helps identifying latent structural patterns that correspond to physically interpretable sea-condition regimes, and provide computationally efficient alternatives to deterministic hydrodynamic simulations. Such approaches align with current efforts in marine science to develop integrated observation–modelling systems that combine physical, ecological, and socio-economic information.

2.2 Introduction

When discussing the impact of climate change on the environment, it is crucial to monitor marine physical parameters as key indicators. Extensive data is available for various measurements related to seas and oceans. In our study, we focus on air pressure, wind, and effective fetch to model their influence on tides and waves. Specifically, analyzing the behavior of these variables helps identifying sea states resulting from meteorological changes. Tides and waves are significantly influenced by air pressure, wind, and fetch. Atmospheric pressure primarily affects tides, with low air pressure corresponding to high tide levels and vice versa. Wind and fetch are closely interconnected: wind speed can either decrease or increase wave heights depending on its direction. The latter is linked to the fetch which represents the stretch of the open sea over which the wind blows, determining wave generation. A greater fetch allows the wind to act upon a larger expanse of the sea surface, resulting in higher waves.

Monitoring these physical parameters over time and incorporating meteorological components as covariates is crucial to comprehend changes in the behavior of the sea over the years. By classifying sea states based on specific weather conditions, we can visualize how the sea behaves throughout the year and identify distinct patterns. Additionally, clustering techniques can help identify anomalies in the seasonal behavior of the sea, providing valuable insights. By examining the relationship between meteorological factors and marine physical parameters, our study contributes to a deeper understanding of how climate change influences the behavior of the sea. The findings can inform decision-making processes related to coastal management, marine safety, and conservation efforts. Understanding the changes in sea states and detecting anomalies are important steps toward adapting to and mitigating the impacts of climate change on marine environments. The Mediterranean Sea is widely recognized as one of the most vulnerable regions to climate change, primarily due to its exposure to climate events [Giorgi, 2006]. In fact, temperatures in the Mediterranean Sea have been observed to be 1.4°C higher than the global warming trend [Marini, 2018]. This study aims to contribute to the existing literature by modeling and deeply understanding the behavior of sea states, which can assist policymakers in addressing the increasing frequency of extreme events such as surges and storms.

Numerous models have been proposed to analyze marine physical parameters, with particular emphasis on waves and tidal currents as they represent significant hazards to coastlines. One widely used model for analyzing currents and wave behavior is the Simulating WAves Nearshore (SWAN) model. Introduced by Booij et al. [1999] as a third-generation wave model, SWAN describes waves using the density spectrum and computes the evolution of the action density. Over the years, this model has been extensively employed, such as Guillou [2017], who utilized SWAN as the foundation for studying the effects of tidal currents on waves by coupling their interactions. Marsooli and Lin [2018] also employed a SWAN model coupled with another model called ADCIRC to simulate storm tides and waves caused by tropical cyclones.

These models are computationally complex and require substantial amounts of historical data. Historical tidal data have been modeled to understand sea level trends. For instance, Foster and Brown [2015] analyzed sea level time series using Ordinary Least Squares (OLS) regression to estimate parameters and applied Auto Regression models to account for data auto-correlation, compute parameter variances, and estimate the number of degrees of freedom. Morucci et al. [2016] examined wave distributions using the Generalized Pareto Distribution, modeling various parameters including mean, variance, and shape, to analyze extreme events along the Italian Coast related to wave height.

While previous studies have analyzed waves and tides separately using rather standard statistical models, our research focuses on coupling the effects of tides and waves on sea levels using a bivariate model. To accomplish this, we employ Generalized Additive Models for Location, Scale, and Shape (GAMLSS) introduced by Rigby and Stasinopoulos [2005]. GAMLSS is a type of semi-parametric regression model designed to overcome the limitations of the Generalized Linear Model (GLM) and Generalized Additive Model (GAM) frameworks. While GAMLSS is commonly used for univariate regression analysis, where a single response variable is explained by multiple explanatory variables, our research utilizes a bivariate regression analysis to conduct a model-based clustering analysis incorporating latent variables. Lagona and Picone [2012] suggest that using a model-based clustering approach based on mixture models can help identify sea regimes, allowing the identification of distinct shapes that data assumes under latent environmental conditions.

By employing mixture models, clusters are identified as mixtures of multivariate distributions, providing insights into the locations and shapes of clusters. Other examples are given in Bulla et al. [2012], Maruotti and Alaimo Di Loro [2023], Huang and Dong [2019, 2020]; see McLachlan et al. [2019] for a comprehensive review on mixture models.

GAMLSS has gained widespread recognition in fields such as environmental and health data regression analysis due to its ability to accurately model non-stationary data using any parametric distribution. One key advantage of GAMLSS is its capability to model not only the mean of the distribution but also the shape and scale parameters, which can be represented as linear or smooth functions of the explanatory variables [Stasinopoulos and Rigby, 2007, Stasinopoulos et al., 2018].

In the literature, GAMLSS has found application in health research as well as the study of floods and wildlife. These models are particularly useful in analyzing data with high variability and bridging the gap between stationary and non-stationary analysis, enabling the study and comparison of both scenarios. For instance, Scala et al. [2022] conducted a regression analysis on river floods in the Sicily Region using GAMLSS, employing both stationary and non-stationary analyses by varying the covariates and comparing the results. They emphasized that natural events cannot be adequately explained by stationary distributions alone, as they fail to account for the natural variability of variables such as temperature, air pressure, precipitation, and floods that are affected by climate change. This highlights the utility of GAMLSS in modeling environmental data, which often exhibits non-stationarity. The study also demonstrated that using physical parameters (such as rainfall) as covariates to model the distribution's parameters yields more reliable results than solely relying on time as a covariate. Furthermore, GAMLSS has been employed to study the impacts of fires in Portugal [Sá et al., 2018] and to investigate the variability or changes in fish species due to biological or environmental factors [Colloca et al., 2017a, Costa et al., 2022].

The objective of our study was to explore the effects of sea physical parameters on two key indicators of sea state: meteorological tides and significant wave height modeling. To achieve this, we modeled both the mean and variance of these variables. We considered three primary factors as covariates that exert significant influence on our response variable: wind speed and direction, air pressure, and geographical fetch. Based on their seasonal behavior and impact on the sea parameters under investigation, we anticipated identifying four clusters that correspond to different seasons throughout the year. Specifically, we expected rougher sea conditions with potential storm surges during winter and autumn due to harsh weather conditions characterized by strong winds, rainfall, and low air pressure. In contrast, we anticipated calmer sea conditions during spring and summer, attributed to milder weather characterized by low wind speeds and high air pressure.

The paper is structured as follows: the first section is dedicated to the methodology and the use of GAMLSS and Finite Mixtures Models frameworks, then a description of the coastline of Ancona is provided along with marine data explanation, highlighting some features characterizing the coast taken into consideration; following analysis and results are presented; and finally, the last section presents the conclusion and suggests future developments of the study.

2.3 Methodology

The objective of the analysis is to detect and classify marine data features accounting for weather conditions; to do so the relationship between the meteorological tide level plus the significant wave height and the meteorological components such as wind speed, wind direction, and air pressure are analysed. To carry out the analysis *Generalized Additive Models for Location, Scale and Shape* (GAMLSS) have been used and they were implemented with *Finite Mixture Models* considering three different probability density distributions as candidate distributions for the data at hand: Gaussian, Log-Normal, and T-Family distribution. In addition, bi-variate analysis has been used to see in detail the different behaviors that the meteorological tides and the significant wave height assume according to different weather conditions. To develop the analysis the `R` software has been used.

2.3.1 Finite Mixture Models

Finite Mixture Models are used in cluster analysis especially when data are not homogeneous, as in this case, and more in general when analyzing environmental data. Finite mixture models can help overcome this problem by fitting more than one probability density function related to a single family, thus identifying more clusters with different distribution parameters. FMMs were first used in the 19th century by Newcomb to model outliers. Later the FMMs were used to analyze various data fitting different distributions such as the Gaussian distribution or the Poisson distribution. Nowadays, finite mixture models are widely spread in statistics because they can fit complex distributions of random phenomena. Indeed, they are applied in many fields such as "agriculture, astronomy, bioinformatics, biology, economics, engineering, genetics, imaging, marketing, medicine, neuroscience, psychiatry, and psychology, among many other fields in the biological, physical, and social sciences" [McLachlan et al., 2019]. One of the strengths of finite-mixture models is that they are very flexible. Nevertheless, their flexibility has some consequences as the number of components and the number of parameters increase. To apply the finite mixtures models there are some assumptions to take into consideration.

- Observations are considered independent, that is a particular value in the data set is statistically independent of the rest. This means that its value is not influenced by any other observations.
- The population is composed of a finite number of unobserved sub-populations (K).
- It is not possible to observe a priori to which an observation belongs.

In our analysis, we assume that observations are considered independent, meaning that each specific value in the dataset is statistically unaffected by any other observations. This assumption contributes to the simplicity and effectiveness of our model, even considering the derived approximation. Notably, a similar approach has been previously implemented by [Lagona and Picone, 2012], where groups of observed variables are assumed to be conditionally independent given a latent class. In particular, the response variable, y belongs to a population that is divided into K components. In one of these components, let's assume in the component k, the probability function of y is $f_k(y)$. In this case, the probability function of y is given by:

$$f(y) = \sum_{k=1}^{K} \pi_k f_k(y) \tag{2.1}$$

As a by-product of the estimation step (the Expectation-Maximization algorithm is often used in a maximum likelihood framework), posterior probabilities of belonging to a specific cluster are obtained, allowing to infer a proper clustering and measuring its uncertainty

2.3.2 GAMLSS

As previously stated, Generalized Additive Models for Location Scale and Shape were introduced by Stasinopoulos and Rigby [2007]. The model assumes different observations y_i for $i = 1,2,3,\ldots,n$ with probability function $f(y_i|\theta_{ik})$ where $\theta_{ik} = (\theta_{ik1}, \theta_{ik2}, \theta_{ik3}, \theta_{ik4}) = (\mu_{ik}, \sigma_{ik}, \nu_{ik}, \tau_{ik})$. The θ_{ik} is a vector of four distribution parameters respectively representing the location, scale, and shape in each component, K. In general, assuming a response variable y_i having length n and assuming $\rho = 1,2,3,4$ as the distribution parameters and let $g_\rho(.)$ be the link functions relating the parameters and the covariates $\mathbf{x}_{i\rho} = (1, x_{1i\rho}, \ldots, x_{P_\rho i\rho})$ to the response variable we can say that

$$\begin{aligned} g_\rho(\theta_{i,\rho k}) &= \eta_{i\rho k} \\ &= \mathbf{x}_{i\rho k} \times \beta_{\rho k} \end{aligned} \tag{2.2}$$

This model allows fitting each distribution parameter, in each mixture component, as a (weighted) linear function of the explanatory variables or as a linear function of the random effects. Although, there are sub-models of GAMLSS that allow fitting distributions also in a non-linear way.

To maximize the likelihood the **Expectation Maximization (EM)** algorithm is used. The latter allows the estimation of parameters in probabilistic models where there are missing data. Indeed, the application of the EM algorithm can be found in many scientific areas where direct access to data is difficult or where data are missing such as genetics, econometrics, and sociological studies [Moon, 1996]. Nonetheless, the EM algorithm can find its application also in other situations such as finite mixtures distributions models or latent variable models. In general, the EM algorithm attempts to estimate the parameters (theta) which maximize the log probability of data. In contrast to the maximum likelihood estimation the EM algorithm is more complex as it deals with missing data; indeed, in the incomplete data case the log probability function can have multiple local maxima. To overcome this problem the EM algorithm reduces the optimization problem into a sequence of sub-problems [Do and Batzoglou, 2008]. The EM algorithm consists of two steps: the Expectation step and the Maximization step. The former estimates the unknown underlying using the observed data set of the incomplete data: it uses the current parameters' estimates and the observed data to assess the complete log-likelihood function. The maximization step is fed with the expectation computed in the E-step and it tries to maximize providing a new estimate of the parameters. These two steps are iterated until convergence. The M-step follows the E-step trying to maximize the expectation computed in the previous step over θ [McLachlan et al., 2004]. The first application of the EM-type algorithm can be referred to Newcomb in 1886 [Newcomb, 1886] when analyzing mixtures of univariate normal distributions. The present formulation of the EM algorithm, instead, has been given by Dempster, Laird, and Rubin: they formulate the algorithm in a general setting and investigated its properties. After the EM algorithm has been widely applied in all statistics fields, also some extensions have been presented by McLachlan and Krishnan [2007].

GAMLSS are good for fitting data where the response variable does not follow a linear distribution or exponential distribution. Nonetheless, these models present some problems, for example, the variable selection becomes a more complicated issue in the GAMLSS framework as more parameters are taken into consideration. According to Rigby and Stasinopoulos, the Generalized Akaike Information Criteria need to be used for variable selection. Another issue is their applications to non-stationary data [Stasinopoulos et al., 2017].

2.4 Marine data description and modelling

The research focuses on the Gulf of Ancona, located on the east coast of Italy in the Marche region. The port of Ancona is situated on the Adriatic Sea, which is a semi-closed sea in the northern part of the central basin of the Mediterranean. This particular region of the Mediterranean Sea has garnered significant interest in numerous studies due to its size, being the largest arm of the Mediterranean Sea, and its connection to six countries.

In the northern part of the Adriatic Sea, a large number of rivers and their deltas contribute sediments that are transported along the coast by currents, adding to the dynamic nature of this sea The prevailing winds along the Adriatic coast are predominantly from the southeast and northeast, including the Scirocco, Mistral, Bora, and Tramontane winds. The exposure to these strong winds, coupled with the sea's depth variation from around 1200 m in the south to less than 30 m in the north, leads to mixed currents and unpredictable behavior [Franco and Michelato, 1992]. For example, the Gulf of Venice exhibits a depth ranging from 100 m to 200 m south of Ancona [Orlic et al., 1992]. Furthermore, some literature suggests that tidal currents in the Adriatic Sea are influenced by Ionian tides, as the two regions of the Mediterranean Sea converge in the Gulf of Otranto, south of Italy. Along the western side of the Adriatic Sea, the Western Adriatic Current (WAC) flows along the coastline from north to south. However, the northern part of the Adriatic Sea, including the shores of Ancona, experiences complex and variable tides. The currents exhibit a permanent counter-clockwise rotation, possibly reflecting a continuation of the behavior of Ionian currents and tides [Zonn and Kostianoy, 2017].

As expected, the Adriatic Sea is also affected by sea level rise due to the impact of climate change. However, in this case, the consequences primarily manifest in the alteration of the coastal landscape. As mentioned earlier, the western coast of the Adriatic Sea, particularly the Italian shoreline, is characterized by numerous rivers and their deltas, with the Po River being the most significant. Sea level rise leads to modifications in the morphology of these landscapes and the incursion of saltwater into inland areas. This contamination adversely affects ecosystems, soils, agriculture, and fisheries, causing significant damage [Carbognin et al., 2009].

In particular, the Gulf of Ancona exhibits an elbow-shaped configuration, exposed to the north, while being shielded by Monte Cernero to the south and the Apennines Mountains to the west. The prevailing wave regime in this area is predominantly influenced by winds from the northeast and southeast directions [APAT, 2004].

The data used for the Ancona site are obtained from both the tide gauge and the buoy, provided by the National Tide Gauge Network and the National Data Buoy Network, respectively. Ancona is among the few locations in Italy where both a tide gauge and a buoy are installed. The tide gauge station, situated in the seaport, measures atmospheric pressure, air and water temperature, and relative humidity every hour. It also records wind speed and wind direction every 10 minutes, as well as the water level. On the other hand, the wave buoy is located offshore and collects data at 30-minute intervals. It provides information on significant wave height, average wind direction, average wind speed, and other parameters that were not included in the analysis. The availability and reliability of these data are ensured by the Italian Institute for Environmental Protection and Research (ISPRA), which coordinates and manages these monitoring networks. Using this data, we can determine that during the analysed period (February 2021 February 2022), Ancona experienced a maximum significant wave height of 4.18 m. The significant wave height is defined as the average height, from trough to crest, of the highest one-third of the waves. It has been extensively studied in environmental research, particularly in the context of harnessing energy from the sea [Zheng and Li, 2015, Ali et al., 2020]. Using this data, we can determine that during the analyzed period (February 2021 – February 2022), Ancona experienced a maximum significant wave height of 4.18 m. The significant wave height is defined as the average height, from trough to crest, of the highest one-third of the waves. It has been extensively studied in environmental research, particularly in the context of harnessing energy from the sea [Zheng and Li, 2015, Ali et al., 2020]. Tides are also of interest, as their behavior is significantly influenced by strong winds and high waves. In the period under analysis, Ancona recorded a maximum meteorological residual of 114.78 m. The meteorological residual of tides represents the deviation between the actual tide level and the effect of the astronomical component on the sea level. In other words, it reflects the impact of meteorological parameters on the sea level. This residual is primarily influenced by atmospheric pressure, wind speed and direction, and wave height. Generally, Ancona exhibits higher tidal peaks during winter compared to summer, as depicted in Figure 2.1.

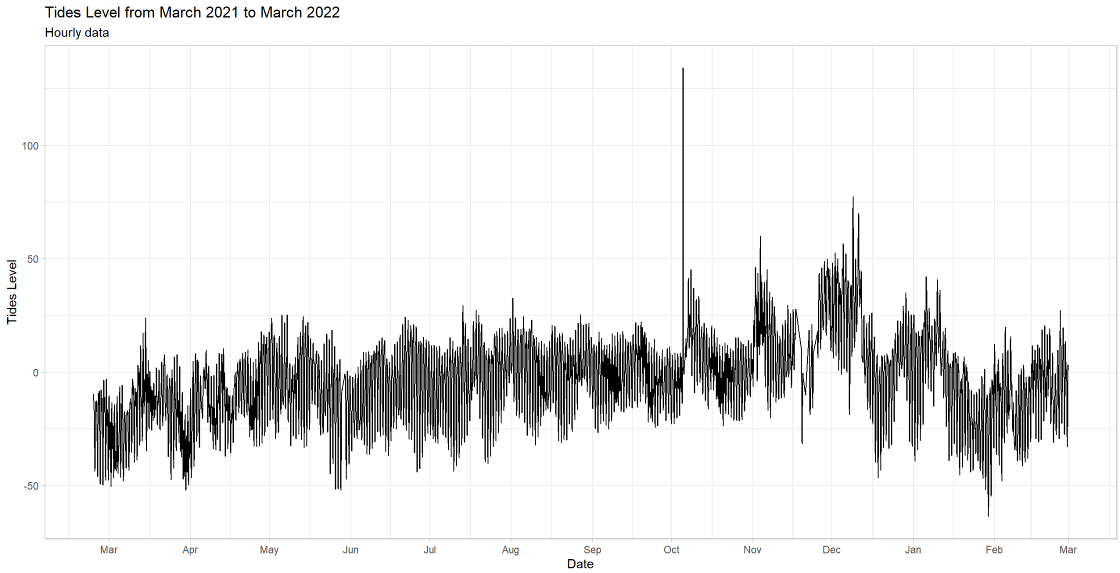


Figure 2.1: Representation of the tides level data during the period February 2021 - February 2022. Data are collected every 10 minutes, here they have been transformed into hourly data by computing the mean.

To see how much effect the meteorological component has on the sea level here are some graphs representing both the tide level and the meteorological residual. In general, we can see that there is always a difference between the two measures.

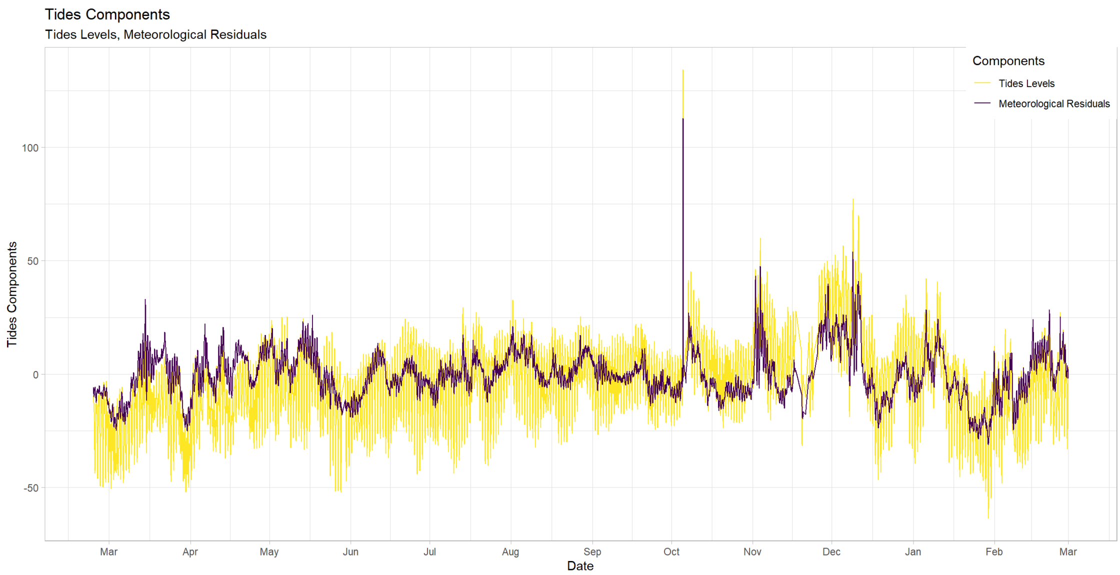


Figure 2.2: Decomposition between tides levels and meteorological residuals

As it is possible to observe in Figure 2.2 during winter the difference between the two components is small, this is because weather influences the levels of tides it mainly contributes to the formation of the tides. On the other hand, during summer the difference is more evident, this is because the formation of tides is caused mainly by other factors (astronomical components) while the weather is usually milder.

Data are collected using sophisticated technologies. In Italy, there is a network of detection tools spread all around the country. There are two different systems for collecting marine data, these are the *Rete Mareografica Nazionale* or *National Tidegauge Network (RMN)* and the *Rete Ondametrica Nazionale* or the *Italian Data Buoy Network (RON)*. The former is well spread all around Italy with 36 measurement stations which collect data every 10 minutes, while the latter has 15 buoys that measure and analyze data every 30 minutes [Canesso et al., 2012]. To carry out the research data were taken into consideration starting from the 23rd of February 2021 to the 28th of February 2022. Moreover, to simplify the calculation data have been aggregated by hour forming a data set of 8658 observations. The analysis was conducted using the `R` software. The data set includes several variables of interest, namely the meteorological residuals, significant wave height, air pressure, wind speed, and direction, and fetch. Hourly aggregation using the mean has been applied to the data, except for wind direction values, which are circular data representing the angle of the wind direction. To calculate the mean of wind direction, the `circular mean` function was utilized, which computes the mean based on the cosine and sine of the angle. Additionally, each wind direction value has been associated with its corresponding fetch. The fetch is divided into 72 categories, with each category representing a 5° angle class. It is important to note that wind direction data have been converted from degrees to radians, as direction is represented by an angle.

To calculate the meteorological residuals of the tide, the `TideHarmonics` package was employed. This package enables the harmonic analysis of tidal and sea level data, estimating over 400 harmonic tidal components. In the plot below (see Figure 2.3), it is possible to have a clear view of the three different components.

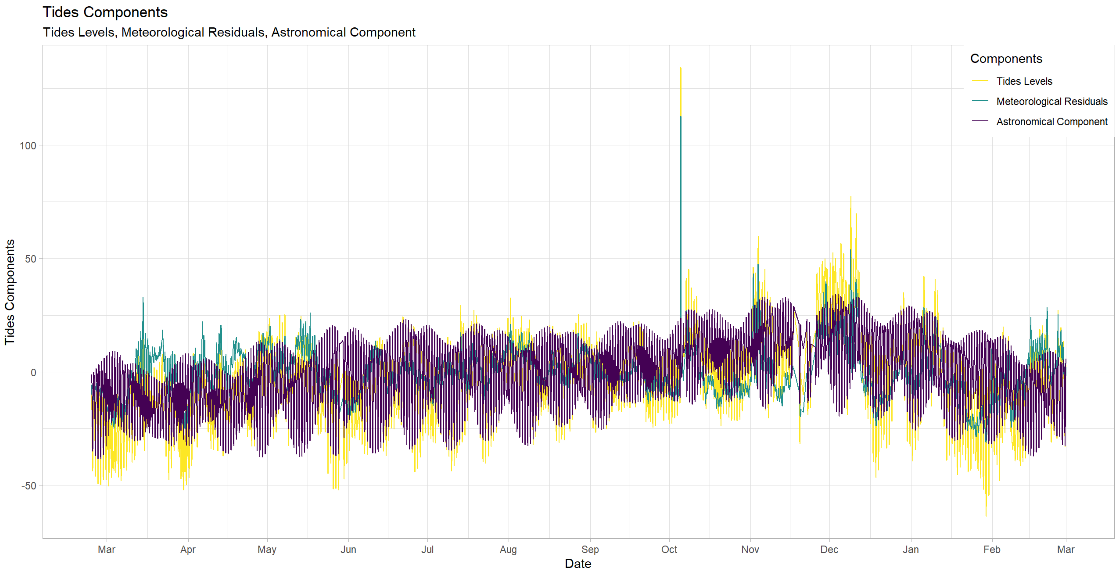


Figure 2.3: Components of the tides: the two different components concurring to the creation of the tides are represented together with the level of the tides. The astronomical component influences tides through the gravitational forces exerted by the Moon and the Sun. The meteorological component is given by the difference between the tides levels and the astronomical component. This shows the remaining factors that influence the formation of tides, that is the meteorological residuals.

The data visualization includes three lines: the yellow line represents the values collected by the station, the purple line represents the estimated astronomical values, and the green line represents the residuals, which are the components of air pressure and wind. It is evident from the observations that the values of the residuals are relatively high, indicating a significant difference between the observed values and the astronomical component. This implies that the meteorological component had a substantial impact on the observed tide. Based on this observation, it can be inferred that during the initial period of the analyzed year, the sea state was rough, likely due to consistently high wind speeds and/or low air pressure. Furthermore, the fetch distances have been scaled by a factor of 10,000 since they were originally measured in kilometers. It is important to note that the fetch considered in the analysis is the "effective fetch," which refers to the effective distance of open water. This choice may result in smoother outcomes. Moving forward, we will use the terms "pres" and "WS" to refer to air pressure and wind speed, respectively.

As aforementioned, the analysis is bi-variate, that is our response variable, is given by *Residuals* and the H_m0 which represents the significant wave height. To proceed with the bi-variate analysis a new code has been implemented to run our analysis for all the distributions we intended to fit: Gaussian, Log-Normal, and T-Family.The Gaussian and Log-Normal distributions involve parameters such as mean (μ_{ik}) and variance (σ_{ik}), while the T-Family distribution introduces additional parameters—mean (μ_{ik}), variance (σ_{ik}), and degrees of freedom (ν_{ik}). For the purpose of interpretation, only the first two parameters have been estimated in the case of the T-Family distribution. Moreover, two different models have been fitted to estimate both the mean and the variance, which account for the effective fetch in different ways. The first is the simpler one as it is given by:

$$\mu_{ik} = \beta_{0k} + \beta_{1k} pres_i + \beta_{2k} WS_i + \beta_{3k} fetch_i \tag{2.3}$$

$$log(\sigma_{ik}) = \delta_{0k} + \delta_{1k} pres_i + \delta_{2k} WS_i + \delta_{3k} fetch_i \tag{2.4}$$

In this case, we consider the fetch associated with each wind direction. The second model, on the other hand, is:

$$\tilde{\mu}_{ik} = \tilde{\beta}_{0k} + \tilde{\beta}_{1k} pres_i + \tilde{\beta}_{2k} weights_i \tag{2.5}$$

$$log(\tilde{\sigma}_{ik}) = \tilde{\delta}_{0k} + \tilde{\delta}_{1k} pres_i + \tilde{\delta}_{2k} weights_i \tag{2.6}$$

A new variable, *weights*, has been assessed to fit the model. It takes into account the wind speed and the fetch associated with the wind direction and it "weights" them to the maximum fetch that the location under investigation has.

From now on we will refer to the models as: **Model 1** and **Model 2**. The bi-variate analysis helps in analyzing *residuals* and *significant wave height* at the same time, but without overlapping their effects, that is we can look at their behavior separately according to changes in our explanatory variables. From 2 to 5 components (or cluster) K, were used. In the end, 4 models per distribution were run, for the two different analyses mentioned before, obtaining 24 different fitted models. To compare them three model selection criteria were used:

1. Akaike Information Criterion (AIC)

2. Bayesian Information Criterion (BIC)

3. Integrated Complete Likelihood (ICL)

The first criterion is assessed as:

$$AIC = -2 \times log\ell + 2K \tag{2.7}$$

In this case, the log-likelihood function ($log\ell$) is penalized by a term that depends on the number of parameters (K). The BIC follows the same pattern as the AIC but it penalizes the log-likelihood function also for the dimension of the sample size n, trying to solve the over-fitting problem.

$$BIC = -2 \times log\ell + K \times \log n \tag{2.8}$$

On the other hand, the ICL criterion is assessed as the BIC penalized by the estimated mean entropy.

$$ICL = BIC - entropy \tag{2.9}$$

The entropy is defined as

$$entropy = -\sum_{g=1}^{K}\sum_{i=1}^{n} p(z_i|y_i, \widehat{\theta_{ik}}, K) \log p(z_i|y_i, \widehat{\theta_{ik}}, K)$$

where z_i denotes the posterior probability of observing the response variable (y_i) in component K, and $\widehat{\theta_{ik}}$ represents the maximum likelihood estimation of the distribution's parameters. This measure characterizes the statistical uncertainty associated with mixture component K, reflecting its capability to create a meaningful data partition (see Biernacki et al. [2000], Celeux and Soromenho [1996] for more details). For all the criteria used the lower one shows the best model.

2.5 Results

Results for all three distributions were interesting, but looking at graphs highlighting the different clusters we conclude that the most interesting one is the Gaussian distribution as the components are better identified. According to the results obtained comparing the criteria for the AIC, the best model is the one with 3 components for both the first and the second analyses, while according to the BIC, the best one is with 3 components for the first model and the one with 4 components for the second model. However, according to the ICL, the best model is the one with 2 components for both first and second analyses fitted to the Gaussian distribution. This is because, as said before the ICL tends to penalize overlapping clusters. Nevertheless, the differences between the criteria are not so high. For this reason, based on the scatter-plot of the posterior probabilities we chose to look into three components results for the first model and into four components results for the second model.

As shown in Figure 2.4 in the first model the three clusters are well separated. For example, looking at the *significant wave height* and the *wind speed* clusters are well highlighted, and they seem to have a positive correlation, thus it is possible to understand that the behavior of the waves depends very much on the wind effects.

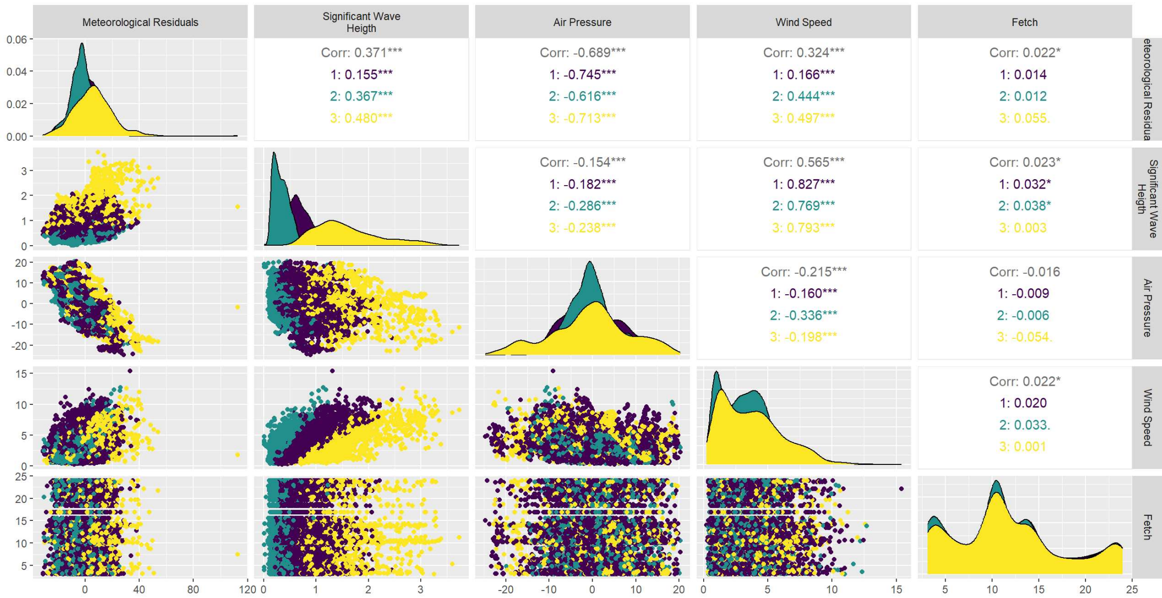


Figure 2.4: Clusters of the Gaussian distribution Model 1 with three components: highlighting of the three clusters from the fitting of Model 1. On the diagonal, it is possible to observe the distribution that each cluster assumes for each variable. On the right part of the matrix, the correlation for each cluster is shown with the level of significance.

The first cluster (purple) represents medium waves, while the second one (green) is capturing low waves, and the third one represents (yellow) high waves. Significant wave height is better modeled than the meteorological residuals of the tides as clusters for the latter are in some ways overlapped and difficult to identify. This may be because of the characteristics of the Adriatic Sea, as aforementioned, that exhibit a complex behavior regarding currents and tides.

Considering both the estimated value of the mean and the variance can provide a more comprehensive interpretation of the results. The estimated effects of the mean and the variance on the two response variables make clearer the impact that each meteorological component has on them.

Table 2.1: Estimated values of the mean and variance for the *Residuals* response variable.

	Estimated values of the mean				**Estimated values of the variance**			
	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)
Cluster 1								
Intercept	-1.38***	0.27	-5.21	0.00	1.70***	0.03	58.61	0.00
Air Pressure	-0.98***	0.01	-70.72	0.00	0.01***	0.00	4.81	0.00
Wind Speed	0.27***	0.05	5.00	0.00	0.05***	0.01	10.48	0.00
Fetch	0.01	0.02	0.57	0.57	0.00	0.00	0.28	0.78
Cluster 2								
Intercept	-6.20***	0.31	-19.85	0.00	1.74***	0.04	49.15	0.00
Air Pressure	-0.80***	0.02	-38.43	0.00	0.01**	0.00	3.22	0.00
Wind Speed	1.16***	0.06	18.05	0.00	0.02***	0.01	3.39	0.00
Fetch	-0.01	0.02	-0.27	0.79	0.00	0.00	0.56	0.58
Cluster 3								
Intercept	-4.18***	0.62	-6.76	0.00	2.46***	0.05	46.82	0.00
Air Pressure	-0.92***	0.03	-35.72	0.00	-0.02***	0.00	-7.30	0.00
Wind Speed	2.23***	0.09	24.46	0.00	-0.06***	0.01	-6.70	0.00
Fetch	0.04	0.03	1.22	0.22	-0.01***	0.00	-4.21	0.00

A detailed examination of Table 2.1 facilitates a more nuanced interpretation of the results. Regarding the residuals, it's worth noting that the estimated mean of the fetch holds no significance in the model, suggesting no influence on the results. This result was expected as the meteorological residuals of tides are not influenced by this variable, but on the other hand, they are affected by meteorological factors. Indeed, at first glance, we can see that the clusters show more or less the same consistent impact for the atmospheric pressure, while they significantly differ for the wind action. Hence, we can say that the three clusters are identified based on the diverse effects of wind speed. The first cluster (purple cloud on the graph) shows an intermediate situation for the meteorological residual of tides. The air pressure has a negative effect of -0.98 which mainly drives the state of the sea. From the matrix, looking at the Meteorological Residuals-Air Pressure plot we can see that the first component highlights points in the middle of the cloud of observations and a strong negative relation between the two variables is evident: the more the air pressure decreases the more the meteorological residuals increase. Also, a slight variance is possible to detect from the graph; indeed, from the analysis, it is significant with a positive relation, that is the variation in the air pressure values causes the increase of the residuals. Since the atmospheric pressure seems to decrease reaching lower values we can assume that this cluster may represent unstable weather, where the formation of clouds, strong wind, and rainfall is possible. What concern the wind, this cluster is the one less influenced by the action of the wind speed with a significant value of 0.27. Mainly, the meteorological residuals of tides are not influenced directly by the wind speed, but the latter has more influence on the creation of waves and currents which indirectly have impacts on the tides. In this case, it seems like the wind speed has a slightly positive effect on the residuals of tides identifying a situation where the wind is not so strong, but it varies a bit. Indeed, the variance of the wind speed is significant and shows a higher value with respect to the variance of the air pressure. Again, this may represent a situation of weather instability where the wind speed changes influence the level of meteorological tides. The second component shows a calm sea state highlighting low values of tides' residuals. Air pressure has always a negative significant impact on the level of the residuals and its variance is always significant in a positive way. This component differs from the previous one because it shows a significant impact of wind speed on the meteorological residuals. The effect is positive with a value of 1.16, nevertheless, the tides' levels are lower. As said before, the wind speed does not have a direct impact on the level of the tides but it influences directly the waves and currents. In this case, instead, the effect of the wind speed on the tides is strongly positive. Though, the level of residuals is low suggesting an interesting dynamic. There are different possible explanations for this observation: it may be that the positive effect of the wind speed on the tides' residuals is mainly offset by the action of the air pressure which tends to lower the sea level, indeed if we look at the graph we notice that the second cluster (green) captures high values of atmospheric pressure which means the force applied on the sea surface is strong and decrease the sea level. Another explanation could be that may be a lag effect, that is there could be a time delay or lag between changes in wind speed and their impact on the residuals not captured by the analysis. This could lead to a situation where the effect of wind speed is positive but not yet fully realized in the current residuals. The third cluster (yellow) represents the rough sea, where the residuals and the waves are high. The effect of the air pressure is still negative and high (-0.92) as for the previous clusters, but in this case, the effect of the wind speed is significant with a very high value of the estimated mean (2.23). As shown in the graph values of the air pressure tend to decrease rapidly and at the same time the wind speed is increasing with a huge effect on the residuals, hence the level of the tides is very high due to the combination of the effects of these two components. This cluster is quite interesting as, along with the variance of the air pressure and wind speed, also the estimated variance of the fetch is significant. Interestingly, the estimated variances show negative effects on the meteorological residuals, contrasting with the previous clusters. This indicates that the variability or fluctuations in these meteorological components contribute to the rough sea conditions and high tides observed in this cluster. That is, the negative estimated values of variance suggest a reduction in the variability of the respective meteorological components. In this case, it implies that there is less variability or fluctuation in air pressure, wind speed, and fetch within the rough sea conditions. Also, the negative values of variance could indicate a more persistent or consistent behavior of the meteorological variables associated with rough sea conditions. It suggests that the rough sea conditions are characterized by a relatively stable or unchanging pattern of air pressure, wind speed, and fetch, resulting in lower variability. In a nutshell, we can say that weather conditions characterizing this situation are stable and optimal to create the rough sea and the persistent or consistent state of these variables contributes to the intensity of rough seas and higher tidal levels.

Table 2.2 shows the estimated effects of the mean and the variance on the second response variable, the significant wave height. In this case, the three clusters completely differ as expected as the generation of waves is driven mainly by the wind speed and not by the atmospheric pressure. The first cluster (purple) represents medium waves. The estimated mean values for this cluster are all statistically significant. The primary factor driving the formation of these medium waves is wind speed, which has a positive effect (0.12). This means that as the wind speed increases, the height of the waves also tends to increase. On the other hand, the air pressure has a negative effect on wave formation, as expected, although its influence is not particularly strong. In other words, when air pressure is high, it tends to

Table 2.2: Estimated values of the mean and variance for the *Significant wave height* response variable

	Estimated values of the mean				**Estimated values of the variance**			
	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)
Cluster 1								
Intercept	0.26***	0.01	34.38	0.00	-2.12***	0.03	-70.38	0.00
Air Pressure	-0.00***	0.00	-7.43	0.00	0.00	0.00	1.25	0.21
Wind Speed	0.12***	0.00	63.67	0.00	0.16***	0.01	28.42	0.00
Fetch	0.00*	0.00	2.39	0.02	0.00	0.00	0.45	0.65
Cluster 2								
Intercept	0.14***	0.00	31.41	0.00	-3.01***	0.04	-81.31	0.00
Air Pressure	-0.00	0.00	-0.02	0.98	0.02***	0.00	8.87	0.00
Wind Speed	0.07***	0.00	53.48	0.00	0.26***	0.01	34.68	0.00
Fetch	-0.00	0.00	-0.88	0.38	-0.00	0.00	-0.62	0.53
Cluster 3								
Intercept	0.65***	0.03	19.97	0.00	-0.95***	0.05	-17.79	0.00
Air Pressure	-0.00**	0.00	-3.06	0.00	-0.01**	0.00	-3.22	0.00
Wind Speed	0.22***	0.01	36.59	0.00	0.04***	0.01	4.08	0.00
Fetch	-0.00	0.00	-0.30	0.76	0.00	0.00	0.37	0.71

suppress wave formation to some extent. Another significant factor is the fetch, which represents the distance over which the wind blows across the water's surface. Despite not having a high impact, it still contributes positively to the formation of medium waves. A larger fetch generally results in higher wave heights. What is particularly interesting is the estimation of variance, which measures the variability within each factor. Among the variables considered, only the variance of wind speed shows a significant impact on wave characteristics. A positive variance suggests that when the wind speed varies more, it leads to an increase in wave height. In this case, the wind speed exhibits substantial variability, ranging from calm conditions (0 m/s) to moderately strong winds (slightly above 10 m/s). This observation is supported by the graph, which shows a clear positive correlation between wave height and wind speed. The wider range of wind speed values indicates that when there is greater variability in wind speed, it tends to generate medium-sized waves with higher significant wave height. In summary, the formation of medium waves in the purple cluster is primarily influenced by wind speed. Higher wind speeds contribute to increased wave height, while air pressure has a dampening effect. The fetch, although not a major factor, still contributes positively to wave formation. Notably, the variability in wind speed plays a crucial role, as higher variability leads to larger waves. The second cluster (green) represents low waves, indicating smooth sea conditions. In this cluster, the estimated mean value of the wind speed is the only component that influences the sea state. It has a slight positive effect (0.07) on the significant wave height. Although the influence of the mean wind speed estimate is not substantial, it still contributes to the formation of waves in this cluster. However, what is particularly interesting in this cluster is the estimated variance of both air pressure and wind speed. Both variables have a positive effect on wave generation. The estimated variance of air pressure does not have a significant impact on wave formation (0.02), but it still contributes to a small extent. On the other hand, the estimated variance of wind speed has a greater impact (0.26) on wave generation. Despite the positive effects of the estimated variances, the waves observed in this component are not particularly high. This suggests that although the variability in air pressure and wind speed contributes to the generation of waves, it does not result in significantly elevated wave heights within this low wave cluster. The third and final cluster (yellow) represents high waves, indicating rough sea conditions. This cluster is particularly intriguing as it captures specific weather and sea conditions. In this case, both the estimated mean values for air pressure and wind speed are found to be statistically significant. For air pressure, the estimated mean value shows a negative effect on wave height, although it is close to zero. This implies that air pressure still has a negative impact on wave formation within this cluster. While the effect is not particularly strong, it suggests that higher air pressure tends to suppress wave height to some degree. On the other hand, the estimated mean value for wind speed in this cluster is the highest among the three clusters (0.22). This indicates a consistent and significant positive impact of wind speed on wave formation. In other words, as the wind speed increases, the height of the waves also increases. The presence of high waves and rough sea conditions in this cluster can be attributed to the higher wind speeds observed. Additionally, both the estimated variances for air pressure and wind speed in this cluster are found to be significant. For air pressure, the variance shows a slight negative impact on wave height. This suggests that greater variability in air pressure has a minor dampening effect on wave formation within this cluster. In contrast, the variance of wind speed has a positive impact on wave height, and the effect is slightly higher. This implies that increased variability in wind speed contributes to higher wave heights within this cluster. The combination of higher mean wind speed and greater variability in wind speed leads to the formation of high waves and rough sea conditions. These findings suggest that the specific weather conditions characterized by higher wind speeds and their variability contribute to the formation of high waves and rough sea conditions within this cluster. To determine if the three clusters exhibit patterns typical of specific months or seasons of the year, we can examine the following graphs.

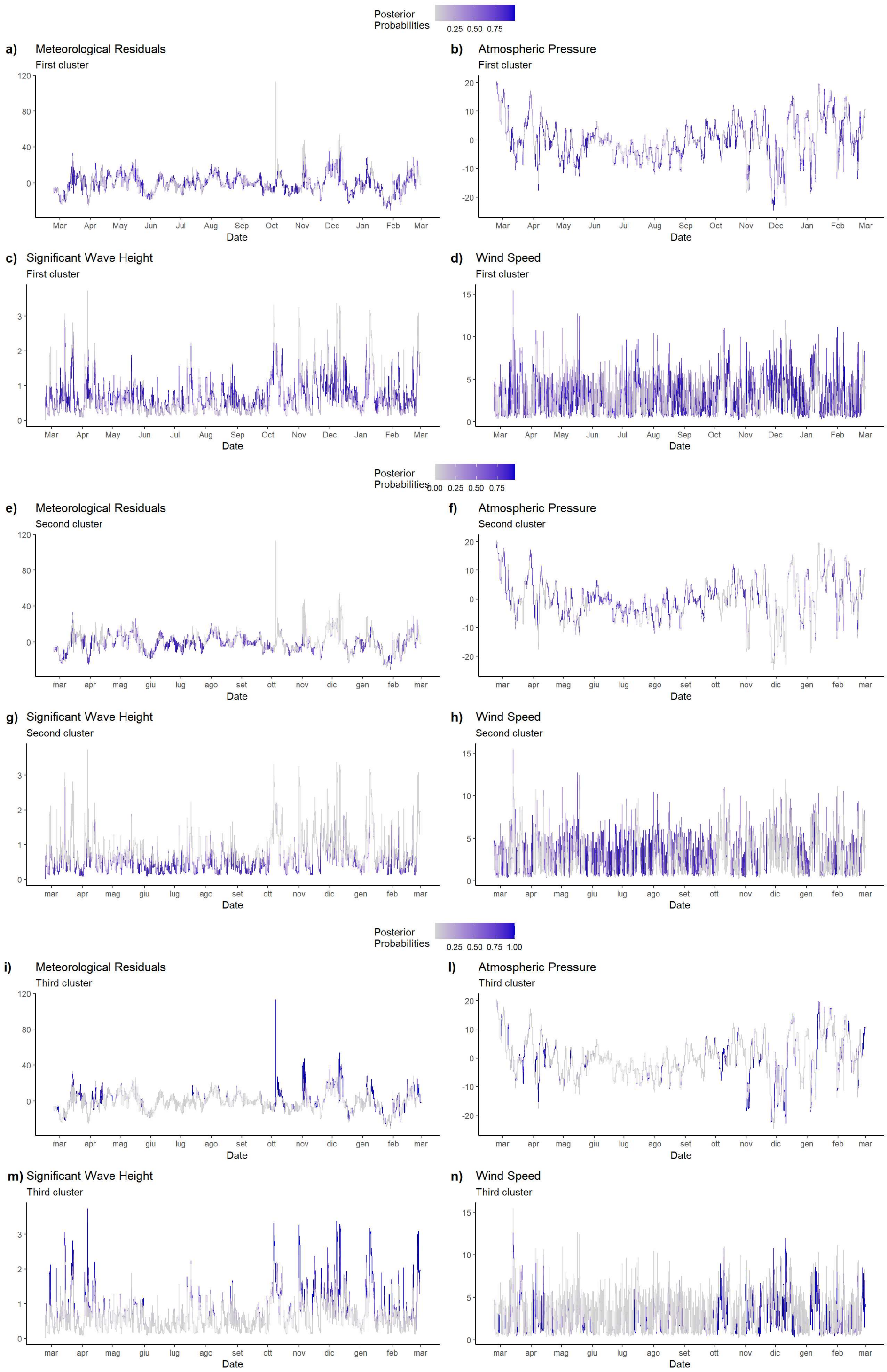


Figure 2.5: Sea state in the three clusters: these graphs represent the components that contribute to the formation of the sea state. They highlight each variable's values that fall in each cluster. Here it is possible to appreciate the different behaviors of the components and how they influence the sea level during the period of observation.

The pattern exhibited by each cluster throughout the year is depicted in Figure 2.5. The first cluster, as previously explained, comprises moderate residuals and waves. Indeed, both components highlight intermediate values which are observed during the entire year of observation. For this reason, it is not possible to identify a special seasonal pattern that characterizes this kind of sea state, on the other hand, we can say that this behavior is quite common during the year. The atmospheric pressure highlighted in this cluster is quite variable showing both decreasing and increasing paths. As mentioned before, this component is the main driver for the residuals' levels. At the same time, the wind speed shows quite variable and high values which contribute in part to the formation of the sea level, mainly the wave generation. The second cluster represents low residuals and waves, indeed the atmospheric pressure and the wind speed are not so high. Observations for this kind of sea state are concentrated in the spring-summer period. The third cluster highlights the highest values for both the residuals and wave heights. These values are associated with the lower values of the atmospheric pressure and the higher values of wind speed, describing well the behavior of the sea under specific weather conditions. What is more interesting is the period captured for this kind of phenomenon, which is the autumn-winter season, when we expect the see to be rough due to adverse weather conditions. As said before, the second analysis has been considered with four components. Recall that in this analysis the wind speed and the fetch have been incorporated into a new index: **weights**.

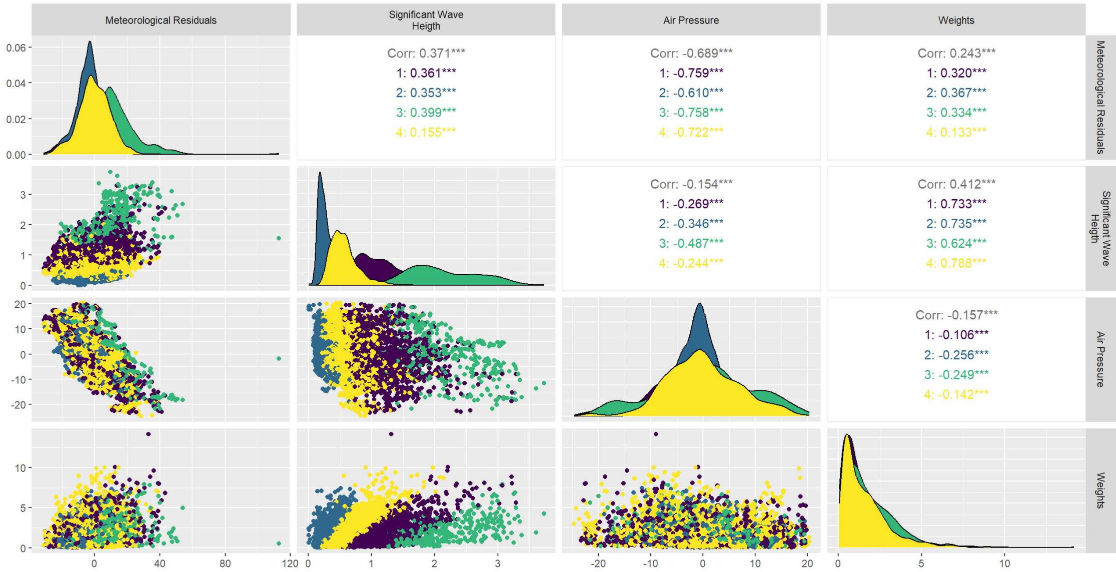


Figure 2.6: Clusters of the Gaussian distribution Model 2 with four components: highlighting of the four clusters from the fitting of Model 2. The diagonal of the matrix shows the distribution of each cluster for each variable. On the right side, it is shown the correlation coefficient for each cluster with its level of significance.

Looking at Figure 2.6 the four clusters can be easily identified for the *Significant Wave Height*. Still, the residuals' clusters are not well separated. In this case, the interpretation of the sea states identified is not straightforward, even if we can state that the second (blue) represents low waves, and the third (green) high waves, the first (purple) and fourth (yellow) components represent medium waves but in different weather conditions.

Table 2.3: Estimated values of the mean and variance for the *Residuals* response variable.

	Estimated values of the mean				**Estimated values of the variance**			
	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)
Cluster 1								
Intercept	-2.30***	0.23	-9.98	0.00	1.95***	0.02	81.97	0.00
Air Pressure	-1.00***	0.02	-54.76	0.00	0.00	0.00	0.45	0.65
Weights	1.83***	0.11	16.53	0.00	0.01	0.01	0.81	0.42
Cluster 2								
Intercept	-5.54***	0.20	-27.31	0.00	1.74***	0.02	72.60	0.00
Air Pressure	-0.82***	0.03	-30.60	0.00	0.01**	0.00	2.93	0.00
Weights	1.24***	0.10	11.95	0.00	0.03**	0.01	3.02	0.00
Cluster 3								
Intercept	4.82***	0.73	6.59	0.00	2.56***	0.05	46.57	0.00
Air Pressure	-0.92***	0.05	-19.35	0.00	-0.02***	0.00	-6.26	0.00
Weights	2.45***	0.21	11.58	0.00	-0.22***	0.02	-9.25	0.00
Cluster 4								
Intercept	-0.93***	0.16	-5.89	0.00	1.79***	0.02	105.69	0.00
Air Pressure	-0.97***	0.02	-64.03	0.00	0.00*	0.00	2.46	0.01
Weights	0.23**	0.08	2.78	0.01	0.05***	0.01	6.88	0.00

In Table 2.3, the mean values estimated for the residuals are the same for all four clusters: both the air pressure and the weights are significant respectively with negative and positive impacts. Although, the atmospheric pressure always has the same effect on the response variable while the *Weights* seems to drive the classification.

The first cluster (purple) represents a range of medium to high residuals. Within this cluster, the air pressure has a significant negative impact (-1.00) on the sea level. This is evident from the graph, which shows a pronounced decrease in air pressure leading to an elevation of the residuals for tides. In other words, lower air pressure is associated with higher sea levels, resulting in larger residuals. Simultaneously, the *weights* variable has a substantial positive effect (1.83) on the response variable. The *weights* variable contributes significantly to the increase in the sea level. It represents the combined influence of various factors, such as wind speed and fetch, on the response variable. Higher values of the *weights* variable correspond to larger sea levels, leading to increased residuals. Furthermore, the variance of the two covariates (air pressure and *weights*) is not significant within this cluster. This indicates that there is not a substantial amount of variability or difference among the values of air pressure and *weights* within this particular range of residuals.

The second cluster (blue) represents a situation where the sea is calm, resulting in low residuals. Within this cluster, atmospheric pressure has a significant negative impact on the residuals (-0.82). The high values of the residuals indicate that high pressure is exerted on the sea surface. In other words, when atmospheric pressure is high, it tends to suppress wave activity, leading to calmer sea conditions and smaller residuals. On the other hand, the weights variable has a high positive effect (1.24) within this cluster. This suggests that the weights play a significant role in influencing the meteorological residuals of the tides. Interestingly, the values captured for the weights within this cluster are not as low as expected. This observation can be attributed to the variance of the variable. Although the estimated value of the weights (0.03) is not particularly high, it still has a noticeable positive impact on the residuals. This suggests that even in a situation of a calm sea, there may be subtle variations in the weights. The presence of steady wind blowing from a specific direction may explain the relatively higher values of the weights within this cluster. This consistent wind flow helps maintain the calmness of the sea surface, resulting in lower residuals for the tides.

The third cluster (green) represents a situation of storms or surges, characterized by the highest values of tides. Within this cluster, the atmospheric pressure consistently has a significant negative estimated value (-0.92), indicating its strong impact on the tides. Although the values of atmospheric pressure vary greatly, ranging from negative to positive, the negative estimated value suggests that it plays a crucial role in the state of the sea and the resulting high values of the residuals. The high waves associated with storms or surges likely contribute to this state. Additionally, the component of the weights in this cluster has a large and significant estimated value (2.45). This indicates that the weights variable has a substantial influence on the tides' residuals in this situation. The weights accounted for in this cluster show some variability, neither too low nor too high. This suggests that factors, such as wind speed and fetch, contribute to the high tides observed during storms or surges. The significant variance of the weights within this cluster further emphasizes their importance, as their variation has a notable impact on the resulting tides' residuals.It is highly plausible that this cluster represents a situation where weather conditions were unstable and adverse, giving rise to storms or surges.

The fourth and final cluster (yellow) represents a situation characterized by low to medium residuals. Within this cluster, both the air pressure and weights variables have estimated mean values that are negative and positive respectively. This suggests that the effects of these variables on the residuals are not consistently in the same direction. In this scenario, the influence of the weights variable is slightly smaller, as indicated by its estimated mean value of 0.23. This suggests that in this particular situation, the strength of the wind is not as significant and does not have a substantial influence on the state of the sea. It implies that the wind is relatively calm and does not contribute significantly to the observed tides or wave heights, resulting in lower residuals. The variances of both the air pressure and weights variables within this cluster are significant and positive. This indicates that changes in these weather components contribute to the variability observed in the residuals. In other words, variations in air pressure and weights, even if they are not consistently in the same direction, still play a role in determining the final tides or wave heights, and thus affect the residuals
As for the previous analysis, the effects of the two meteorological variables are clearer when looking at the significant wave height. The latter is very well divided according to the different sea states identified.
For every cluster, the estimated coefficient of the mean resulted statistically significant for both explanatory variables: negative for the air pressure and positive for the weights. Though the values estimated for the atmospheric pressure are not so high and do not differ from each other, this suggests the classification is led by changes in the *Weights* variable.

The first cluster (purple) represents an intermediate situation characterized by medium and high waves. In this scenario, the air pressure has a relatively minor influence on the sea state. The estimated values for both the mean and variance are close to -0.01, indicating that changes in air pressure have minimal impact on wave generation within this cluster. The variance is significant negatively in this case, it may suggest that the atmospheric pressure was slightly decreasing (-0.11), usually the atmospheric pressure decrease when there is a change in weather conditions, often signaling worsening weather. On the other hand, the weights have a slightly greater positive effect for both the estimated mean and the estimated variance

Table 2.4: Estimated values of the mean and variance for the *Significant Wave Heights* response variable

	Estimated values of the mean				**Estimated values of the variance**			
	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)
Cluster 1								
Intercept	0.69***	0.01	66.63	0.00	-1.31***	0.02	-57.71	0.00
Air Pressure	-0.01***	0.00	-9.24	0.00	-0.01***	0.00	-3.86	0.00
Weights	0.20***	0.01	33.06	0.00	0.11***	0.01	10.28	0.00
Cluster 2								
Intercept	0.18***	0.00	67.79	0.00	-2.89***	0.03	-114.10	0.00
Air Pressure	-0.00***	0.00	-7.34	0.00	0.01*	0.00	2.50	0.01
Weights	0.07***	0.00	33.32	0.00	0.29***	0.01	23.51	0.00
Cluster 3								
Intercept	1.61***	0.04	39.90	0.00	-0.68***	0.06	-11.31	0.00
Air Pressure	-0.02***	0.00	-8.07	0.00	-0.02***	0.00	-5.82	0.00
Weights	0.22***	0.02	14.27	0.00	-0.07**	0.03	-2.63	0.01
Cluster 4								
Intercept	0.36***	0.00	92.02	0.00	-2.10***	0.02	-120.05	0.00
Air Pressure	-0.01***	0.00	-12.82	0.00	-0.00	0.00	-1.10	0.27
Weights	0.12***	0.00	46.90	0.00	0.19***	0.01	22.26	0.00

(0.20 and 0.11). Indeed, from the graph, we can see that there is a positive relation between the waves and the weights variable indicating that the higher the value of Weights the higher the waves. At the same time, the variance of the explanatory variable influences the sea state as its variability may amplify its effect on wave formation.

The second component (blue) represents low waves. In this scenario, the impact of air pressure is negatively significant but close to zero, indicating that air pressure has little influence on wave height. However, the variance of air pressure is positively significant (0.01), suggesting that its variability plays a role in wave formation. From the plot, we can observe that the highlighted waves are very small, coinciding with a slight decrease in air pressure within this cluster. The effect of wind on wave height is positive but relatively small (0.07), and its variance is also positively significant (0.29). This implies that wind contributes slightly to wave generation in this cluster, and the variability in wind conditions has a notable impact on wave formation. Notably, the variance of the *Weights* variable is considerably higher in this cluster compared to the other three clusters. This indicates relatively stable weather conditions, as the high variability in the *Weights* variable suggests consistency or a lack of significant changes in the associated factors influencing wave generation. These results, then, may represent a situation of a smooth sea, where the air pressure is low and stable, while the wind is blowing not so strong, but it may be changing its direction or strength as the graph shows an increase in the *Weights* values which lead to little higher waves, this explains the significance of the variance of the weights.

The third group (green) represents rough sea conditions with high waves. In this scenario, atmospheric pressure has a small but significant negative impact on significant wave height (-0.02), indicating that higher atmospheric pressure is associated with slightly lower wave heights. The variance of atmospheric pressure remains negatively significant (-0.02), suggesting that its variability plays a role in shaping wave conditions. The values of atmospheric pressure in this group are relatively high and appear to be decreasing, indicating the significance of its variability in influencing wave generation. On the other hand, wind has a significant positive impact on wave height (2.45). The plot reveals that even a small increase in the *Weights* values leads to considerably higher waves, indicating that the wind in this scenario is strong and blowing in the appropriate direction to generate high waves. This suggests that the fetch, which refers to the distance over which the wind blows, is likely to be substantial. Although the variance of the wind variable is negatively significant (-0.22), indicating some variability in wind conditions, the wave heights remain high. This implies that the variability in the *Weights* variable, as shown in the graph, is not sufficient to counterbalance the strong effect of wind speed and direction on wave generation.

The last cluster (yellow) represents sea conditions characterized by low to medium waves. In this case, only the estimated mean of atmospheric pressure is negatively significant (-0.01), while the estimated variance is not. The plot indicates a slight negative relationship between atmospheric pressure and wave height, suggesting that lower pressure is associated with higher waves. It appears that as the atmospheric pressure decreases, the wave height tends to increase. On the other hand, the *Weights* variable has both the estimated mean and variance positively significant (0.12 and 0.19). The estimated mean indicates a positive effect, implying that higher wind components correspond to higher wave heights. Additionally, the positively significant variance suggests that the variability in weather conditions contributes to an increase in wave height. From the plot, we can observe that the *Weights* variable exhibits considerable volatility in this cluster, resulting in higher waves. This component likely describes changes in the sea state, transitioning from low to medium waves due to fluctuations in wind speed and direction.

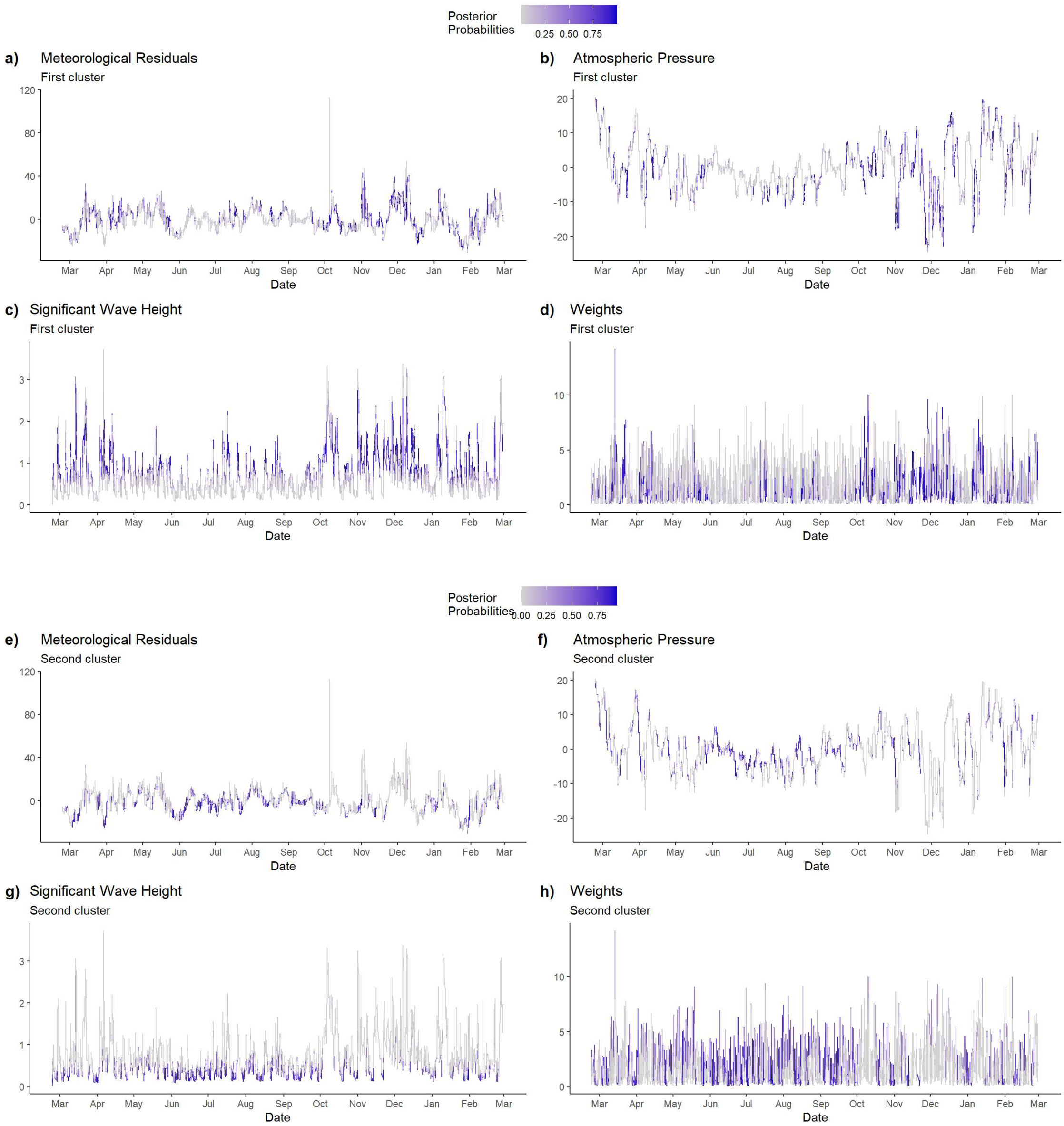


Figure 2.7: Sea state in clusters 1 and 2: these graphs represent the components that contribute to the formation of the sea state, highlighting variable values in each cluster.

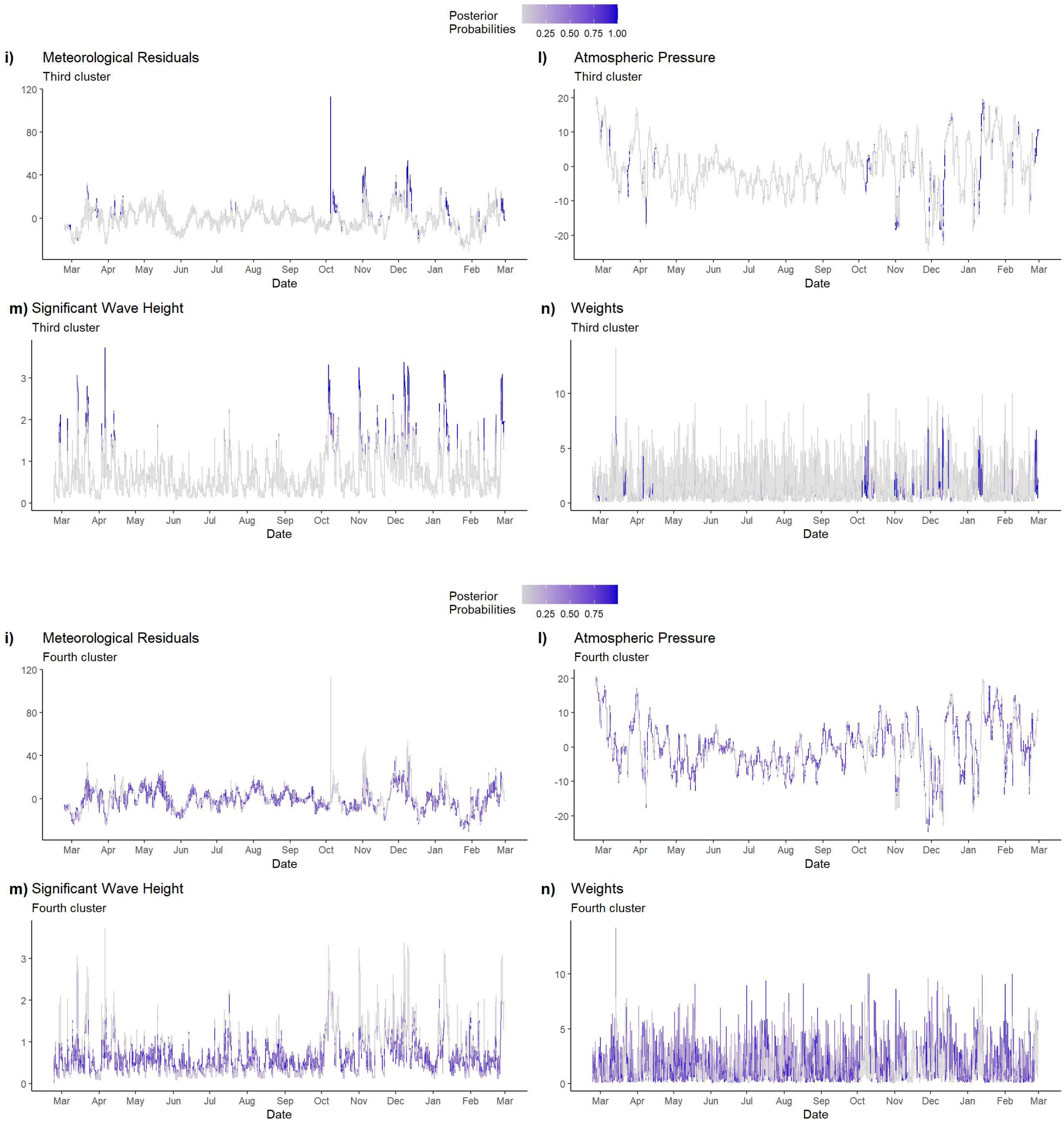


Figure 2.8: Sea state in clusters 3 and 4: these graphs represent the components that contribute to the formation of the sea state, highlighting variable values in each cluster.

Based on the previous findings, Figures 2.7 and 2.8 provide a comprehensive view of the patterns exhibited by each cluster over the course of the year. The first cluster, as previously described, represents an agitated sea state. This is evident from the significant wave height and residuals, which consistently demonstrate high values within this cluster. Moreover, the atmospheric pressure consistently registers lower values, while the *Weights* components exhibit moderate values. It is noteworthy that this agitated sea state scenario persists throughout most of the year, with the exception of June, July, and September. This indicates that for the majority of the year, the Adriatic Sea maintains an intermediate sea state. This implies that the sea conditions are generally moderate, apart from specific periods when the sea state becomes more agitated. The second cluster represents a distinct scenario characterized by a very calm sea state. In this cluster, both meteorological residuals and wave heights exhibit consistently low values. Additionally, the air pressure tends to show higher values, while the weights variable remains relatively constant. This particular scenario is predominantly observed during specific months, namely May, June, July, and August. It aligns with our expectations, as these months correspond to the summer season when weather conditions generally promote a calm sea state. The third cluster represents a distinct set of weather conditions, as evidenced by the relatively limited number of observations captured within it. Within this cluster, both meteorological residuals and wave heights display remarkably high values. Meanwhile, the atmospheric pressure exhibits considerably low values, while the Weights variable does not appear to have particularly high values. This finding may seem contradictory to our theoretical understanding of sea behavior. However, upon closer examination of the plot for the Weights variable, we can observe that the highlighted values often occur immediately after or before peaks in its values are registered. This suggests a time lag in the analysis, indicating that changes in wind speed and their subsequent impact on the residuals and wave heights might not be fully captured by the current analysis. This time lag in the analysis implies that the effects of changes in wind speed on the sea state may take some time to manifest in the observed residuals and wave heights. As a result, the analysis may not accurately capture the immediate influence of wind speed on the sea conditions. The fourth and final cluster depicts a sea state characterized by low to medium sea levels. In this cluster, the atmospheric pressure tends to assume high values, resulting in increased pressure exerted on the sea surface. As a consequence, the sea level is lower. However, despite the elevated atmospheric pressure, the sea state in this cluster is not calm. This can be attributed to the noteworthy variability and relatively high values of the weights factors. These factors indicate the significant influence of wind components on the sea state. Consequently, even with high atmospheric pressure, the actions of the wind play a substantial role in maintaining a non-calm sea state. Interestingly, this particular sea state pattern is observed consistently throughout the year, suggesting its typical occurrence in the Gulf of Ancona. Regardless of the season, the fourth cluster consistently represents the prevailing sea conditions in the region.

2.6 Conclusions

In conclusion, the research findings have provided valuable insights that align with our theoretical understanding of sea state behavior, particularly in the context of the Adriatic Sea. By employing the GAMLSS approach and thoroughly analyzing the various meteorological components, we have successfully identified and associated their effects on the sea surface. This segmentation of effects allows for the classification of sea level rise conditions, distinguishing between normal conditions and those conducive to surges. Considering the model selection criteria, the three- and two-component models demonstrated the best performance, with slight differences observed among the results. However, the three-component model utilized in the initial analysis has been selected as the most suitable choice due to its ability to effectively capture and represent the different states of the sea. This can be observed in the cluster graphs where the highlighted clusters correspond well with the various sea conditions. Furthermore, when incorporating the variable *weights*, which encompasses important meteorological factors such as wind speed, wind direction, and fetch, the four-cluster classification emerged as the most appropriate. This result highlights the interconnected nature of these meteorological components and their significant influence on sea levels, particularly within the Adriatic Sea. Given the unpredictable nature of tides in this region, the wind components play a major role in shaping the sea state and influencing wave formation. The inclusion of the "weights" variable allows the model to accurately identify and classify distinct sea states based on different weather conditions, with particular emphasis on wind speed and direction.

It is important to acknowledge that while wind speed and direction are influential factors, the applicability of these findings may vary in other geographical locations. In different regions, other factors, such as the meteorological residuals of tides, may also play a crucial role in determining sea levels. Consequently, the influence of air pressure becomes more significant in defining the sea state in such cases. This analysis serves as a valuable foundation for future research endeavours. For instance, further exploration could involve the inclusion of additional data from multiple years, enabling the development of a robust model for sea state forecasting. Additionally, this analysis can be expanded to investigate the effects of climate change on sea levels, offering important insights into the potential impacts of changing climatic conditions on coastal areas.

In conclusion, the research has shed light on the complex relationship between meteorological components and sea state behaviour. The findings provide a solid basis for further investigations and hold potential for applications in sea level forecasting and the study of climate change impacts on coastal regions.

Chapter 3

A zero-inflated hidden semi-Markov model with covariate-dependent sojourn parameters for analysing marine data in the Venice lagoon[1]

[1]Ricciotti, L., Picone, M., Pollice, A., & Maruotti, A. (2025). A zero-inflated hidden semi-Markov model with covariate-dependent sojourn parameters for analysing marine data in the Venice lagoon. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 74(2), 506–529.

3.1 Paper Summary

In this third empirical chapter of the thesis, based on Ricciotti et al. [2025], the focus shifts from modelling continuous sea-state variables to modelling marine count data—specifically, the occurrence of extreme tide events (*acqua alta*) in the Venice Lagoon and the activation of the MOdulo Sperimentale Elettromeccanico (MOSE), defence system from flooding for the Lagoon. The study presents a novel statistical modelling framework built to capture the temporal dynamics, zero-inflation, and regime-dependence of such events, thereby furthering the thesis's aim of applying advanced statistical learning methods to coastal and marine systems. Flooding events in the Venice Lagoon represent a major coastal hazard of both environmental and socio-economic concern. For this reason, the MOSE system has been built positioned at the entrances of the Lagoon. Accurate modelling the occurrence and assessing whether the activation of the MOSE is of help preventing flooding aids flood risk assessment, early-warning systems and infrastructure planning and management.

The proposed modelling strategy is based on a hidden semi-Markov model (HSMM) with zero-inflated Poisson distribution allowing the sojourn (duration) distribution's parameters to depend on covariates. The core model includes: a latent semi-Markov process evolving in time across a finite number of hidden states, representing latent risk regimes for flooding; the count of flooding events conditional on the latent state is modelled via a zero-inflated Poisson distribution; the sojourn of each latent state is allowed to assume different distributions and it is parametrised such that its parameters depend on covariates; the transition probabilities between states and the emission parameters also may depend on covariates. This model structure allows a rich description of the dynamics of flooding events in the lagoon: not only how many events and when, but under which latent state, for how long, and how environmental drivers influence state entry, exit and emission behaviour. The empirical case study uses data from the Venice Lagoon. The response variable is the counts of exceedances of predefined flooding limits (i.e., events when tide/sea level exceeds a threshold). Covariates include relevant environmental drivers that may influence both the emission process and the sojourn dynamics of the latent risk states.

This chapter advances to the hazard-modelling layer, connecting sea and meteorological conditions to actual flooding events and proposing a latent-state model that can be used for forecasting and risk assessment. Moreover, the emphasis on covariate-dependent state durations foreshadows the modelling of temporal dynamics and state-dependent behaviour that will later appear in the spatio-temporal analysis carried out in Chapter 4 and in the economic-policy setting in Chapters 5, 6, 7. The analytic capacity to characterise latent regimes, durations, and their drivers underpins subsequent modelling of fisheries variability and coastal defence investment dynamics. Beyond the immediate case study of the Venice Lagoon, this work contributes to environmental statistics and coastal risk modelling in several ways. It shows how hidden semi-Markov frameworks can be adapted for zero-inflated count data—a common feature in environmental hazard data (e.g., rare event counts). The explicit modelling of state-duration (sojourn) dependence on covariates is a methodological advancement, enabling richer temporal dynamics than memoryless (geometric) Markov models. From a coastal management perspective, the ability to identify and characterise latent risk states, their durations, and driving conditions opens the door to regime-specific monitoring, forecasting, and intervention strategies.

3.2 Introduction

The repercussions of climate change are intensifying, marked by the escalating frequency, severity, and unpredictability of extreme events. Projections indicate an inevitable rise in sea levels by the century's end, amplifying the vulnerability of coastal and lagoonal regions. Among these, the historical city of Venice and its lagoon in the Northern Adriatic Sea stand as poignant examples, constantly exposed to flooding events caused by marine tides, known as *acqua alta* events. This necessitates a pragmatic assessment of environmental conditions under which these events are likely to happen [Alberti et al., 2023].
Over the last century, human interventions have significantly altered the lagoon's landscape. Notably, the construction of the MOSE (Experimental Electromechanical Module) has substantially reshaped the lagoon inlets [Toso et al., 2019]. Its increasingly frequent operation promises further transformations, posing profound impacts on societal and ecological fronts. Thus, it is rather important to evaluate its efficacy in limiting *acqua alta* events as its repeated use may also lead to changes into the lagoon ecosystem.

We aim at contributing to the literature on flooding events in the Venice lagoon by modelling the time series of tides exceeding 80 cm from January 1st 2020 to June 12th 2024. We choose to motivate tide levels modelling from the perspective of peaks over threshold. The emphasis is thus on modelling the daily counts of threshold exceedances, which we favour as this approach allows for the identification of high tides within a manageable framework: by focusing on exceedance counts, we can effectively capture the frequency of high-tide events while avoiding potential biases introduced by selecting only the extreme tide values (for a discussion, see e.g. Jonathan and Ewans [2013]).

This task is usually important in domains where extreme values (i.e., the tail of the distribution) are highly relevant, such as earthquakes, and hurricanes too.

Four main features characterize our data:

- A prevalent occurrence of zero counts.

- Dependence of count process parameters on observed environmental covariates

- Time-varying heterogeneity, indicating data drawn from diverse environmental conditions needing proper modeling.
- The influential role of the MOSE in attenuating *acqua alta* events.

Count data exhibiting excessive zeros are common (see e.g. Feng [2021] for a recent review), often stretching the limits of standard count distributions like the Poisson model. To address this, we initiate our analysis with a zero-inflated Poisson distribution. Utilizing a regression framework, we seamlessly link distribution parameters to observed environmental variables, employing canonical links akin to generalized linear models [Turner et al., 1998].

To capture data dynamics influenced by diverse environmental conditions and time-dependent factors, we introduce a Hidden semi-Markov model (HSMM; Yu [2010], Bulla et al. [2010], Langrock and Zucchini [2011], Hadj-Amar et al. [2023]), allowing for flexible sojourn times in the hidden Markov states. This model approximates the data distribution through a mixture of densities, contingent on the evolution of a hidden semi-Markov chain. This approach effectively controls the data dynamics. The general idea of modelling time series by finite mixtures with time-varying parameters is not new. Often, the latent process driving the parameters is assumed as a Markov chain, and the hidden Markov model (HMM) is considered in both the count data literature (Adam et al. [2019], Chen et al. [2019], Ötting et al. [2021], Berentsen et al. [2022]), even when zero-inflation arises [Wang and Alba, 2006, Alfò and Maruotti, 2010, DeSantis and Bandyopadhyay, 2011] and marine data analysis [Bulla et al., 2012, Lagona et al., 2015]. HMMs are flexible, well-established models useful in a diverse range of applications [Bartolucci et al., 2012, Zucchini et al., 2017, Bartolucci and Farcomeni, 2019]. However, one potential limitation of such models lies in their inability to explicitly structure the sojourn times of each hidden state. HSMMs are more useful in the latter respect as they incorporate additional temporal structure by explicit modelling of the sojourn times. However, even though the transition between HMMs and HSMMs is mathematically straightforward, the complexity of the model increases considerably. Conceptually, one needs to consider all possible state sequences at the same time as all possible sojourn times for each state. This renders HSMMs computationally intensive and as a result, the literature on HSMM applications is considerably smaller than that relating to HMMs. Nevertheless, recently, cutting-edge methods have been developed within the HSMMs framework [Economou et al., 2014, Pohle et al., 2022], and with a focus on environmental data analysis [Merlo et al., 2022, Maruotti et al., 2021, Stoner and Economou, 2020]. These proposals however rely on homogeneous sojourn time distributions. While homogeneity can be a realistic assumption in selected applications, it becomes a major drawback when a study aims at estimating the influence of time-varying covariates on the sojourn time distribution. We achieve this goal by modelling sojourns by state-specific Logarithmic regressions where time-varying covariates tune the sojourns and a transition to a different regime. Our empirical contribution lies in deploying non-homogeneous HSMMs to link the environmental covariates to the count data process representing the number of *acqua alta* events, and to model the MOSE system's impact explicitly. This model meticulously considers the system's influence on the number of *acqua alta* events, sojourn times and state transitions, crucial in understanding and quantifying its effects on mitigating high-level tides.

Despite its complexity, parameters of HSMMs are generally estimated using the method of maximum likelihood. The likelihood equations have a highly nonlinear structure; however, there exists a convenient explicit expression for the log-likelihood that can be easily evaluated even for very long sequences of observations. This makes it possible to estimate the parameters by direct numerical maximization of the log-likelihood function [MacDonald, 2014, MacDonald and Bhamani, 2018].

The main results related to the application of HSMMs to the Venice lagoon data can be summarized as follow. We clearly identify three different environmental conditions, related to the observed covariates, corresponding to the steady-state, moderate risk-state, and a more risky state in the lagoon, in terms of likelihood of *acqua alta* events. Environmental covariates have time-varying effects, i.e. their effects are state-specific, with atmospheric pressure and wind speed (and partially wind direction) playing a crucial role. The MOSE has an impact on the different ingredients of the modelling and its efficacy is discussed.

The paper is structured as follows. In Section 3.3 we introduce the empirical problem, describing the main characteristics of the Venice lagoon and the data collected *ad hoc* for this research, focusing on the evolution of environmental conditions in the lagoon. Section 3.4 introduces the novel HSMM, along with some computational aspects. The main results are discussed in Section 3.5 and a discussion with some still open questions and further developments to be considered in the next future can be found in Section 3.6.

3.3 The Venice Lagoon and the MOSE

The Venice Lagoon is a unique and iconic geographical feature located in the northeastern part of Italy. It is a shallow coastal area that stretches along the Adriatic Sea and encompasses the city of Venice and its surrounding islands. The lagoon is highly susceptible to flooding, due to its low-lying topography, the gradual subsidence of the land, and the periodic high tides from the Adriatic Sea. This leads to frequent *acqua alta* events, occurring when sea level at the station of Punta della Salute reached or exceeded 80 centimeters. The phenomenon of *acqua alta* has captured significant attention in the literature, due to its substantial impact on the city's economy and its escalating frequency. Historical data show that very high tide events doubled in 2002-2021 compared to 1966-2001 (refer to `https://www.comune.venezia.it/it/content/durata-media-delle-maree2` for additional details), thus increasing the risk of *acqua alta*. Various factors contribute to these events. For instance, Lionello et al. [2021] conducted a detailed analysis, emphasizing that typical floods in Venice result from a combination of low air pressure and strong winds such as, Sirocco (South-East) and Bora (North-East). Additionally, Pham et al. [2023] examined the Venice Lagoon under a climate change scenario, employing different models, including a waves model, whose results emphasize the influence of Sirocco and Bora winds on waves in the region. Again, according to Zanchettin et al. [2021] the most influencing wind direction over the Lagoon is the Sirocco wind, which leads to the accumulation of Ionian surface waters towards the northern Adriatic.

Thus, it is of fundamental importance to include wind characteristics (i.e. speed and direction) into any modelling specification. Indeed, from a graphical inspection of wind characteristics during *acqua alta* events (see the wind rose in Figure 3.1), it becomes evident that the Bora wind prevails with higher wind speeds, whilst the Sirocco wind could also play a role, though being less common than the Bora.

To address *acqua alta*, Venice has implemented various engineering solutions, the latest being the MOSE system—a set of 78 movable barriers at the three lagoon

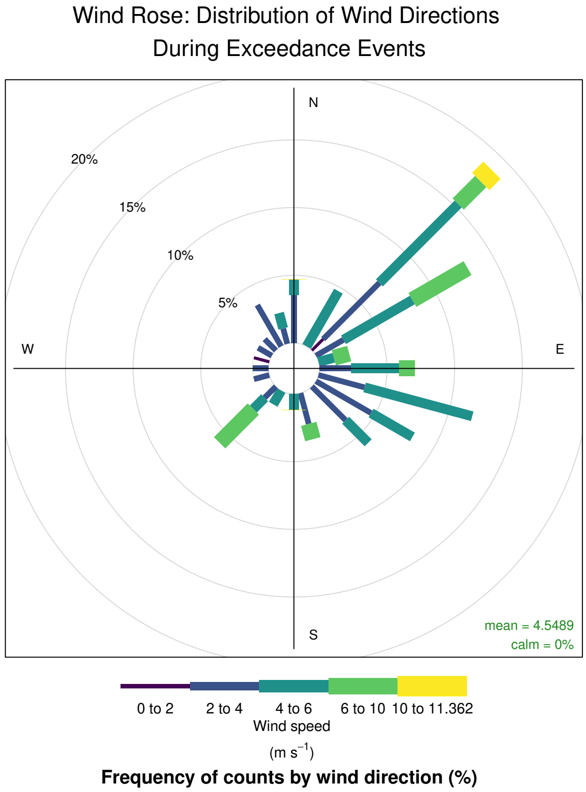


Figure 3.1: Wind direction distribution during *acqua alta* events, with wind speeds catagories.

entrances (see Figure 3.2). The MOSE system was devised to mitigate the recurrent *acqua alta* events in Venice, notorious for ravaging the city's cultural and economic heritage over time [Zanchettin et al., 2021].

A technical committee bases its decision to activate the MOSE on tide predictions, weather conditions, and sea level measurements. Barriers are deployed to keep the lagoon below the safety level when water level predictions reach a specified threshold and when these exceedances persist for an extended duration throughout the day or experience a substantial increase over the course of the day. Extreme events, involving the flooding of the city's streets and buildings, have intensified in frequency and severity in recent decades [Umgiesser et al., 2021, Lionello et al., 2021, Alberti et al., 2023]. Consequently, the research community has devoted efforts to assess coastal erosion processes, their impact on seawater quality, and the surrounding ecosystems [Pham et al., 2023].

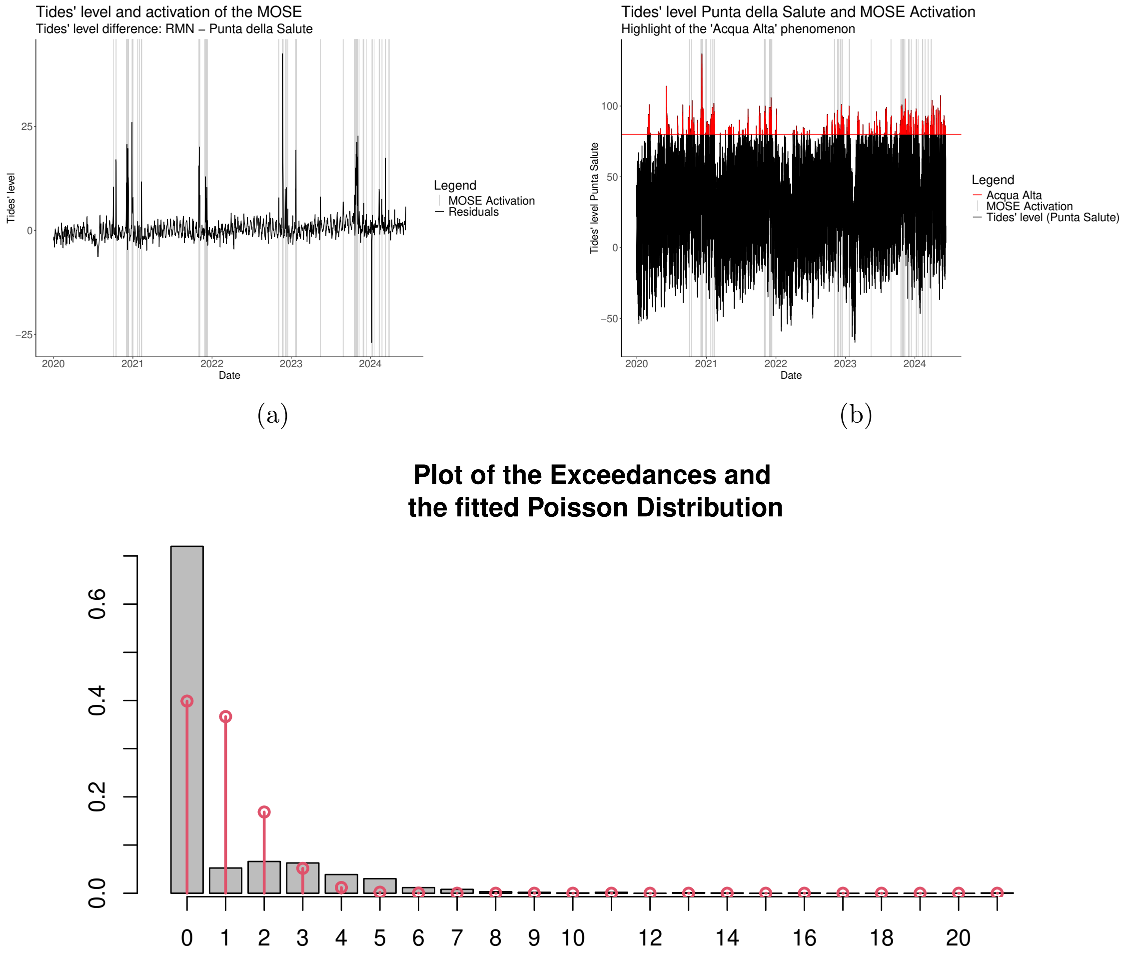
Figure 3.2: Map of the Venice Lagoon, with locations of the MOSE's three barriers securing the lagoon inlets.

3.3.1 Marine data

Meteorological and marine data were gathered from two distinct sources: the Italian Institute for Environmental Protection and Research (ISPRA) and the Municipality of Venice.

The dataset spans from January 1st 2020 to June 12th 2024. This particular time-frame holds significance in our research, as it coincides with the first use of MOSE in October 2020. Meteorological data are sourced from the ISPRA's tide gauge, strategically positioned just beyond the boundaries of the lagoon. This placement ensures that meteorological conditions remain authentic, unaffected by the morphological features of the city. The variables considered include wind speed and direction, relative humidity, atmospheric pressure, and water temperature. Environmental measurements for the considered variables are available at different time scales. Dealing with mixed-data sampling, known as MIDAS [Guérin and Marcellino, 2013], has not been discussed in the HSMM literature yet, and requires additional modelling complexity. Here, instead, to keep the analysis simple, though still reliable, we collapse and aggregate the data at the day level, resulting in 1625 time observations. Overall, missingness is not a relevant issue, as approximately the 1.5% of observations of the entire dataset is not recorded. However, whilst dealing with missingness in the outcome variable is not a cumbersome task [Maruotti, 2015], the same is not true when covariates values are missing (see e.g. Lagona and Zhang [2010]). Thus, we perform a multiple imputation strategy to obtain complete data for both the response and covariates. This is obtained by using the *Amelia* package (Honaker et al. [2011]), jointly with the *TideHarmonics* package [Stephenson, 2016] in `R`. For each variable, at the corresponding time scale, we performed 100 imputations. The number of imputations was fixed after controlling for the variability in the means of the imputed missing data. The `ftide` function of the *TideHarmonics* package was further considered to account for the cycle of the tide levels. Whilst the inclusion of linear variables (i.e. wind speed, temperature, humidity and atmospheric pressure) is straightforward, dealing with circular variables (i.e. wind direction) in a regression setting requires further attention [Pewsey et al., 2013]. To incorporate wind direction into the model projections of the angles on the South-North axis (cosine of the angle, where conventionally North is taken as 0 degree and angles increase clockwise) and West-East axis (sine of the angle) have been considered in the modeling process. In addition, a binary variable indicating whether the MOSE has been activated on that specific day or not has been added. At last, dummy variables for the years have been added, using 2020 as the benchmark. This approach allowed us to account for the varying strategy of activation of the MOSE system across different years. It is important to notice that the MOSE's functionality is still under testing, thus decisions on its use evolved during the years based on testing results. The change in the management of the MOSE was primarily due to the full operability of the MOSE system from 2021 and the increasing frequency of *acqua alta* events.

Figure 3.3a illustrates the difference among tidal levels, captured by gauges both inside the lagoon and outside the lagoon (depicted by the solid black line). The gray lines highlight instances of MOSE activation. The plot shows that MOSE activation coincides with exceptionally high tidal levels outside the lagoon, mostly in autumn and winter. During MOSE activation, the ISPRA tide gauge often records higher levels outside the lagoon than inside (Punta della Salute). This disparity, prompting MOSE activation, demonstrates the system's protective function by shielding the lagoon from rising tides. Figure 3.3b illustrates the *acqua alta* phenomenon when



(a) (b)

(c)

Figure 3.3: Tides characteristics. Figure 3.3a shows the difference between the tidal levels recorded by the RMN outside the lagoon and those captured inside the lagoon (Punta della Salute); light gray vertical lines represent MOSE activations. Figure 3.3b shows tidal levels at Punta della Salute; the red line indicates *acqua alta* and light gray vertical lines represent MOSE activations. Figure 3.3c shows the marginal distribution of *acqua alta* events and the fitted Poisson distribution in red.

tide levels exceed 80 cm. Finally, Figure 3.3c presents a visual representation of the fitted Poisson distribution for our count data. distribution of the counts. It is noteworthy that the marginal distribution of zero counts is much lower than the observed count, indicating that the Poisson distribution would not be able to account for the observed excess of zero counts. Consequently, it becomes evident that the use of the Zero-Inflated Poisson distribution, is necessary.

As discussed in Section 3.3, a key factor contributing to high water is the prevailing weather condition, primarily encompassing atmospheric pressure and wind patterns. Atmospheric pressure plays a significant role in influencing sea levels, interacting with the astronomical component. To dissect the impact of each, tidal levels can be decomposed into the astronomical component, and the meteorological residuals (or residuals only), the latter being the component influenced only by meteorological factors.

Variable	**U.m.**	**Mean**	**Standard deviation**	**Min**	**Max**
Wind Speed	m/s	4.17	3.10	1.02	47.11
Wind Direction	Radiant	1.17	1.47	0.00	6.28
Water Temperature	Celsius	16.91	6.59	6.73	28.97
Relative Humidity	%	79.04	10.50	42.87	100.00
Atmospheric Pressure	hPa	1016.2	7.89	988.4	1041.2

Table 3.1: Descriptive statistics of daily measurements of the considered environmental variables.

Figure 3.4 presents wind roses divided by distinct classes of residuals, categorized based on the influence of air pressure. These categories are delineated based on the corresponding influence of air pressure. The visual representation reveals a noteworthy pattern: minimal residuals coincide with high atmospheric pressure levels, suggesting a suppressing effect on tidal levels. In contrast, elevated residuals are associated with low air pressure, and are predominantly influenced by winds coming from the northeast direction. This underscores the significant impact of wind patterns on the variability of residual levels, providing valuable insights into the intricate interplay between atmospheric conditions and tidal dynamics.

3.4 Statistical model

3.4.1 The Zero-Inflated Hidden semi-Markov model

We assume that the observed data are of the form $\{Y_t, \mathbf{X}_t\}_{t=1}^{T}$ where $\mathbf{Y} = \{Y_1, \ldots, Y_T\}$ are non-negative integers, denoting the number of hours of *acqua alta* events at day t; and $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$ is a $Q \times T$ matrix, with Q representing the number of available covariates, collecting atmospheric data. The observed data are driven by a

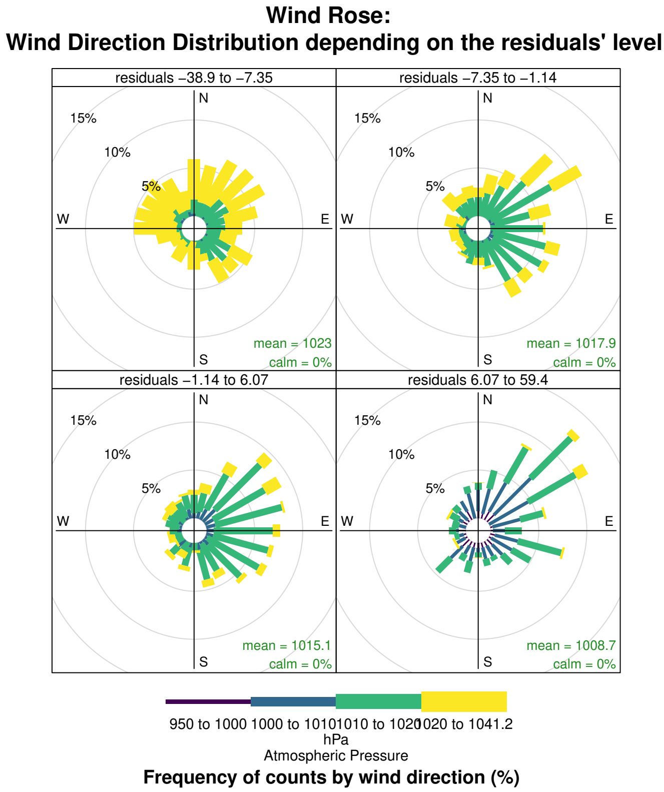


Figure 3.4: Wind rose diagram illustrating the distribution of wind directions categorized by different residual levels, each associated with the corresponding atmospheric pressure effects.

finite-state semi-Markov chain, defined on the state space $\Omega = \{1,2,\ldots,K\}, k \in \Omega$, at time t. A hidden semi-Markov model is a particular kind of time-dependent mixture. The underlying hidden process $\{S_t; t = 1,\ldots,T\}$ is constructed as follows. A homogeneous Markov chain with K states models the transitions between different states, with initial probabilities $\pi_k = \Pr(S_1 = k)$ and transition probabilities

$$\pi_{t,k|j} = \Pr(S_t = k \mid S_{t-1} = j, S_t \neq j), \quad t = 2,\ldots,T \tag{3.1}$$

with $\sum_{k=1}^{K} \pi_{t,k|j} = 1$ and $\pi_{t,k|k} = 0$, i.e. the diagonal elements of the transition probability matrix are zeros. We collect the initial probabilities in the K-dimensional vector $\boldsymbol{\pi}$, whereas the transition probabilities are collected in the $K \times K$ transition matrix $\boldsymbol{\Pi}_t$.

We explicitly model the sojourn-time distribution, associated with each hidden state as

$$d_k(u) = \Pr(S_{t+u+1} \neq k, S_{t+u-v} = k, v = 0, \ldots, u-2 \mid S_{t+1} = k, S_t \neq k), \qquad (3.2)$$

i.e. the sojourn of the hidden process of length u from $t+1$ until $t+u$ in state k. Before and after this sojourn in state k, the process has to be in some different state. The upper bound of the time spent in state k is defined by D_k. We will assume that the state sojourn distributions are concentrated on the finite set of time points $\{1, \ldots, D_k\}$, where D_k may also increase up to the entire length of the observed time series. Thus, rather than having $d_k(u)$ implicitly defined by the elements along the diagonal of $\mathbf{\Pi}$, as for HMMs, we model $d_k(u)$ explicitly. In other words, the difference with HMMs is that the probabilities of self-transitions $\pi_{k|k}$, $k = 1, \ldots, K$, which determine the time spent in each state, are now separately modeled by the sojourn distribution $d_k(u)$. The combination of a Markov chain, modeling state changes, and sojourn-time distributions defines the hidden process and illustrates the main difference between the HMM (in which the sojourn-time distribution is implicitly geometric) and the HSMM. The semi-Markovian hidden process $\{S_t\}$ of a HSMM does not have the Markov property at each time t, but is Markovian at the times of state changes only.

Let the number of different states visited consecutively by the hidden semi-Markov chain be $R+1$. Thus, the number of sojourn times, denoted by $u_1, \ldots, u_R$ fulfills $u_1 + \cdots + u_R = T$. The relationship between the sojourn times $u_1, \ldots, u_R$, with $R \leq T$, and the state sequence $s_1, \ldots, s_T$, can be written in a simplified way by reducing that sequence to the sequence of states $\widetilde{s}_1, \ldots, \widetilde{s}_R$ which have been visited:

$$\begin{aligned}
\widetilde{s}_1 &:= \{s_1, \ldots, s_{u_1}\}, \\
\widetilde{s}_2 &:= \{s_{u_1+1}, \ldots, s_{u_1+u_2}\}, \\
&\vdots \qquad\quad \vdots \\
\widetilde{s}_R &:= \left\{s_{u_1+\cdots+u_{R-1}+1}, \ldots, s_T\right\}.
\end{aligned}$$

Before and after each sojourn, the process has to be in a different state. The Markov property is so transferred to $\widetilde{s}_1, \ldots, \widetilde{s}_R$ which constitutes a hidden Markov chain.

Let $f_{t,k} = \Pr(Y_t = y_t \mid S_t = k; \mathbf{X}_t = \mathbf{x}_t)$ be the conditional distribution of observed counts. Because the data show zero-inflation, we set $f_{t,1}$ to be the probability mass function of a Zero-Inflated Poisson distribution, allowing for a mix of *sampling zeros* (those generated from the underlying Poisson process) and *structural zeros* (the remaining zeros that are not originating from any process). Specifically, let p_t be the structural zero proportion and $\lambda_{t,k}$ be the mean of the Poisson process at time t in the latent state k; for $k = 1$, we have

$$f_{t,1} = p_t(\mathbf{x})\mathbf{1}_{y_t=0} + \{1 - p_t(\mathbf{x})\}\frac{\lambda_{t,1}^{y_t}(\mathbf{x})\exp\{\lambda_{t,1}(\mathbf{x})\}}{y_t!} \tag{3.3}$$

and for $k = 2, \ldots, K$ we have

$$f_{t,k} = \frac{\lambda_{t,k}^{y_t}(\mathbf{x})\exp\{\lambda_{t,k}(\mathbf{x})\}}{y_t!}. \tag{3.4}$$

For each time t, and hidden state k, we model p_t and $\lambda_{t,k}$ as functions of the available covariates $\mathbf{x}_t$,

$$\log\left(\frac{p_t(\mathbf{x})}{1 - p_t(\mathbf{x})}\right) = \mathbf{x}_t^{'*}\boldsymbol{\delta} \tag{3.5}$$

and

$$\log(\lambda_{t,k}(\mathbf{x})) = \mathbf{x}_t^{'*}\boldsymbol{\beta}_k \tag{3.6}$$

with $\boldsymbol{\delta}$ and $\boldsymbol{\beta}_k$ being a vector of (state-specific) regression coefficients and $\mathbf{x}_t^* = (1, \mathbf{x}_t)$. We use the Logarithmic distribution for the sojourn distribution in equation (3.2) with probability mass function

$$d_k\,(u; \mu_k) = \frac{\alpha_k \mu_k^u}{u}, \quad u = 1{,}2, \ldots, \tag{3.7}$$

where $0 < \mu_k < 1$ $\alpha_k = -[\log(1 - \mu_k)]^{-1}$. We also attempted using different sojourn distributions. Specifically, we ran all models using the shifted-Poisson and non-parametric distributions for the sojourn time, but the results for the fitted models were not satisfactory in terms of the model selection procedures described in the next section.

3.4.2 Dealing with endogeneity: the conditional approach

The impact of the MOSE on the number of exceedances can not be considered exogenous. Methods to deal with the endogeneity of the MOSE binary variable should be taken into account to prevent endogeneity bias (see e.g. [Deb et al., 2006, Alfò et al., 2011]). To properly address this issue we introduce a conditional approach, resulting in a concomitant-variable model, that can be summarized as follows. The joint density of the main outcome, the binary MOSE variable and the latent states is defined as

$$f(y_t, MOSE_t, S_t) = \prod_{t=1}^{T} f(y_t \mid MOSE_t, S_t) f(S_t \mid MOSE_t) f(MOSE_t)$$

where $f(y_t \mid MOSE_t, S_t)$ is defined according to equations (3.3) and (3.4), $f(S_t)$ represents the semi-Markov chain defined by equations (3.1) and (3.2) and $f(MOSE_t)$ is the distribution of the endogenous binary variable. Nevertheless, we are not interested in modelling the endogenous binary variable, but rather to account for its effect on the main outcome to avoid misleading inferences. Accordingly, we work with the conditional model $f(y_t, S_t \mid MOSE_t)$. The conditional model implies that the endogenous binary variable is included as a further covariate in the linear predictors (3.5) and (3.6)

$$\log\left(\frac{p_t(\mathbf{x})}{1 - p_t(\mathbf{x})}\right) = \tilde{\mathbf{x}}_t'^{*} \boldsymbol{\delta} \tag{3.8}$$

and

$$\log(\lambda_{t,k}(\mathbf{x})) = \tilde{\mathbf{x}}_t'^{*} \boldsymbol{\beta}_k \tag{3.9}$$

with $\tilde{\mathbf{x}}_t^{*} = (1, \mathbf{x}_t, MOSE_t)$.

As the endogenous variable also affects the Hidden semi-Markov distribution, to complete the model specification, we define non-homogeneous transition probabilities and parameterize them as a function of the MOSE variable

$$\pi_{t,k|j} = \frac{\exp(\zeta_{0jk} + \zeta_{1jk} MOSE_t)}{1 + \sum_{h:h\neq j}^{K} \exp(\zeta_{0jh} + \zeta_{1jh} MOSE_t)}$$

where $\boldsymbol{\zeta}_{jk} = \{\zeta_{0jk}, \zeta_{1jk}\}$ is a vector of regression parameters, estimating the specific influence of the MOSE variable on each transition probability. The aim of such a specification is to understand how the evolution of the weather states is affected by atmospheric data.

We further model the μ_k parameter in 3.7 as a function of the MOSE as well, leading to $\mu_{t,k}$

$$\text{logit}(\mu_{t,k}(\mathbf{x})) = \tilde{\zeta}_{0k} + \tilde{\zeta}_{1k} MOSE_t.$$

with $\tilde{\boldsymbol{\zeta}}_k = \{\tilde{\zeta}_{0k}, \tilde{\zeta}_{1k}\}$.

The resulting model can be seen as an extension of the concomitant-variable model [Dayton and Macready, 1988, Wedel, 2002] to the Hidden semi-Markov case. Furthermore, it is important to remark that previous approaches using HMMs or HSMMs have included covariates in the observation model [Punzo et al., 2021, 2018, Maruotti and Punzo, 2017] or in the specification of transition probabilities [Maruotti and Rocci, 2012, Ailliot et al., 2015], but the inclusion of covariates in the model for state durations has been considered only recently [Lagona and Mingione, 2024, Ren and Barnett, 2023].

3.4.3 Maximum likelihood estimation

We estimate the model's parameters using maximum-likelihood estimation via the *forward* recursive representation of the likelihood; the number of latent states, K, is kept fixed and model selection criteria are used to select the most suitable value for K. Let $\boldsymbol{\Theta}$ be the short-hand notation for all the nonredundant model parameters corresponding to the vectors $\{\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\zeta}, \tilde{\boldsymbol{\zeta}}\}$ and let us define the forward variable

$$\alpha_t(k; \boldsymbol{\Theta}) = \Pr(Y_1, Y_2, \ldots, Y_t, S_t = k; \boldsymbol{\Theta}). \quad t = 1, \ldots, T; k = 1, \ldots, K.$$

We further assume that the maximum number of sojourn times is known, common to all hidden states (i.e $D_k = D$) and any state sojourns exceeding D are artificially censored [Maruotti et al., 2021]. Then it can be shown that the forward variable is defined according to the following recursions

$$\alpha_1(k; \boldsymbol{\Theta}) = \pi_k d_k(1) f_{1,k}$$

$$\alpha_t(k; \boldsymbol{\Theta}) = \pi_k d_k(t) \prod_{h=1}^{t} f_{h,k} + \sum_{u=1}^{t-1} \sum_{j \neq k} \alpha_{t-u;}(j; \boldsymbol{\Theta}) \pi_{t-u,k|j} d_k(u) \prod_{h=t+1-u}^{t} f_{h,k}, \quad t = 2, \ldots, D$$

and

$$\alpha_t(k; \boldsymbol{\Theta}) = \sum_{j=1}^{K} \sum_{u=1}^{D} \left[\alpha_{t-u}(j; \boldsymbol{\Theta}) \pi_{t-u,k|j} d_k(u) \prod_{h=t-u+1}^{t} f_{h,k} \right], \quad t = D+1, \ldots, T.$$

Thus, the likelihood is $\ell(\boldsymbol{\Theta}) = \sum_{k=1}^{K} \alpha_T(k; \boldsymbol{\Theta})$.

Model parameters can be estimated by direct numerical likelihood maximization (other approaches are available, see e.g. Yu [2010], O'Connell and Højsgaard

[2011]). [Barbu and Limnios, 2006] show that, under certain regularity conditions, the maximum likelihood estimators for finite-state discrete-time HSMM parameters are consistent and asymptotically normal. In detail, the maximum likelihood estimator is strongly consistent as $T \to \infty$ if:

a) for any states $j, k \in \{1, \ldots, K\}$ there is a positive integer u such that $\Pr(S_{t+u} = k \mid S_t = j) > 0$;

b) the conditional state sojourn distributions $d_k(\cdot)$ have finite support $\forall k \in \{1, \ldots, K\}$.

In the class of HSMMs with a finite state space, assumption a) means that the Markov chain is irreducible. This holds when all the states communicate with each other. Assumption b) automatically holds when we use the discrete non-parametric state duration distribution in the Hidden semi-Markov model because we explicitly assign probability mass to a finite collection of possible sojourns, as in [Langrock and Zucchini, 2011]. In our case, as the Logarithmic distribution has infinite support, we censor the distribution at a maximum duration $D = 200$ to satisfy the assumption.

In order to numerically maximize the likelihood with respect to the parameters one needs to take care of some technical problems, like e.g. underflow and, because we are using an unconstrained maximization algorithm, it is necessary to reparameterize the model in terms of unconstrained parameters. We use a general-purpose optimization based on quasi-Newton method, using the `optim` function in `R`. In the absence of covariates, we define the unconstrained working parameters, which are the logit of parameter $\mu_{t,k}(\mathbf{x})$ for each Logarithmic-series sojourn distribution, the logit of prior probabilities, the logit of each nonzero structural zero proportions, the log of each state-dependent Poisson means, and the generalized logit of the transition probability matrix, and then transform the estimates of the working parameters to estimates of the natural parameters; details are also given, without considering the covariates, in Zucchini et al. [2017, Ch.3]. As we are moving in a regression framework, we need to estimate the regression parameters only, linked to the natural parameters via appropriate link functions that guarantee that constrains on each set of parameters are met.

3.5 Results

3.5.1 Operational aspects

As the initial step in our empirical analysis, we undertake model selection by comparing several models: the simple zero-inflated Poisson regression, along with extensions using hidden Markov and hidden semi-Markov structures. These models are explored both with and without the inclusion of the endogenous MOSE variable. We fit each competing model across a range of states, K, varying from 2 to 4, resulting in a total of 14 fitted models.

Comparison between these models relies on evaluating the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) to determine the optimal value for K, with lower AIC and BIC values indicating better model fit. Alternative model selection criteria are also considered in the literature (see, e.g., Pohle et al. [2017]). Specifically, AIC selects $K = 3$ states for the conditional Zero-Inflated HSMM with a Logarithmic sojourn distribution. On the other hand, the BIC selects K= 2 components as the best model for the HSMM considering the MOSE and with a Logarithmic sojourn distribution. Overall the HSMM including the MOSE variable has a better performance with respect to the HMM.

In complex modeling frameworks, an efficient and reliable initialization of parameters is fundamental [Maruotti and Punzo, 2021]. Addressing computational issues in parameter estimation for hidden Markov models, Bulla and Berzel [2008] emphasize the significance of reasonable parameter initialization, especially during numerical maximization of likelihood. They suggest a short-EM strategy as a hybrid/two-step algorithm, a notion supported by Di Mari and Maruotti [2022]. In alignment with these insights, we adopt a similar strategy here. Initially, we fit a simple Zero-Inflated Poisson Hidden semi-Markov model employing a random start initialization strategy, estimating parameters via an EM-based algorithm [O'Connell and Højsgaard, 2011], while disregarding the role of covariates. This deliberate approach is chosen due to the sensitivity of both HMMs and HSMMs to latent distribution perturbations [Maruotti and Rocci, 2012], particularly in transition probability matrix and sojourn distribution of the underlying Markov chain. Subsequently, these estimates serve as starting values for the more complex model incorporating covariates. This hybrid/two-step algorithm **combines** advantageous elements of the EM and direct numerical maximization algorithms.

Table 3.2: Model selection criteria and comparisons.

Model	K	MOSE	Sojourn	log-likelihood	AIC	BIC
Zero-Inflated Poisson						
		✗	✗	-1583.834	3211.669	3330.321
		✓	✗	-1581.347	3210.693	3340.131
Zero-Inflated Poisson hidden Markov						
	2	✗	Geometric	-1439.127	2950.255	3144.412
	3	✗	Geometric	-1408.579	2921.157	3201.607
	4	✗	Geometric	-1389.813	2919.627	3297.155
	2	✓	Geometric	-1408.429	2898.858	3119.982
	3	✓	Geometric	-1395.172	2914.072	3248.727
	4	✓	Geometric	-1385.839	2945.677	3414.891
Zero-Inflated Poisson hidden semi-Markov						
	2	✗	Logarithmic	-1428.568	2929.136	3123.293
	3	✗	Logarithmic	-1390.400	2884.801	3165.251
	4	✗	Logarithmic	-1383.902	2907.804	3285.333
	2	✓	Logarithmic	-1404.564	2891.129	**3112.252**
	3	✓	Logarithmic	-1364.838	**2853.676**	3188.058
	4	✓	Logarithmic	-1353.660	2881.319	3350.533

3.5.2 Empirical results

As stated in Section 3.5.1, overall the best model selected based on the BIC and AIC is the conditional zero-inflated HSMM with a Logarithmic sojourn distribution and the MOSE vaiable as a further predictor. Looking at the estimated parameters, we found that the interpretation and the meaning of the three states selected by the AIC are clearer and more complete than the outcomes with two states. For this reason, in the following, we present results for the Hidden semi-Markov model with $K = 3$. In Appendix A results for the $K = 2$ HSMM counterpart are summarised. For the zero-inflated Poisson HSMM with $K = 3$ states, Table 3.3 presents parameter estimates alongside standard errors computed using a numerically differentiated Hessian matrix (see also Bartolucci and Farcomeni [2015]). Our analysis identifies three distinct states characterized by different weather conditions influencing the *acqua alta* (see Figure 3.5), where observations are assigned to each state via the Viterbi algorithm [Viterbi, 1967, 2006]. The steady-state (State 1) is visited for

1167 days (71.82% of the considered times) and for 1099 days (94.17%) we observe no events; the moderate risk-state (State 2) is instead visited for 345 days and with 66 zero event (19%); the risky-state (State 3) is visited for only 113 days with less than 5% proportion of zero events. In detail, most of the zeros are clustered in State 1, as expected. While various factors contribute to defining environmental states with different *acqua alta* risks in the Venice lagoon, meteorological forcing stands out as a predominant influence. During these events, a distinctive atmospheric pattern unfolds. From Figure 3.5, it is clear that the steady-states was the norm, over most the observed time period. Similarly, the other two risky states were likely to manifest during autumn and winter. Recently, in the last 12 months, the moderate-risk state was dominant, as a clear indication of a general climate change, where the steady-state, characterized by a zero or very low risk of *acqua alta* events, has been substituted by a new more risky typical situation as the one identified by the moderate-risk state.

In the first half of 2024 (January to June), moderate-risk and high-risk states together accounted for 50% of the days (43% for moderate-risk and 7% for high-risk). We expect these states to become predominant by the end of the year, as most *acqua alta* events occur in autumn and winter, indicating a significant shift in the general climate. In contrast, from 2020 to 2023, the steady state, characterized by zero or very low risk of *acqua alta* events, was observed on 74% of days, with moderate-risk and high-risk states accounting for 19% and 7% of days only, respectively. Let us characterize the states more in detail by looking at the coefficients presented in the Table 3.3 refer to: the probability of zero exceedances ($\boldsymbol{\delta}$), the Poisson rate in the steady-state ($\boldsymbol{\beta_1}$), in the moderate risk-state ($\boldsymbol{\beta_2}$), and in the risky-state ($\boldsymbol{\beta_3}$). At the bottom of the table, there are six additional sets of parameters: the estimates of sojourn times parameters ($\tilde{\zeta}_{0k}, \tilde{\zeta}_{1k}$), and the estimates of transition matrix parameters (ζ_{0jk}, ζ_{1jk}). The probability of zeros, as estimated by the logit model (3.5), significantly depends on all meteorological covariates except for the wind direction's components. As values of the wind speed increase, the probability of occurrence of zero exceedances decreases. Same reasoning can be applied to water temperature and relative humidity. Their interpretation is quite straightforward, as it can be associated to seasonality: colder periods of the year show more events of exceedances and vice-versa. As pressure values increase,

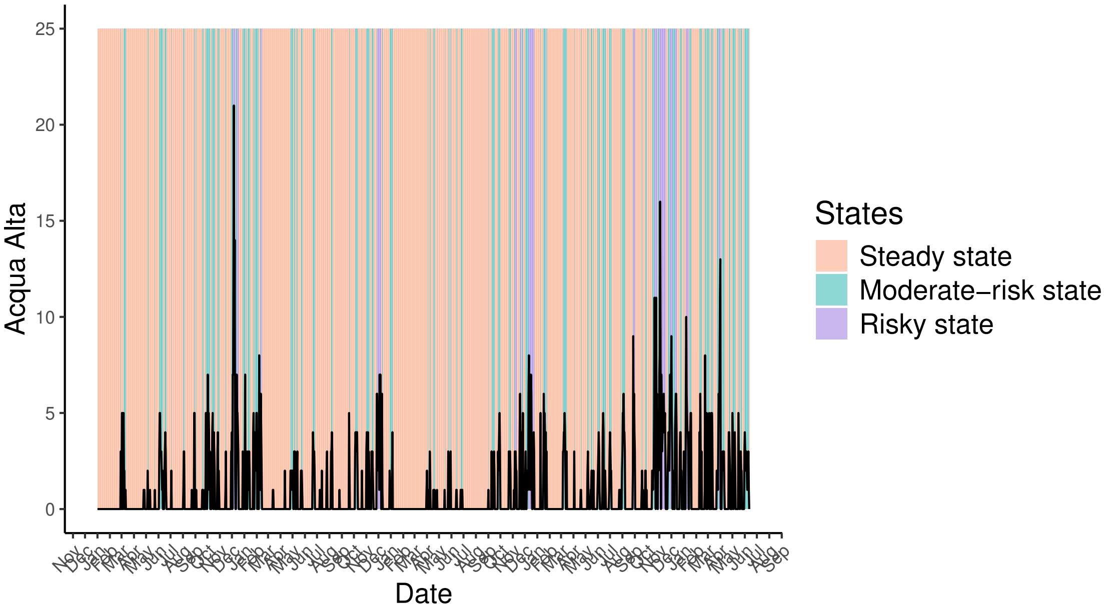


Figure 3.5: Observed counts and inferred classification obtained by the Viterbi algorithm.

Table 3.3: Estimated coefficients for the zero-inflated hidden semi-Markov model with $K = 3$ and endogeneity, as selected by the AIC.

Variable	$\boldsymbol{\delta}$		$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$		$\boldsymbol{\beta}_3$	
	Coefficients	Standard Errors	Coefficients	Standard Errors	Coefficients	Standard Errors	Coefficients	Standard Errors
Intercept	**6.654**	1.413	0.425	0.591	**0.341**	0.149	**1.685**	0.199
Wind Speed	**-1.309**	0.367	**-0.241**	0.087	**0.053**	0.025	0.037	0.028
Temperature	**-0.171**	0.096	0.054	0.042	0.017	0.009	-0.026	0.016
Humidity	**-0.147**	0.078	0.071	0.038	**0.012**	0.006	-0.002	0.007
Pressure	**0.246**	0.063	**-0.063**	0.024	**-0.073**	0.009	**-0.044**	0.007
Sin(Wind Direction)	0.034	0.615	**-0.444**	0.221	**-0.179**	0.072	0.109	0.094
Cos(Wind Direction)	-0.091	0.563	-0.021	0.225	**0.154**	0.065	**-0.233**	0.099
Year = 2020	Benchmark							
Year = 2021	**-2.339**	1.059	-0.409	0.321	0.124	0.169	**-0.507**	0.217
Year = 2022	-0.929	1.369	**-1.853**	0.917	**-0.515**	0.192	**-0.504**	0.198
Year = 2023	**-3.674**	1.502	**-1.236**	0.402	**0.290**	0.143	-0.052	0.216
Year = 2024	**-3.734**	1.647	-0.538	0.413	0.268	0.158	0.031	0.196
MOSE	-2.223	1.822	-0.731	0.389	**-0.474**	0.151	**-0.411**	0.132
			$(\tilde{\zeta}_{01}, \tilde{\zeta}_{11})'$		$(\tilde{\zeta}_{02}, \tilde{\zeta}_{12})'$		$(\tilde{\zeta}_{03}, \tilde{\zeta}_{13})'$	
			Coefficients	Standard Errors	Coefficients	Standard Errors	Coefficients	Standard Errors
Intercept			**13.089**	1.765	**4.373**	0.613	**6.175**	1.652
MOSE			**-11.591**	2.035	1.152	3.011	-2.379	2.295
			ζ_{013}, ζ_{113}		ζ_{023}, ζ_{123}		ζ_{032}, ζ_{132}	
			Coefficients	Standard Errors	Coefficients	Standard Errors	Coefficients	Standard Errors
Intercept			**-2.428**	0.512	-10.191	35.425	-9.372	35.414
MOSE			**3.449**	1.583	18.399	35.418	-2.026	35.414

there is a corresponding increase in the probability of not exceeding the *acqua alta* threshold. This is rather obvious, as higher air pressure correlates with lower sea levels. These results help in the understanding of steady-state conditions. Years' dummy variables regarding 2021, 2023 and 2024 show there have been less zero counts compared to the year 2020. This is somehow expected and in line with the idea of a general increase of the sea level over time, due to climate changes. Indeed, just to put these results under a more general view, over the past 60 years, the number *acqua alta* events in the last 9 years was more than twice those observed between 1870 and 1949, with a further increase in the last year.

Even under environmental conditions as those characterizing the zero-inflated state, we can still observe some counts as modelled by the Poisson distribution, corresponding to hours of *acqua alta* in correspondence of not specific extreme events, but rather to conditions likely to be observed in the lagoon. Three meteorological covariates emerge as significant factors influencing the count distribution parameter under the steady-state. Mainly, meteorological tides are influenced by air pressure and wind action. Air pressure results negatively significant, meaning that the higher the air pressure the lower counts of exceedances. Particularly, this state involves winds from the North-East (Bora) and South-East (Sirocco). These two types of wind are especially significant in Venice, as they are capable of pushing water masses into the lagoon, causing the sea level to rise. The Bora is a winter wind, and its effects can easily combine with other seasonal factors. The Sirocco blows throughout the year, mainly during the warm season, so its impact on the tide is often mitigated by factors such as atmospheric pressure. Moreover, the sine of the angle of the wind direction shows a negative impact on the counts in the steady-state, meaning that a decrease in the sine of the angle corresponds to an increase in the number of exceedances. This corresponds to winds shifting from East to South and East to North directions. It is, thus, useful to visualize the most frequent wind directions to better appreciate the characterization of the steady-state (see Figure 3.6). Being the steady-state, both Bora and Sirocco are observed, to further remark the typical wind characterization of the Venice lagoon. To get deeper into this aspect, we conducted an investigation by incorporating an interaction term between wind speed and wind direction in the model. However, the results indicated a preference for the initially analyzed model. In the moderate risk-state, some higher counts are clustered, indicating days with persistence of *acqua alta* events. These conditions can be observed all over the year, with exception for the summer season. Conversely to the steady-state, wind speed has a positive effect, meaning that an increase in wind speed leads to an increase in the number of counts. This observation must be linked to the interpretation of the cosine and

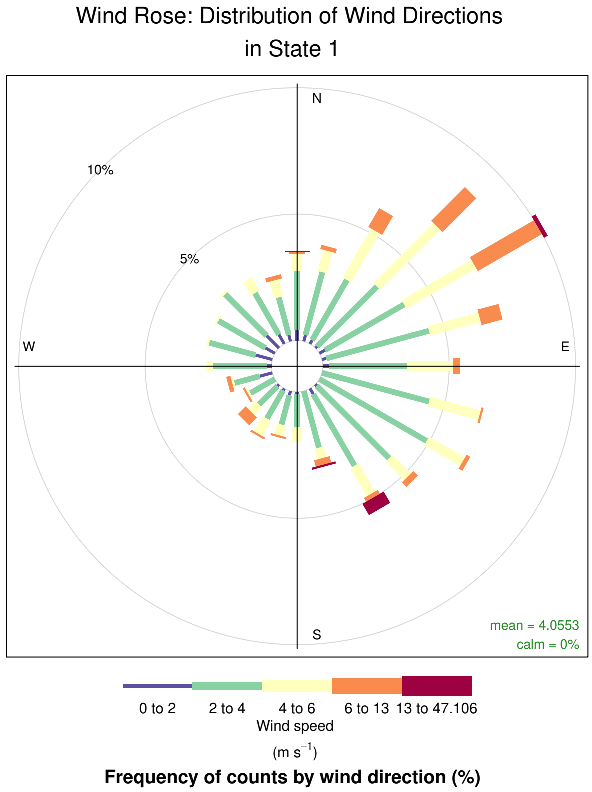


Figure 3.6: Wind Rose displaying prevalent wind directions and wind speed in State 1.

sine of the angle of the wind direction. The latter is significant with a negative effect on the counts of *acqua alta* events, while the cosine of the angle has a positive impact, implying that wind directions shifting from East to North (Bora wind) contribute to increase the expected number of *acqua alta* events. This becomes clearer by looking at the wind rose (Figure 3.7), which shows that the predominant wind responsible for this state conditions within the lagoon is the Bora wind (NE) with mainly high wind's speed.

Atmospheric pressure shows negative significant estimate as in the steady-state. On the other hand, relative humidity has positive impact on the exceedances counts. The risky-state is rare but shows the higher number of hours of *acqua alta* during a day, happening mainly during the winter and autumn season. In this case only two environmental components are significant. The air pressure, which holds negative and the cosine of the angle which presents a negative sign. The negative coefficient of the cosine indicates that the lower the cosine the higher the counts and, as discussed before, this can be attributed to wind directions going from North to

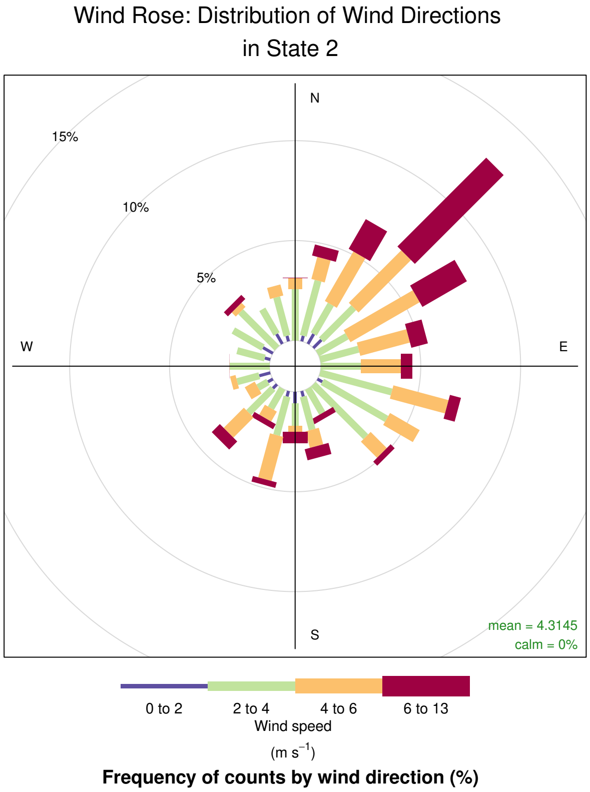


Figure 3.7: Wind Rose displaying prevalent wind directions and wind speed in the moderate risk-state.

East, precisely identifying the Bora wind. This can be further proved by the wind rose presented in Figure 3.8. Here we can see that this state is mainly driven by the Bora.

These results remark the importance of wind directions in generating *acqua alta* events, and identify the South-East (Sirocco) and North-East (Bora) winds as steady-states of the lagoon characterization. Nevertheless, in-depth analysis shown in Figures 3.6, 3.7 and 3.8 emphasizes the significant impact of robust Bora winds on *acqua alta* events in the Venice Lagoon, with a minor impact of the Sirocco on the more extreme events. The identification of environmental states with different *acqua alta* risks is supported by several researches describing heterogeneous conditions in the Venice lagoon (see e.g. Umgiesser et al. [2021], Trigo and Davies [2002]) and multiple conditions as generating processes. Moreover, the time dynamics is stressed and properly modelled in a probabilistic framework.

In this respect, the years-specific dummy variables further contribute to the

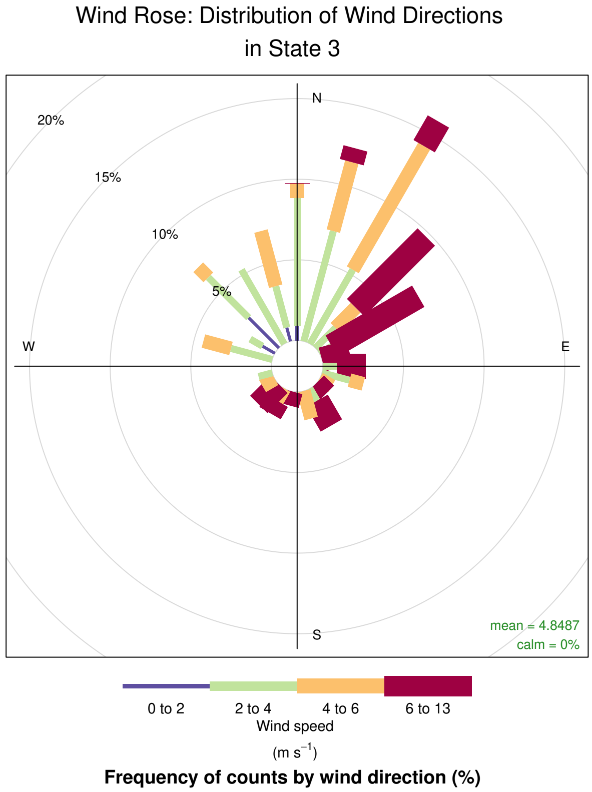


Figure 3.8: Wind Rose displaying prevalent wind directions and wind speed in State 3.

understanding of the evolution of *acqua alta* due to general climate changes, not directly measured by the considered covariates. Two major aspects can be identified. Firstly, the probability of an *acqua alta* event is much higher than 2020, with a clear sharp increase in 2023 and 2024. On the other hand, however, the number of extreme events, as clustered by the risky state is not much different from the beginning of the observation period. What is of interest is, instead, that medium-risk events, i.e. some consecutive hours of *acqua alta* events, but not extreme, increase in the last two years. Let us now focus on the impact of the MOSE, as there is a relevant debate on its efficacy, also due to its potential impact on the lagoon ecosystem as well. The activation of MOSE is a noteworthy factor in reducing the number of occurrences of *acqua alta* events. Indeed, the MOSE significantly reduces the number of *acqua alta* hours in the moderate-risk and risky states. On the other hand, it does not contribute to reduce the probability of an event, as it strongly depends on environmental covariates only, as discussed above. The role played by the MOSE does not reduce to limiting the number of hours of *acqua alta*. Additionally, when examining the parameters' estimates of the sojourn distribution, we observe a substantial influence of MOSE on the time spent in the steady-state ($\tilde{\zeta}_{11} = -11.591$). Specifically, MOSE significantly decreases the duration of permanence in this state. At first instance, this may sound like a negative impact of the MOSE in the management of the risk in the lagoon. To appreciate this result, it is crucial to emphasize that the decision to activate the MOSE barriers is based on a short-term prediction. This finding demonstrates the effectiveness of the short-term forecast model used to make decisions on the activation of the MOSE. Every time that the forecasting model suggests a worsening of the environmental conditions, which may cause significant levels of *acqua alta*, the MOSE is activated, indicating that the predictions made by the technical committee in Venice are reliable and effective in guiding timely interventions. Of course, this implies a shorter time into the steady-state, where mainly no occurrences were clustered. This can be further supported by looking at the coefficients of the multinomial regression linked to transition probabilities. The coefficient for the MOSE variable ($\zeta_{113} = 3.449$) shows a higher probability of moving from the steady-state to the risky one, when the MOSE is active. In this respect, due the binary nature of the MOSE variable, it is easy and of potential interest to display the transition probability matrices when the MOSE is active and when it is not. Thus, we have the following transition probability matrix

$$\begin{bmatrix} 0 & 0.919 & 0.081 \\ 0.999 & 0 & 0.001 \\ 0.999 & 0.001 & 0 \end{bmatrix}$$

if the MOSE is not active and

$$\begin{bmatrix} 0 & 0.265 & 0.735 \\ 0.000 & 0 & 1.000 \\ 1.000 & 0.000 & 0 \end{bmatrix}$$

if the MOSE is active. In the matrices, each row represents the current risk state and each column represents the next risk state. For instance, in the first matrix the element in row 1, column 2 (0.919) indicates a high probability of transitioning from state 1 (low risk) to state 2 (moderate risk), while the element in row 1, column 3 (0.081) suggests a much lower probability of transitioning from state 1 to the worst risk state (state 3). Of course, it is not the MOSE activation that causes a change in the risk associated to the inferred hidden states, but it is rather a further evidence of the good use made by technicians of the MOSE, that is activated if and only if a relevant worsening in the environmental conditions is expected. Indeed, to summarize the results from the two transition probability matrices above:

- if the MOSE is not active, the transition to the worse risk state (as seen in row 1, column 3 of the first matrix) is unlikely, i.e. not expected to be observed;
- if the MOSE is not active, any sudden or unexpected high risk event is followed by the steady-state (represented by high probabilities in row 2 and row 3 transitioning to state 1 in the first matrix), i.e. the MOSE is activated if there is the belief of a persistent worsening of the environmental conditions such that it justifies its cost of use of roughly $300,000, with workers dispatched to the artificial island, as patrol boats cordon off maritime traffic;
- if the MOSE is active, there is the belief that conditions would worsen and as such we observe high probabilities of moving towards the most risky state (reflected in row 1, column 3 and row 2, column 3 of the second matrix);
- following the latter consideration, the MOSE contributes to move towards the steady-state after a period of high risk (seen in row 3, column 1 of the second matrix).

3.6 Discussion

Latent-variable models, as HSMMs, offer a useful strategy to interpret complex environmental data, by modelling multiple features of the data under a single framework, often separating the analysis in multiple sub-problems, e.g. time-dynamics, the influence of covariates on the observed process parameters, etc. Our proposal is a latent-variable model for zero-inflated time series of counts that extend the literature on HSMMs to the concomitant-variable framework. In the considered case study on the *acqua alta* events in the Venice lagoon, the model is capable of offering a clear-cut description of risk periods in terms of intuitively appealing environmental states. It provides a classification that reflects the impact of

environmental covariate on the events of interest, with a focus on the orography of lagoon. It finally captures the influence of the MOSE on the evolution of the latent-variable, i.e. the hidden semi-Markov chain, in both the transitions across different risky states and the duration of each state. Though tailored to issues that arise in marine studies, the proposed model can be easily extended to a wide range of real-world cases, where the interest is not only on the segmentation of the data according to time-varying latent classes, but also on the influence of covariates on sojourn times within each latent class.
The model is fully parametric and therefore exposed to misspecification issues. Non-parametric extensions can in principle be developed either at the observation level of the model, i.e. on the count data distribution, or at the latent level, assuming a non-parameteric sojourn [Langrock and Zucchini, 2011], separately estimating the whole set of mass probabilities under each state. This strategy is computationally tractable only when the study involves a long time series with short dwell times and it does not guarantee against wiggly dwell time distributions with implausible gaps and spikes. Recently, [Pohle et al., 2022] suggest penalized likelihood methods to estimate the dwell time distribution in a (homogeneous) HSMM. This idea could in principle be integrated in our proposal, at the price of additional computational burden.
We compared the HSMM and the HMM, and though we did not discuss differences in the estimated parameters in depth, we would like to remark a few points. The HMM is often perceived as a simpler model than the HSMM; this is partially true, as the explicit modelling of the sojourn distribution increases the computational burden. Nevertheless, nothing is known a-priori about the sojourn distribution of the latent process and its choice is part of the model selection process, and this calls for the use of the HSMM. The geometric distribution, as implicitly assumed by the HMM, is just one of the available sojourn alternatives and should be investigated along with other count distributions. We noticed that both the inferred clustering and the regression parameters estimates strongly depend on the assumed hidden structure (this is in line with the results of Maruotti and Rocci [2012], who show that biased estimates of the parameters driving the hidden structure would lead to misleading inferential results). Results, available upon request, about the estimated parameters according to the HMM with $K = 3$ qualitatively and substantially differ

from those presented in Section 3.5, and are in general not straightforward to be interpreted, as a further of the need of estimating a more suitable hidden structure, starting from a proper sojourn distribution for the data at hand.

From an empirical perspective, our model explores the complex dynamics associated with the MOSE project's influence on water levels in the Venice lagoon. An hourly dataset would be more useful in providing a detailed understanding of the MOSE activation dynamics. It is rather clear that, by considering closures at the inlets, we unveil that the MOSE is an effective defense mechanism against future high-water events, further remarking the efficacy of the MOSE in mitigating flooding risks. Nevertheless, there are other environmental aspects which warrant further attention. Incorporating more comprehensive data on the ecological impacts of the MOSE project will enhance our understanding of its broader consequences on the lagoon's ecosystem. Changes in water flow patterns, sediment transport, and salinity levels may impact the lagoon's delicate ecosystem. Ongoing research should address these ecological dimensions, providing an overall view of the MOSE project's influence on the Venice lagoon.
In conclusion, our modeling approach sheds light on the intricate connection of environmental factors, state transitions, and the impact of the MOSE project on *acqua alta* events in the Venice lagoon. While our findings represent a significant step forward, continuous refinement and expansion of our methodology remain crucial for a more comprehensive understanding of the evolving dynamics in this unique and vulnerable ecosystem.

Chapter 4

Spatio-Temporal Variability in Mediterranean Sardine Abundance: A Bayesian Approach[1]

4.1 Paper Summary

This chapter represents a pivotal transition in the thesis: from the analysis of physical and hydrodynamic marine conditions and hazard dynamics to the modelling of living marine resources and ecosystem productivity. The paper addresses one of the most pressing ecological and economic challenges in Mediterranean fisheries: the overexploitation and environmental sensitivity of small pelagic species, particularly the European sardine (*Sardina pilchardus*). Through the development of a hierarchical Bayesian spatio-temporal model, the study contributes novel insights into the dynamics of sardine landings across the Mediterranean basin, providing standardized indices of relative abundance, also known as Catch per Unit Effort (CPUE), and uncovering spatial and temporal trends, crucial for sustainable fisheries management. The Mediterranean Sea is one of the most biologically diverse and socio-economically significant marine ecosystems in Europe, supporting millions of livelihoods and a complex network of small-scale fisheries. However, over

[1]Ricciotti, L., Paradinas, J., Pollice, A., & Martino, S. Spatio-Temporal Variability in Mediterranean Sardine Abundance: A Bayesian Approach. Stochastic Environmental Research and Risk Assessment, *Under Review.*

half of the assessed fish stocks in the region remain exploited beyond sustainable limits, with sardines and anchovies among the most affected. Within this context, the chapter focuses on investigating the spatio-temporal variability of sardine landings using fishery-dependent data, which offer high spatial and temporal resolution. The analysis is based on data from the Joint Research Centre's (JRC) Data Dissemination Tool, integrating information collected under the EU Data Collection Framework (DCF).

From a methodological perspective, the study demonstrates how Bayesian spatio-temporal models can effectively standardize CPUE indices using fishery-dependent data. By integrating effort variables, vessel and gear characteristics, and spatial and temporal structures, the models produce relative abundance indices that account for major sources of bias and variability. The use of Stochastic Partial Differential Equations (SPDE) within INLA ensures computational efficiency and full posterior inference, enabling uncertainty quantification crucial for stock-assessment applications. Bayesian spatio-temporal modelling with INLA and SPDE are introduced here and reappear in later chapters for the analysis of environmental, economic, and policy data. Conceptually, this chapter represents the living resource component of the thesis: quantifying the dynamics of a key species that lies at the intersection of ecological sustainability and economic activity. Together with subsequent chapters on coastal-defence funding, it contributes to a comprehensive statistical evaluation of Mediterranean coastal systems encompassing environmental, biological, and financial dimensions. The study also demonstrates the strength of combining open European data sources with modern Bayesian computation, offering a replicable framework for other species or regions.

4.2 Introduction

The Mediterranean Sea is one of the most biologically diverse and economically significant marine ecosystems in Europe. Hosting over 500 marine species, it plays a crucial role in sustaining both biodiversity and human livelihoods. Among its many contributions, the fisheries sector stands out as a key economic driver, supporting local economies, employment, and food security throughout the region. In 2021, the Mediterranean and Black Sea produced nearly 2 million tonnes of seafood, worth USD 20.5 billion. Marine capture fisheries contributed 1.06 million tonnes and USD 7.8 billion, highlighting their importance to coastal communities and food systems. The fishing fleet in the region comprises over 84,000 vessels, 87% of which operate in the Mediterranean Sea. In particular, small-scale vessels represent the majority of the fleet, accounting for 82% of all vessels and providing 61% of on-board employment, yet contributing about 26% of total revenue [General Fisheries Commision for the Mediterranean, 2023].

Despite its ecological and economic value, the Mediterranean Sea is facing mounting challenges. Over 50% of assessed fish stocks in the region are currently being fished beyond biologically sustainable limits, with critical overexploitation reported for key species such as sardines, anchovies, European hake, red mullet, and swordfish [Altmayer, 2024]. Although some improvements have been observed in recent years, such as a 15% reduction in overexploited stocks between 2020 and 2021, 58% of assessed stocks are still fished outside biologically sustainable limits, and the average fishing pressure remains twice the sustainable level [Guadagnoli and López, 2024]. Among the most heavily exploited species in the Mediterranean are the European sardine and European anchovy, which together constitute a significant share of total landings. Particularly, sardines account for 17.8% (36.455 tonnes) of landings in the Western Mediterranean, 11.4% (20.328 tonnes) in the Central Mediterranean and 47.4% (57.589 tonnes) in the Adriatic Sea [General Fisheries Commision for the Mediterranean, 2023]. These small pelagic species are particularly vulnerable due to their short life cycles, sensitivity to environmental change, and the intensity of fishing pressure they face [Martín et al., 2012, Hidalgo et al., 2022]. Beside the overexploitation problem, the Mediterranean is also facing big changes due to global warming. Several studies have documented how warming waters, habitat degradation, and shifting ecological regimes are negatively impacting the productivity and distribution of key species [Moullec et al., 2019, Galappaththi et al., 2022].

These persistent pressures reflect broader challenges, including overfishing, habitat degradation, and the impacts of climate change, all of which threaten the resilience and productivity of the region's marine ecosystems. Colloca et al. [2017b] provide an overview of the fishing exploitation in the Mediterranean Sea pointing out the Western Mediterranean and the Adriatic Sea as the less sustainable ones. Moreover, Fiorentino and Vitale [2021] suggest different ways to face the overfishing and to make the fishery processes more sustainable. These include the adoption of species-specific technical measures and the use of monitoring technologies such as Vessel Monitoring Systems (VMS) and Automatic Identification Systems (AIS) to support the enforcement of spatial restrictions. Further, Singh and Weninger [2024] highlight that in multi-species fisheries, allowing some level of discretion in harvest decisions, rather than imposing rigid rules, can improve economic efficiency and better align harvest patterns with ecological and market variability. Last, Tahvonen et al. [2018] show that management targets based only on Maximum Sustainable Yield (MSY) may deviate from economically optimal outcomes.

In 2021, the General Fisheries Commission for the Mediterranean (GFCM) established a management plan to face overexploitation of small pelagic species in the Adriatic Sea for the time span 2022-2029 [General Fisheries Commission for the Mediterranean (GFCM), 2021]. This plan is divided in two phases: a transitional one based on the fixed annual reductions of the total catch of sardine and anchovy and a long-term one starting in 2025 [General Fisheries Commission for the Mediterranean (GFCM), 2024a], where a yearly catch per species is set annually based on the status of the stocks and a harvest control rule (HCR). The adoption of this management plan aims at keeping the exploitation levels at MSY by the end of 2029. Last reports published by the GFCM about sardines and anchovies in the Mediterranean [General Fisheries Commission for the Mediterranean (GFCM), 2024b] showed that applied models to the data can help in understanding the level of exploitation of the species but there still are problems in estimating standardized Catch Per Unit Effort (CPUE) index. For instance, sardines stocks result overexploited especially in the Adriatic Sea where a reduction of biomass and fishing mortality are experienced despite a reduction of catch. Indeed, sardines landings are experiencing a decrease from 2014 onwards [General Fisheries Commission for the Mediterranean (GFCM), 2024b].

Given this context, the present study focuses on the European sardine. The main objective is to investigate the spatial and temporal patterns of sardine catches across the Mediterranean Sea, with the goal of identifying trends, variations influencing their distribution and abundance over time, and producing standardised catch per unit effort (CPUE) indices. Because sardine fishing activity is not uniformly distributed across the basin, we concentrated our analysis on specific regions where this species is most intensively caught. In particular, we examined 23 Geographical Sub-Areas (GSAs) located in the Western Mediterranean, encompassing zones from the Spanish coasts through the Tyrrhenian Sea to the Adriatic Sea. Previous research has explored various aspects of sardine population dynamics and fisheries in the Western Mediterranean. For instance, Calvo et al. [2024] proposed Bayesian longitudinal linear mixed models to capture temporal dynamics in both artisanal and industrial sardines fisheries, confirming the persistent declining trend in Mediterranean sardine landings. Also, other studies have investigated the spatial distribution of sardines [Gordó-Vilaseca et al., 2021, Tugores et al., 2010]. However, there remains a clear need for modelling approaches that jointly account for spatial and temporal dynamics within a robust statistical framework. Additionally, standardizing catch and effort data remains essential to obtain unbiased abundance indices, and a variety of statistical approaches—ranging from generalized linear models to more flexible methods like GAMs and GLMMs—have been developed to address this challenge [Maunder and Punt, 2004]. Recent reviews have consolidated best practices for CPUE standardization, highlighting key methodological considerations such as model selection, covariate handling, treatment of zero-inflated data, and uncertainty quantification [Hoyle et al., 2024]. In fisheries science, data are generally classified into two main categories: fishery-independent and fishery-dependent data. As discussed in detail by Pennino et al. [2016], these data types differ significantly in their sources, design, and potential biases. Fishery-independent data are collected through scientific surveys specifically designed for research purposes. Although they often offer limited spatial and temporal coverage due to logistical and resource constraints, they are considered more reliable and less biased, as data collection is conducted under controlled and standardized conditions. In contrast, fishery-dependent data originate from commercial fishing operations, such as logbooks, landings, and observer programs. These data sources are inherently more susceptible to biases, stemming from non-random sampling, under-reporting, or changes in fisher behaviour. However they typically provide broader spatial and temporal resolution and offer valuable details on the fishing effort and the composition of the catch. Understanding these differences is crucial for interpreting results and selecting appropriate modelling strategies. We focus exclusively on fishery-dependent data, which are well-suited to the scope of our analysis. To address these gaps, our work integrates spatial and temporal components, providing a powerful and flexible tool for modelling sardine abundance dynamics and estimate reliable CPUE indices. To carry out the analysis, we developed spatio-temporal models within a hierarchical Bayesian framework and used the Integrated Nested Laplace Approximation (INLA) to compute posteriors [Rue et al., 2009, 2017], implemented via the `inlabru` package in the `R` software environment [Lindgren et al., 2024, Bachl et al., 2019]. INLA has been widely applied in various domains, including environmental science, epidemiology, economics [Moraga et al., 2021, Panunzi et al., 2025], and increasingly in fisheries research. Several recent studies illustrate how INLA has been effectively applied to analyse fisheries data of varying types. For example, Soto et al. [2023] used INLA to investigate the drivers of discarding in fisheries. INLA has also been employed to examine the influence of environmental variables on species distribution; for instance, Deen et al. [2025] studied the abundance of summer flounder in relation to environmental changes. Additionally, several studies have used the INLA framework to estimate relative abundance indices for marine species [Cavieres and Nicolis, 2018, Staeudle et al., 2024]. In our study, we implemented the Stochastic Partial Differential Equation (SPDE) approach, which provides a computationally efficient approximation of spatially continuous Gaussian random fields by leveraging Gaussian Markov random fields. This significantly reduces computational costs, making it particularly suitable for large spatial and spatio-temporal datasets. As a result, SPDE has become a popular method in applied research [Fichera et al., 2023, Cameletti et al., 2013, Fioravanti et al., 2023]. For instance, Rubino et al. [2024] used the SPDE approach to investigate the effects of temperature on demersal fish species in the Central Mediterranean Sea. Building on this line of research, our study applies advanced Bayesian spatio-temporal methods to fisheries data with the aim of identifying key drivers of catch variability and supporting sustainable fisheries management in the region. The following sections present the data (Section 2), outline the proposed modelling framework (Section 3), and describe its application to a case study of sardine landings in the Western Mediterranean (Section 4). We end with a discussion in Section 5.

4.3 Materials and Methods

4.3.1 The Western Mediterranean sardines' fisheries data

We consider data from the Data Dissemination tool provided by the Joint Research Centre (JRC) of the European Commission. It provides access to data submitted by EU Member States to the European Commission under the Data Collection Framework (DCF) [European Parliament and Council, 2017]. Particularly, we focused on the EU Fisheries Dependent Information (FDI), which is published as the result of the 2024 DCF data call [European Commission: Joint Research Centre, Scientific, Technical and Economic Committee for Fisheries (STECF), 2025]. Specifically, we collected spatial data on landings and efforts [Gibin et al., 2024]. Landings represent the total tons of fish landed while efforts refer to the number of days with at least one fishing activity (fishing days). The original data time-span goes from 2013 to 2023 and covers the 27 EU Member States' fisheries activities. Here we focus on the Mediterranean Sea; specifically, on 23 Geographical Sub-Areas (GSAs) covering mostly the Western part of the Mediterranean Sea (Figure 4.1).

Data have a spatial resolution of $0.5° \times 0.5°$ grid and each observation is linked to its corresponding GSA. Note that, the area of each grid cell is variable, reflecting both the curvature of the Earth and the partial overlap of some cells with the coastline. All spatial coordinates are expressed in the World Geodetic System 1984 (WGS 84) coordinate reference system. Moreover, we reduced the dataset to the years 2017–2022, since in these years both fishing effort and sardine landings are consistently represented, while in previous years data are scarce.

Beside spatial location and year, the recorded information includes vessel length, gear type (i.e. the type of equipment used to catch fishes), species caught, year,

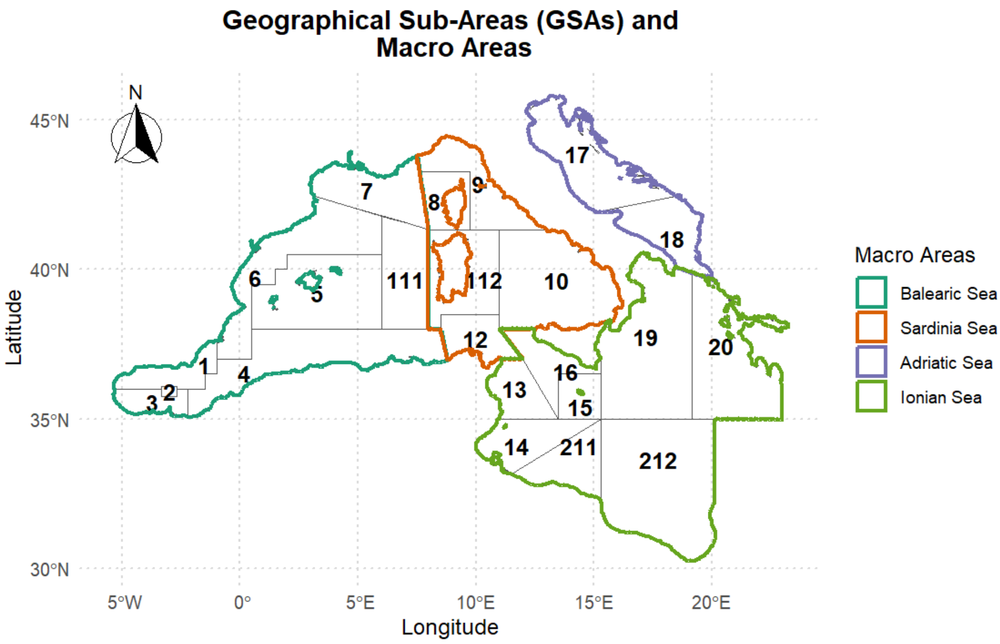


Figure 4.1: Geographical Sub-Areas (GSAs) of interest. The thicker, coloured, borders indicate the 4 macro-areas considered in Model 4.

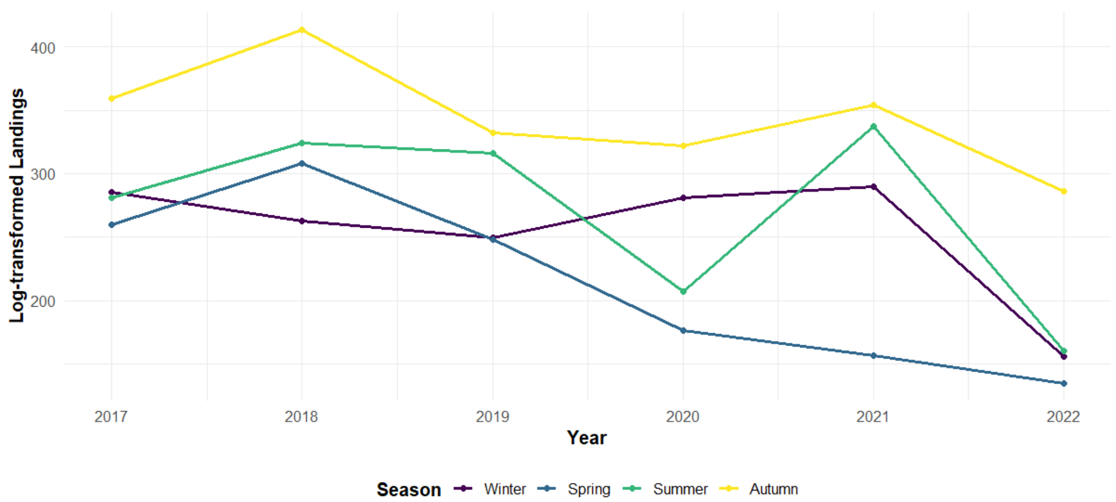


Figure 4.2: Seasonal trends in log-transformed sardine landings from 2017 to 2022.

quarter (with winter being the first quarter), fishing days, landings and area of the grid's cell (km^2).

Figure 4.2 illustrates the temporal and seasonal trends in log-transformed sardine landings from 2017 to 2022. Overall, landings have declined across all seasons over the study period, with a particularly sharp drop in 2022. Despite this declining trend, a consistent seasonal pattern is evident: higher landings tend to occur during the warmer seasons, summer and autumn, compared to winter and spring. This suggests that seasonal environmental conditions or fishing practices may play a role in shaping catch patterns.

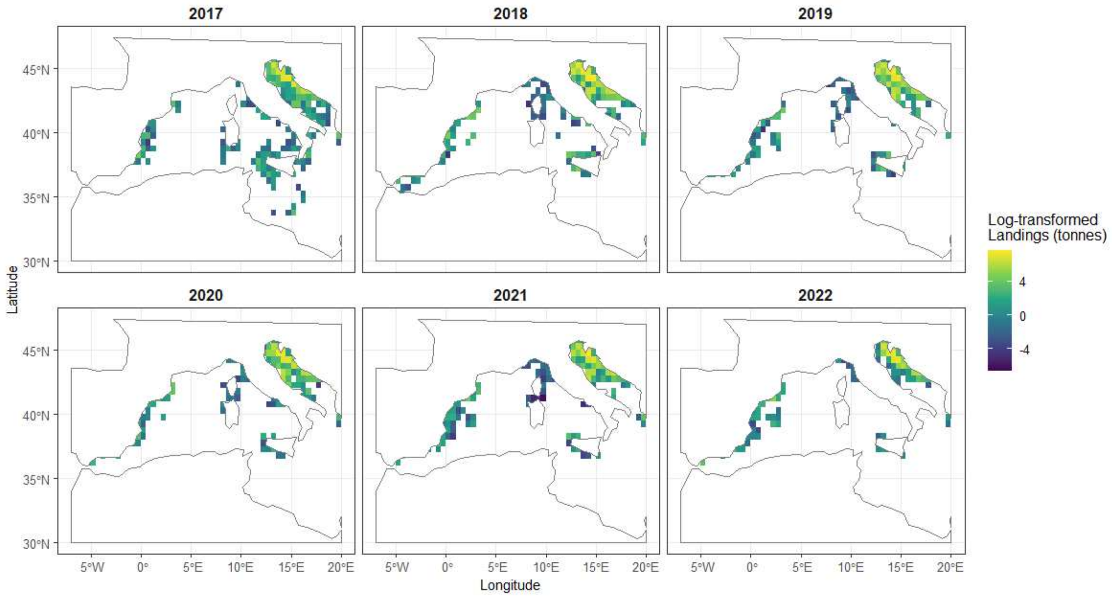


Figure 4.3: Spatial distribution of sardines' log-landings across different years.

Figure 4.3 shows the spatial distribution of the sardines' landings in the area of interest from 2017 to 2022. The highest concentration of sardines' landings is in the Northern Adriatic Sea, and some along the Spain and Sicilian coasts. This distribution is quite usual since sardines are caught mainly in shallower waters, thus along the coastline.

Figure 4.4 represents the spatial distribution of fishing days. The latter, as defined in the context of EU fishery-dependent data collection, refer to the number of days during which a vessel is active at sea and engaged in fishing operations. This metric serves as a proxy for fishing effort and is commonly used in fisheries analyses to standardize catch and landings across vessels and time periods. As for the sardines landings, the highest effort can be found in the Northern Adriatic Sea,

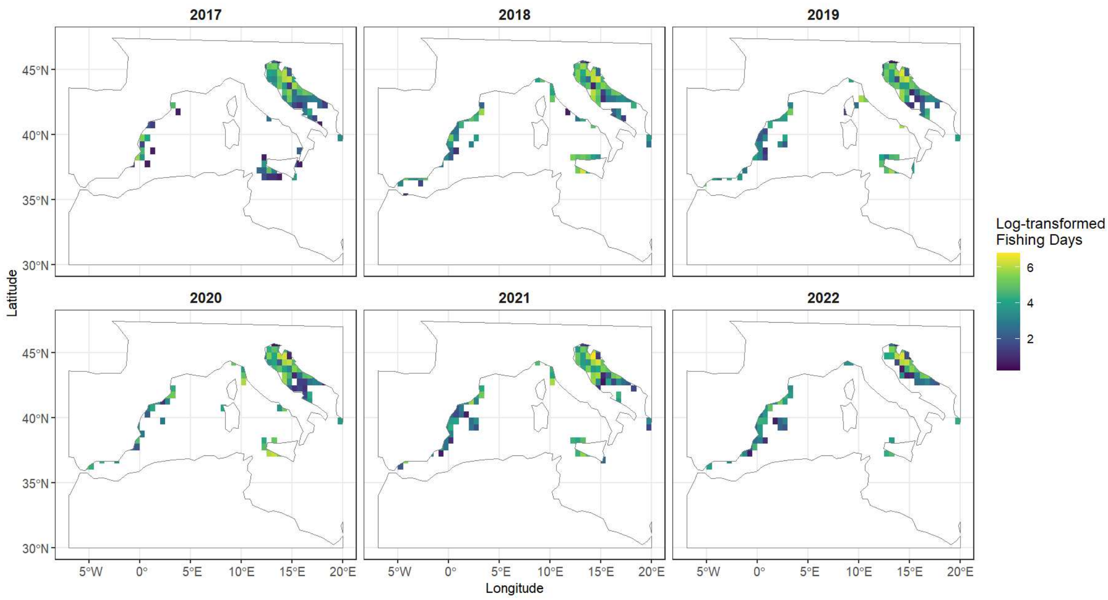


Figure 4.4: Spatial distribution of log-fishing days across different years.

and along the Spain and Sicilian coasts.

We selected three gear types: purse seine (PS), midwater otter trawl (OTM), and midwater pair trawl (PTM), which are used for sardine fishing in the Mediterranean. Figure 4.5 (left panel) presents the distribution of sardine landings with respect to the three gear types. Landings associated with purse seines and pelagic trawls show the highest intensity of catches. In contrast, midwater otter trawls (OTM), shows very limited intensity, indicating that these gear types play a minor role in sardine exploitation. Figure 4.5 (right panel) shows the distribution of landings with respect to vessel length. It clearly show that vessels between 12 and 40 meters in length are responsible for the majority of sardine catches across the Mediterranean. In contrast, landings associated with very small vessels (< 6 m) and very large vessels (> 40 m) reflect limited operational range or fleet size, or different targets. This distribution highlights the dominant role of medium-sized vessels in regional sardine fisheries.

Together, these data provide a comprehensive spatio-temporal overview of sardine landings across the Mediterranean, stratified by gear type and vessel characteristics. Notice that, by using landings as the response and effort along with vessel and gear characteristics as covariates, the fitted models provide standardized

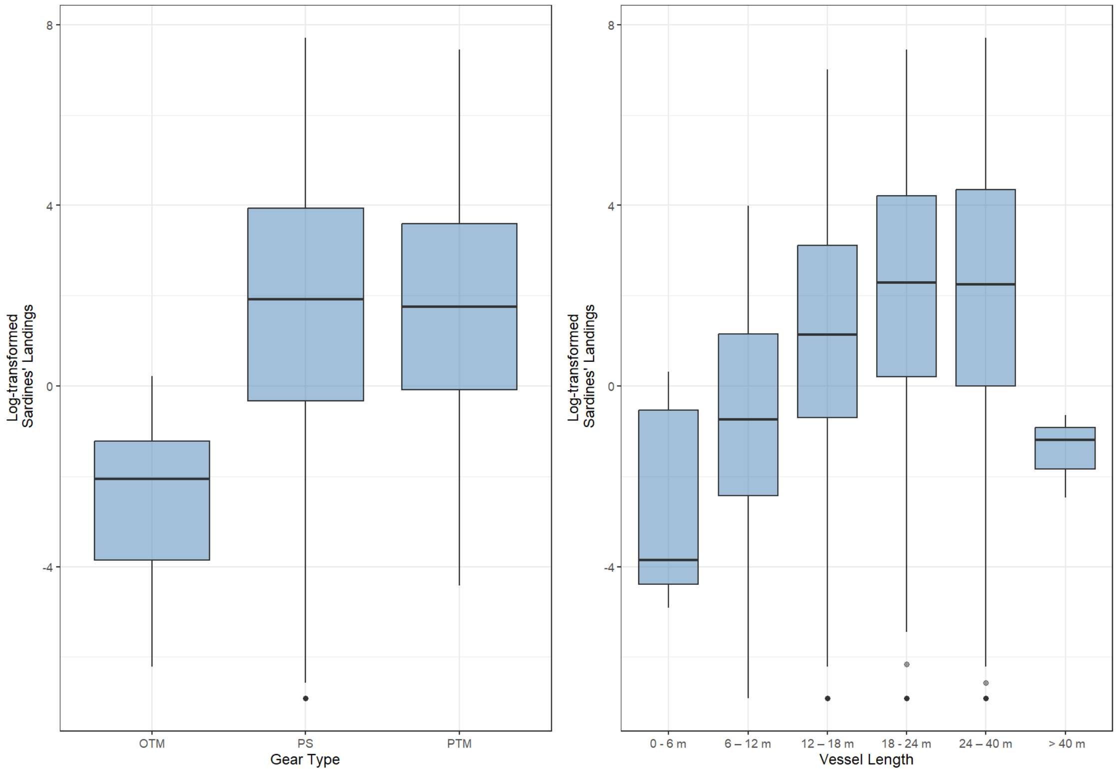


Figure 4.5: Boxplots showing the distribution of log-transformed sardine landings across different gear types (left panel) and vessel length classes (right panel).

CPUE indices, offering a relative measure of sardine abundance over time that accounts for differences in fishing activities and operational practices. Building on this foundation, our analysis aims at exploring spatial, and temporal trends, and potential drivers underlying these observations. Particularly, since the exploratory analysis reveals that spatial distribution basically stable with respect to time, we focused on exploring temporal dynamics with different models.

4.3.2 Bayesian Spatio-Temporal Model using SPDE

We consider $Y_t(s)$ sardine landings at location s and time $t = 1, \dots, T \cdot Q$, where $T = 6$ denotes the years (2017–2022) and $Q = 4$ represents the quarters. The response variable is assumed to follow a Gaussian distribution:

$$\log(Y_t(s)) \sim \mathcal{N}(\mu_t(s), \sigma_y^2),$$

where σ_y^2 is the residual variance, and $\mu_t(s)$ represents the expected log-landings. The mean $\mu_t(s)$ accounts for both spatial and temporal variability and includes key covariates related to fishing effort and fleet characteristics. We evaluated four different models. These models share the fixed and spatial effect while have distinct specifications of the temporal component. Each model includes the area of each grid cell as an offset term to account for differences in spatial coverage. The models are defined as:

$$\mu_t(s) = \beta_0 + \beta X_t(s) + \delta(d) + \omega(s) + f(t) \tag{4.1}$$

where β_0 is a common intercept, the matrix $X_t(s)$ represents the fixed effects of the quarter, gear type, and vessel length factor covariates, and β is a vector of unknown parameters to be estimated. We allow the number of fishing days covariate (d) to have a non-linear effect by modeling $\delta(d)$ as a second-order Random Walk (RW2). Spatial dependences are captured by $\omega(s)$, a Gaussian spatial field with Matérn covariance structure defined as:

$$\text{Cov}(\omega(s), \omega(s')) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa\|s - s'\|)^{\nu} K_{\nu}(\kappa\|s - s'\|), \tag{4.2}$$

where σ^2 is the marginal variance, ν is the smoothness parameter which we set to 2, κ is the scale parameter related to the range ρ of the spatial correlation, and K_ν is the modified Bessel function of the second kind. This Gaussian random field (GRF) is modelled using the Stochastic Partial Differential Equation (SPDE) approach [Lindgren et al., 2011]. This method allows for a computationally efficient approximation of Gaussian random fields by representing them as Gaussian Markov random fields defined over a triangulated mesh.

The temporal component $f(t)$ is defined differently in each of the four models: Model 1 assumes a simple linear effect of time. Model 2 increases the complexity by allowing for non-linear time effect. This is achieved by assuming a RW2 prior for $f(t)$. Model 3 and Model 4 are meant to investigate the hypothesis that the temporal trend is not constant. To do this, in Model 3, we allow each GSA to have it's own non-linear temporal trend. We assume that, in the prior, time trends in each GSA are independent on each other, but they share the same smoothness parameter. Some of the GSA have very little observed data. This can make result in high uncertainty for some of the estimated trends. To mitigate this, in Model 4, we aggregate the GSAs into 4 macro-areas: the Balearic Sea, Sardinian Sea, Ionian Sea, and Adriatic Sea. Each macro-area has its own temporal trend. As for Model 3 we use a RW2 model, no correlation is assumed for the 4 macro-areas and the 4 RWs share the same smoothing parameter.

To complete the model specification within a Bayesian framework, we defined prior distributions for all parameters. Fixed effects, were assigned vague Gaussian priors (standard deviation equal to 100). For the range and standard deviation of the spatial field $\omega(s)$ we used Penalized Complexity (PC) priors [Fuglstad et al., 2019]. Specifically, the prior on the range ρ_ω is defined such that $\mathbb{P}(\rho_\omega < 3) = 0.5$, while the prior on the standard deviation σ_ω satisfies $\mathbb{P}(\sigma_\omega > 1) = 0.5$.

Regarding the temporal components, Model 1 specifies it as a linear effect with Gaussian prior $\mathcal{N}(0, 1000)$. In Models 2, 3, and 4, the temporal effect is modelled as a second-order random walk (RW2), with precision parameter τ given a PC prior so that $\mathbb{P}(\rho_\omega < 1) = 0.5$. The models were implemented in R using the `inlabru` package [Bachl et al., 2019], which provides a flexible framework for specifying spatial and spatio-temporal latent models.

4.4 Results

The models provide standardized measures of relative abundance (CPUE), enabling comparisons of sardine distribution across space and time. Model performance was assessed using two widely accepted model selection criteria: the Deviance Information Criterion (DIC) and the Watanabe-Akaike Information Criterion (WAIC), as reported in Table 4.1. Among the four, Model 3 achieved the lowest values for both DIC and WAIC, indicating the best fit to the data.

Both estimated covariate effects and spatial random field are virtually identical

Model	DIC	WAIC
Model 1	14681.28	14663.90
Model 2	14650.09	14633.07
Model 3	**14586.92**	**14564.97**
Model 4	14601.04	14583.99

Table 4.1: Model selection criteria, DIC and WAIC for the four fitted model.

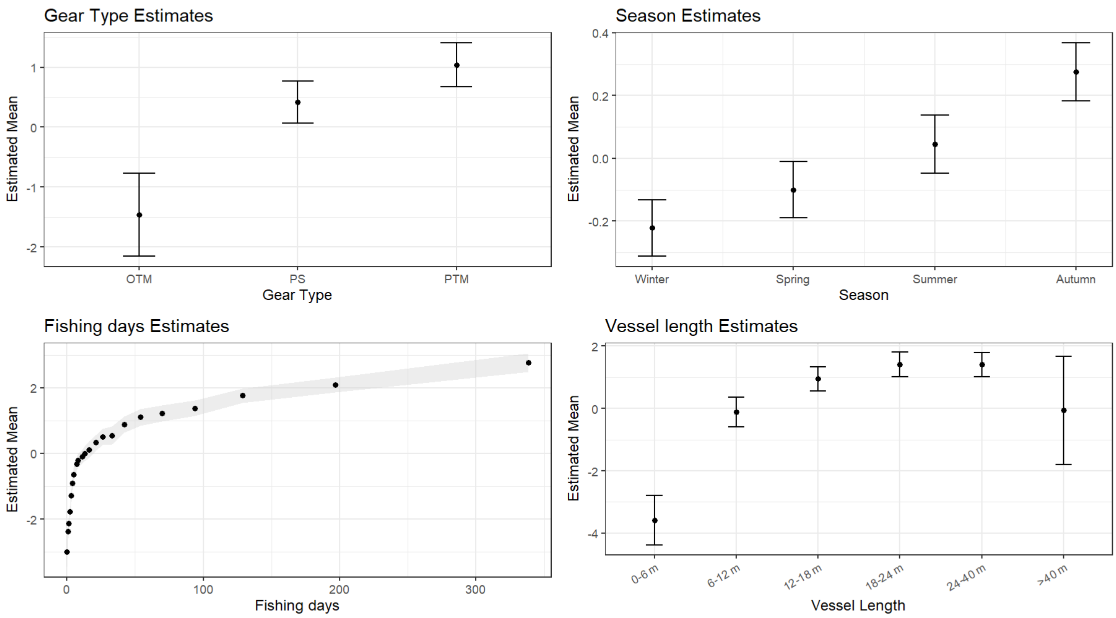


Figure 4.6: Posterior estimates of the covariates in Model 3. The four panels display the estimated posterior means with their associated 95% credible intervals.

for the four models. We therefore present these results only for Model 3. At a later stage we present the estimated time effects for all four models. Estimates for the models' hyperparameters are available in the Appendix.

Figure 4.6 presents the posterior estimates and 95% credible intervals for the covariate effects in Model 3. The top-left panel illustrates the effect of gear type: PS and PTM show substantially higher posterior means, with PTM being the most effective, suggesting that these gears are more commonly employed and successful in targeting sardines. The top-right panel shows the estimated effect of seasons, with autumn having the highest estimates, followed by summer, and winter the lowest. This pattern aligns with sardine biology, as spawning occurs in winter, while higher abundances in warmer months increase catch rates [Basilone et al., 2021]. The bottom-left panel shows the effect of the number of fishing days, which represents an estimate of effort. The relationship appears non-linear and exhibits diminishing returns: catch increases rapidly with the number of fishing days, but begins to flatten-out after approximately 100 days. This suggests that while effort is initially rewarded, beyond a certain threshold, additional fishing days do not yield proportionally higher catches. The bottom-right panel presents the posterior mean estimates associated with vessel length. Vessels between 12 and 40 meters exhibit the highest posterior means with relatively narrow credible intervals, indicating that these size classes are most strongly associated with higher sardine landings. Overall, the results suggest that medium-sized vessels are most aligned with higher sardine landings within the fishery.

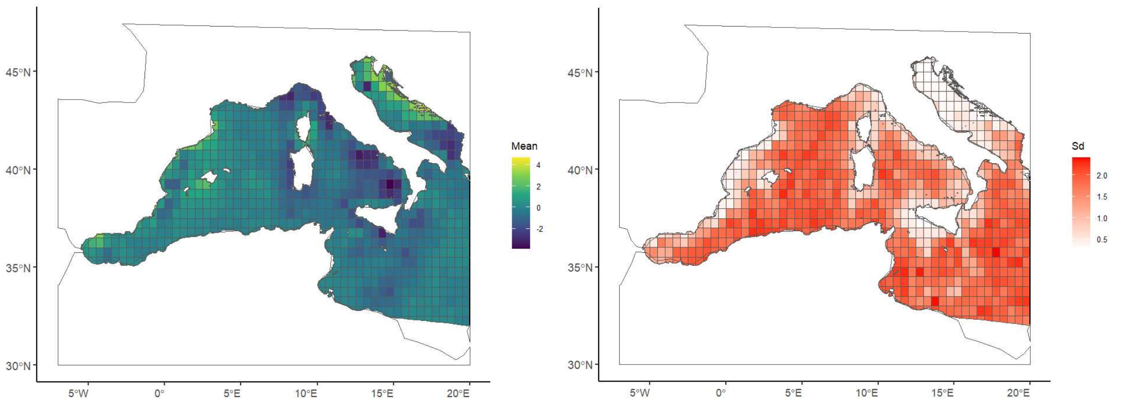


Figure 4.7: Spatial effect estimates for Model 3. The left panel displays the posterior mean estimates, while the right panel shows the corresponding posterior standard deviations.

The spatial dynamics are illustrated in Figure 4.7, which maps the posterior mean (left panel) and standard deviation (right panel) of the spatial random effect for Model 3. The posterior mean shows distinct spatial patterns in sardine fisheries, with higher values concentrated along specific coastal regions, such as parts of the Adriatic and the Balearic Seas, suggesting areas of higher expected catch. Conversely, regions in the western Mediterranean and parts of the southern basin exhibit lower values, indicating reduced catch potential. The right panel highlights the spatial uncertainty, with lower standard deviations in data-rich areas, while

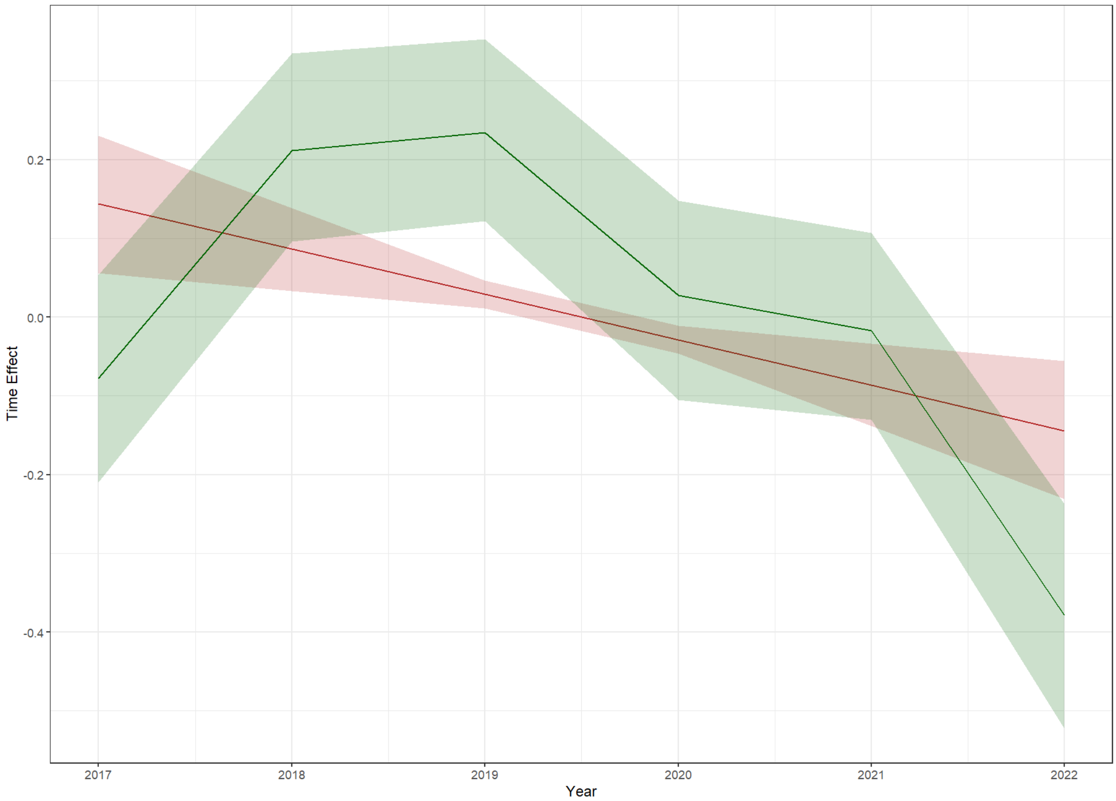


Figure 4.8: Time effects for Models 1 and 2. Posterior means for the effect captured by Model 1 (red line) and for the non-linear effect of Model 2 (green line). The shaded areas represent the 95% credible intervals.

higher uncertainty is observed in the southern and eastern parts of the basin, likely due to sparser data coverage. We compare now the estimated time trends for the four models: while Models 1 and 2 are simpler in structure, they still offer valuable insights regarding the temporal dynamics.

Figure 4.8 shows the different time effect for Model 1 (red line) and Model 2 (green line). Model 1 suggests a negative linear trend over time. This signal is present also in Model 2, but the more complex model allows to capture more details: for the first two years there has been an increase, while for last years a steeper non-linear declining trend.

Overall, several southern and eastern GSAs (e.g., GSA16, GSA19, GSA20, GSA211, and GSA212) exhibit a clear declining trend, suggesting a consistent reduction in sardine abundance or fishing activity over the study period. In contrast,

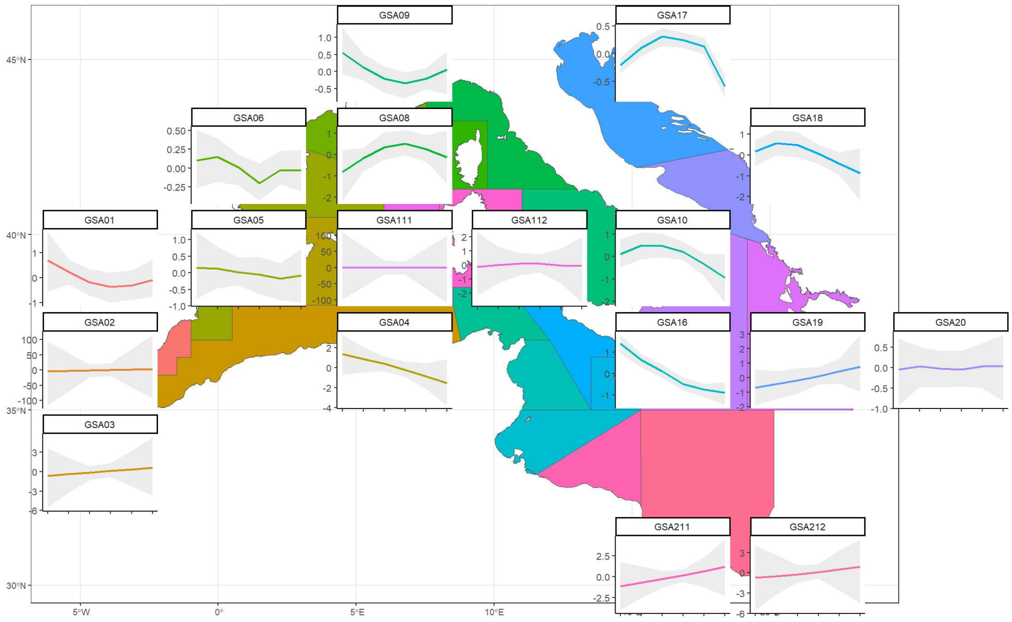


Figure 4.9: Posterior mean estimates and 95% credible intervals (shaded bands) of the temporal effect on sardine catches, modelled independently for each GSA (Geographical Sub-Area).

some northern and western GSAs (e.g., GSA17, GSA18) show a hump-shaped pattern, possibly indicating localized increases in abundance or fishing effort during the middle years, followed by a subsequent decline. Other regions (e.g., GSA05, GSA06, GSA08) display relatively stable trends with limited interannual variability. These spatially heterogeneous temporal patterns likely reflect the combined influence of local environmental conditions, management measures, and varying fishing pressures. The uncertainty bands—wider in some GSAs (e.g., GSA20)-highlight areas where the model exhibits greater uncertainty, potentially due to limited data availability or higher temporal variability.

Model 4 aggregates the 23 GSAs into four broader spatial units, thereby reducing uncertainty in the estimated temporal effects through the pooling of information across regions. Figure 4.10 displays the temporal effects for Model 4. The temporal patterns reveal distinct dynamics across regions. The Adriatic and Ionian Seas exhibit a consistent and marked downward trend, indicating a decline in sardine abundance or fishing activity over the study period. This may reflect environmental degradation, stock depletion, or regulatory impacts specific to those macro-areas. In contrast, the Sardinian and Balearic Seas show relatively stable trends with

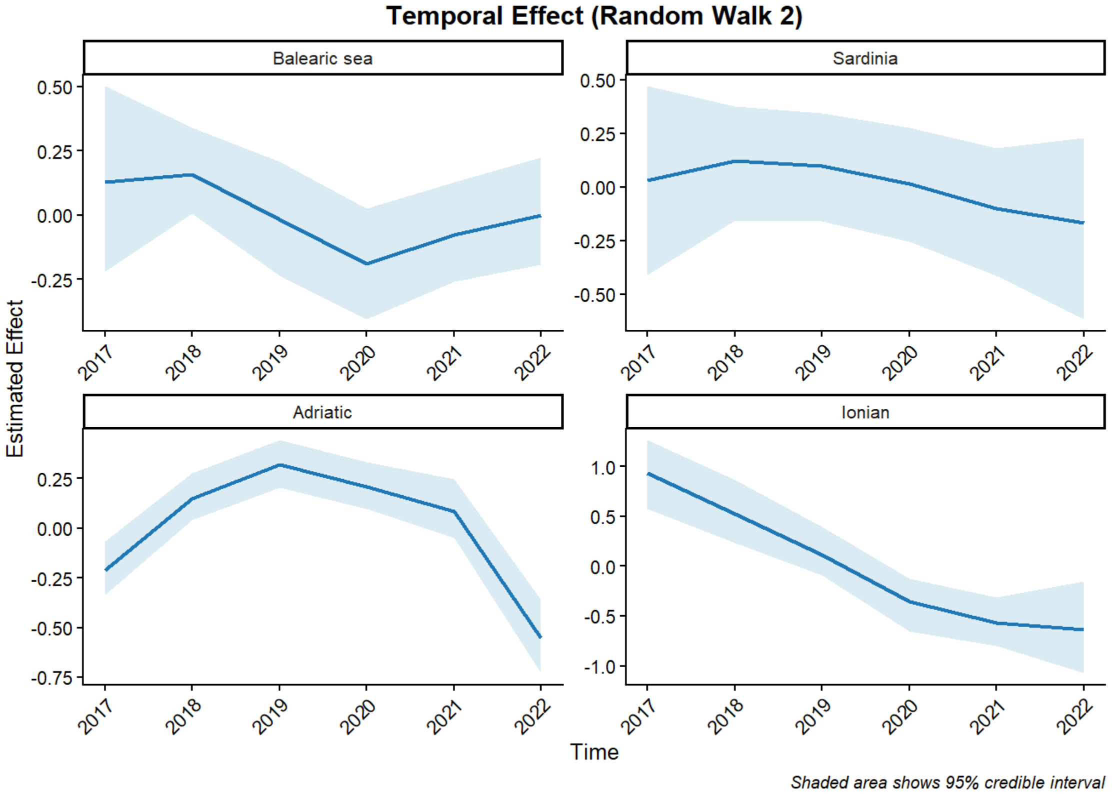


Figure 4.10: Posterior mean estimates and 95% credible intervals (shaded bands) of the temporal effect on sardine catches, modelled independently for macro-areas.

modest fluctuations and no pronounced directional change, suggesting a more stable sardine population or effort in these regions. The smoother temporal patterns in Model 4 reflect the effect of spatial aggregation, which reduces noise but does not capture finer-scale dynamics. Overall, this approach captures broader temporal signals while still retaining geographic specificity, offering a more parsimonious yet informative structure compared to modelling each GSA separately.

4.5 Discussion

This study provides a comprehensive spatio-temporal analysis of sardine fisheries in the Mediterranean Sea, highlighting the complex interplay between environmental, seasonal, and fishing-related factors. By fitting hierarchical Bayesian models with differing temporal structures, we were able to characterize and understand the dynamics of sardine catches across both space and time. The suggested models are able to effectively standardize sardine commercial catches into catch per unit effort (CPUE) by accounting for: three different effort variables (fishing days, gear and vessel length); spatial heterogeneity in catches; and both seasonal and annual temporal trends. Particularly, annual trends varied across space, which may suggest the use of a spatial stock assessment model [Punt, 2019] to manage the sardine fishery in the Mediterranean. In that case, and depending on the spatial resolution of the stock assessment model, either Model 3 or Model 4 estimates could serve as CPUE indices. In case a spatial stock assessment model was not considered, the CPUE index should be retrieved from Model 2. Our results confirm that sardine catches are seasonally driven, with the strongest positive effects occurring during autumn, followed by summer and spring. Winter exhibited a negative effect, suggesting this season is least favourable for sardine fishing. These patterns align well with known biological cycles, where sardines typically reproduce in winter and are more actively fished during the warmer months [Basilone et al., 2021]. Including the fishing effort variable, measured as the quantile of fishing days, as a smoothed random effect was important for capturing latent effort-related variability. Although its precision estimate revealed a strong smoothing effect, the wide credible interval suggests some remaining complexity that may require further modelling. Both Model 3 and Model 4 include a spatial random effect modelled via SPDE, and the resulting spatial field was consistent across models. The spatial structure was characterized by a moderate marginal standard deviation and a relatively short range of spatial correlation, indicating that localized spatial heterogeneity plays a significant role in shaping sardine distributions. The spatial effect clearly identifies areas of higher and lower expected sardine abundance, which may correspond to favourable environmental conditions or localized fishing pressure. The main difference between the two models lies in the specification of the temporal random effect. Model 3 captures temporal variation with separate RW2 processes for each GSA, while Model 4 aggregates these into four macro-regions (Balearic, Sardinian, Ionian, and Adriatic Seas). The temporal trends from Model 3 reveal regional differences, particularly in the eastern and southern parts of the basin, where declines or more variable patterns are observed. In contrast, the smoother trends from Model 4 provide a more general overview, potentially facilitating communication with stakeholders, though possibly at the expense of regional detail. The precision estimates for the temporal components highlight these differences: while both models suggest substantial smoothing, Model 3's higher uncertainty may reflect the increased complexity of modelling time separately for each GSA. Model 4 offers greater parsimony but may under-represent localized trends.

Our findings have clear implications for the spatial and temporal targeting of sardine fisheries. The strong seasonal effects support seasonal closures during reproductive periods to protect stock regeneration. The spatial heterogeneity underscores the need for area-specific management measures rather than uniform regulations. Moreover, understanding regional temporal trends can help forecast future availability under changing environmental conditions or fishing practices.

Chapter 5

An integrated georeferenced dataset of public investments for soil defence in Italy[1]

[1]Ricciotti, L., & Pollice, A. (2024). An integrated georeferenced dataset of public investments for soil defence in Italy. Data in Brief, 57, 110909.

5.1 Paper Summary

In this chapter of the thesis, we shift focus from modelling physical marine and coastal processes toward the policy-infrastructure dimension of coastal ecosystems. The core objective of the chapter is to build and release a harmonised, georeferenced dataset of public works investments directed to soil-defence across Italian municipalities, thereby providing a foundational resource for spatial-statistical and policy-oriented analyses. In the landscape of coastal and marine ecosystem evaluation, infrastructure investments for soil and coastal-defence are a critical component: they reflect the human responses to flooding, landslide risks, erosion, and sea-level pressures. Yet, despite the importance of investment data, publicly available financial datasets are often fragmented, inconsistently coded, lacking geospatial linkage. In Ricciotti and Pollice [2024c] we addressed this gap by assembling data from three major Italian open-data platforms: the OpenBDAP platform of the Ministry of Economy and Finance, the OpenCoesione portal on cohesion-policy financed interventions, and the ReNDiS database of the Italian Institute for Environmental Protection and Research (ISPRA), dedicated to soil-defence works. By linking these sources and by adding municipal-level geographic referencing, the dataset provides a spatially explicit, harmonised view of soil-defence investments across Italy.

Within the thesis structure, this paper plays a crucial role in establishing the public investment infrastructure dimension. After the earlier chapters focusing on physical sea-state modelling and hazard modelling, this chapter moves into the macro-policy/investment sphere. It supplies the raw dataset and processing logic necessary for the subsequent analytic Chapters 6 and 7. Beyond the immediate uses in this thesis, the dataset contributes to broader research and policy fields: it facilitates evidence-based policy by enabling visualisation and statistical summary of where soil-defence funds are going, enabling policymakers to identify under-invested areas, regional disparities, or misalignment between risk exposure and funding. It allows cross-disciplinary linkage of infrastructure investment data with environmental science, hydrology, coastal engineering, land-use planning, and socio-economics, supporting integrative approaches to coastal resilience. From a methodological standpoint, the process of merging multiple public open datasets, geographic referencing at the municipal level, and documenting the variable structure exemplifies good practice in open data harmonisation and spatial-statistical infrastructure. As for statistical learning applications, the dataset is ideally suited for spatial-econometric modelling, clustering of municipalities by investment patterns, regression of investment on risk/driver variables, network analysis of fund flows, and even machine-learning classification of high vs. low-investment areas conditioned on environmental exposures.

5.2 Value of the Data

- The dataset on public works for soil defense in Italy is the result of merging the data from three different repositories: OpenBDAP, OpenCoesione and ReNDiS. This merging process enhances the richness and completeness of the dataset providing a comprehensive overview of soil defense financial efforts across the country. By consolidating information from multiple sources, the dataset becomes valuable for policymakers and researchers alike, offering a unified and comprehensive resource for analysis and decision-making.

- The data holds significant value for various stakeholders, especially Italian policymakers and researchers, due to its specificity. This specificity is crucial as the data provides detailed information for each individual project and can be aggregated at the municipality level. The dataset's uniqueness lies in its comprehensive coverage of each municipality, representing a notable advancement in soil defense research. Unlike previous studies that often-lacked granularity, this dataset delves into localized investment details. As a result, it enables deep insights into investments made and outcomes achieved, supporting informed decision-making, resource allocation, and the development of innovative solutions to mitigate soil-related risks. Additionally, the availability of such detailed and comprehensive data promotes collaboration and knowledge-sharing among stakeholders involved in soil defense.

- This dataset on soil defense investments in Italy, incorporates also the temporal dimension. Through temporal analysis, researchers can track trends in funding allocations and policy emphasis over time, providing insights into the evolution of soil defense strategies and priorities. This temporal approach allows for the identification of critical periods of investment growth or decline, highlighting shifts in governmental priorities or environmental concerns. Moreover, researchers can assess the long-term effectiveness of soil defense initiatives by examining how investment patterns correlate with changes in soil consumption indicators over time. By leveraging this temporal perspective, researchers can not only understand current investment trends but also anticipate future needs and challenges, enabling proactive decision-making to safeguard Italy's soil resources for generations to come.

- This dataset can seamlessly integrate with other datasets on various levels. For instance, it can be coupled with environmental data to explore how environmental factors influence the allocation of funding for soil defense public works. Additionally, socio-economic indices can be utilized to analyze whether contextual factors impact expenditure patterns. This interoperability enables researchers and policymakers to conduct comprehensive analyses, uncovering correlations and insights that can inform more effective decision-making and resource allocation strategies in soil defense initiatives.

5.3 Background

Public works encompass infrastructures and services that are typically financed and undertaken by government authorities to serve the collective needs of the society. They relate to a diverse range of facilities and initiatives, including transportation systems, water and sewage treatment facilities, energy infrastructure, public buildings, and other essential structures. These are designed to enhance the overall well-being of the community by fostering economic development, ensuring public safety, and improving the quality of life for residents. Particularly, public works for soil defence mitigate the hydro-geological risks posed by various environmental factors, such as erosion, landslides, and flooding. They aim at safeguarding both infrastructures and ecosystems. The motivation behind constructing a dataset of public works for soil defence stems from the necessity to address some key challenges in the research on public policies for environmental management. Existing datasets often lack consistency and specificity, impeding researchers' ability to analyse trends and assess the effectiveness of public investments comprehensively [Zuiderwijk et al., 2012]. As the digital landscape evolves, there's a growing demand for comprehensive, organized data to facilitate evidence-based decision-making [Commission, 2003, Burdick et al., 2015, Welter et al., 2023, Delina et al., 2024]. By aggregating detailed information on soil defence investments across Italy, including temporal and spatial dimensions, this dataset fills a critical gap, offering researchers a nuanced understanding of evolving funding patterns and policy priorities.

5.4 Data Description

The dataset provides comprehensive information on soil defense public works conducted throughout Italy Figure 5.1. Specifically, it includes details on the total funding allocated to each intervention, along with precise location data indicating the municipality where each project was implemented.

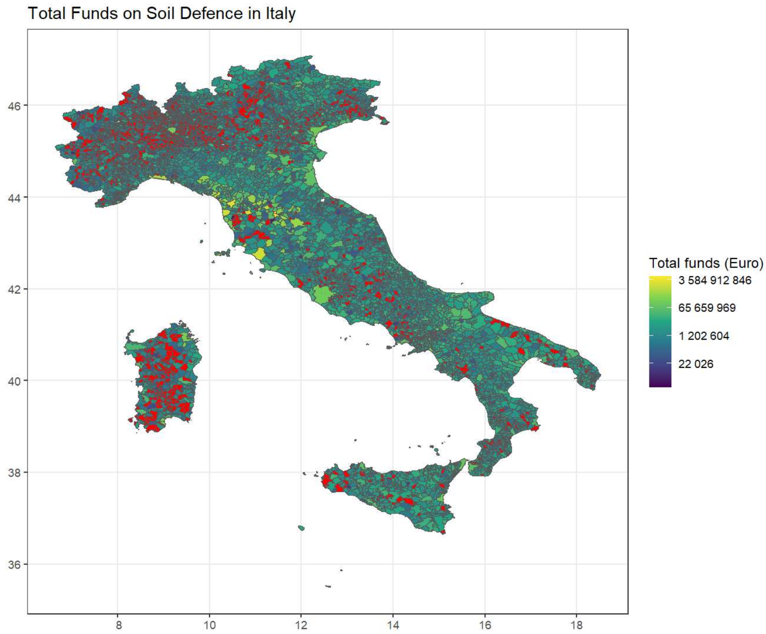


Figure 5.1: Italy map on soil defence total investment. Red areas show municipalities without information on soil defence investments.

The dataset is available on Zenodo [Ricciotti, 2024], and its latest static version is updated as of February 2024. This particular version was created on March 16, 2024, at 4:50 PM. It is composed of four different folders named after the sources from where data are taken plus one called “Data Region”, where the merged version of the three datasets can be found. Each folder contains: datasets for each Italian region, hence 20 files. Names of each dataset are composed of "sd_" which stands for "soil defense", followed by the name of the source, thus "obdap", "oc" or "rendis, followed by _regionname. In addition, the main folder includes a file called "Italy.csv" which contains the data on soil defense for the entire country, and a file named "metadata" where names of the variables and information on them can be found. Available datasets are in CSV format and contain 28 variables. These include two descriptive variables regarding the intervention, one of which details the

Unique Project Code (CUP) associated with each intervention. Additionally, there is information on the title and description of the intervention, along with the rationale behind its necessity. Sometimes, specific location details such as the address of the works are also provided. A financial variable denotes the total investment amount, while six geographic variables capture region, province, and municipality codes and denominations. Sixteen variables track key dates in the public works process. There's one variable indicating the source of the data. Furthermore, two geo-reference variables are included; one indicating point coordinates, and the other delineating the polygonal shape of each municipality area (see Table 5.1).

In the repository there is also a file called metadata where information regarding each variable can be found. The temporal expanse of the collected data is particularly notable, extending from the late 90's to the current date. Specifically, there are relatively few observations up to the early 2000's, attributed to the challenges of tracking interventions and a lack of robust data infrastructure. Municipalities exhibit varying spans of available information, reflecting the uneven temporal coverage of the dataset. This variability in temporal data coverage further complicates comprehensive analysis and monitoring of soil defence interventions across different regions. Additionally, the dataset can be considered complete and consistent starting from a specific year. Specifically, from 2000 onwards, the data can be regarded as consolidated for the entire national territory. Figure 5.1 shows some missing municipalities. In particular, there are 1699 municipalities which do not have any records. It is possible to see that the majority of these municipalities are very small, and mainly located in the hinterland. Several factors might explain why hinterland municipalities lack specific data on soil defence interventions. First, geographic and demographic characteristics might contribute to the absence of recorded information. Furthermore, limited resources and budgets, lower population densities, and accessibility issues can hinder the implementation and documentation of such measures. Additionally, there might be a lack of awareness or expertise regarding soil defence strategies in these areas. Also, it may be that economic priorities do not emphasize environmental interventions. Historical gaps in data collection and reporting infrastructure also contribute to the absence of information. The lack of data on soil defence interventions is crucial as it underscores potential gaps in environmental monitoring and the need for improved documentation and support

Variable	Description	Format
CUP	Unique Project Code	Character
Intervention	Title and description of the Intervention	Character
Finance Total	Total amount of investments financed for the Intervention	Numeric
COD_REGION	ISTAT Region Code	Character
DEN_REGION	Region Denomination	Character
COD_PROVINCE	ISTAT Province Code	Character
DEN_PROVINCE	Province Denomination	Character
COD_MUNICIPALITY	ISTAT Municipality Code	Character
DEN_MUNICIPALITY	Municipality Denomination	Character
FeasibilityStudyStartingDate	Start date of the Feasibility Study Phase	YYYY-MM-DD
FeasibilityStudyEndingDate	End date of the Feasibility Study Phase	YYYY-MM-DD
PreliminaryDesignStartingDate	Start date of the Preliminary Design Phase	YYYY-MM-DD
PreliminaryDesignEndingDate	End date of the Preliminary Design Phase	YYYY-MM-DD
DefinitiveDesignStartingDate	Start date of the Definitive Design Phase	YYYY-MM-DD
DefinitiveDesignEndingDate	End date of the Definitive Design Phase	YYYY-MM-DD
ExecutiveDesignStartingDate	Start date of the Executive Design Phase	YYYY-MM-DD
ExecutiveDesignEndingDate	End date of the Executive Design Phase	YYYY-MM-DD
EffectiveDesignStartingDate	Start date of the Effective Design Phase	YYYY-MM-DD
EffectiveDesignEndingDate	End date of the Effective Design Phase	YYYY-MM-DD
WorksExecutionStartingDate	Start date of the Works Execution Phase	YYYY-MM-DD
WorksExecutionEndingDate	End date of the Works Execution Phase	YYYY-MM-DD
ConclusionStartingDate	Start date of the Conclusion Phase	YYYY-MM-DD
ConclusionEndingDate	End date of the Conclusion Phase	YYYY-MM-DD
InterventionClosed	Date of the Intervention Closing	YYYY-MM-DD
Operability	Date of Operability of the Intervention	YYYY-MM-DD
Source	Source of the data	Character
geom_C	Centroids of the Municipality	Point
geom_A	Areal shape of the Municipality	Polygon / Multi-polygon

Table 5.1: Variables of the dataset.

for these regions.

5.5 Experimental, Design, Materials and Methods

Three sources are used to retrieve the soil defense public works data: the OpenBDAP, the OpenCoesione, and the ReNDiS repositories. The OpenBDAP database is maintained by the Italian Ministry of Economics and Finance (MEF), specifically by the Ragioneria Generale dello Stato. Although it primarily houses data related to the whole Italian Public Administration, our interest specifically centers on the dataset concerning investments made for public works, with a particular emphasis on soil defense. This dataset, updated monthly, encompasses a range of information about the interventions, including their nature, project code, investment sectors, descriptions of the interventions, implementation timelines, sub-sector details, and other pertinent attributes. From a financial perspective, the dataset offers insights into the costs associated with interventions and distributes financial flows among national, European, local authorities, and private entities. This allocation of financial resources is presented at an aggregated level, distinguishing between sources such as European Fundings, National Fundings, contributions from Local Authorities, foreign countries or private entities. Moreover, the dataset covers data registered mainly from the year 2000 to the present day, showing the dates of the most meaningful steps of public works' projects. The second source is the OpenCoesione. It draws its information from Cohesion Policy data, in which EU member states commit to delivering cohesive programs and data at both national and European levels. This resource is updated bi-monthly, offering a broad view of public investments sourced from both Italian and European funding channels. Even in this case, the data collected are specific to single projects and are referenced at the municipal level revealing the various resilience measures undertaken by each municipality, along with the corresponding allocated budgets. This database offers a detailed view of the economic investments regarding all public infrastructures: it collects data from both national and European programs and provides a comprehensive breakdown of the financing allocation between national and European resources. Specifically, it delineates the total amount of financing and specifies the respective contributions from specific national and European funds. Also in this case, dates about the legislative process of public works are reported. As previously mentioned, these two repositories gather data on various types of public works, including school infrastructure, transportation systems, and more. However, our analysis exclusively concentrates on soil defense interventions. To achieve this focus, the data has been filtered using the sub-sector code provided by both datasets. Specifically, the sub-sector code "05" signifies soil defense investments, enabling us to isolate relevant interventions. Last repository is the 'Repertorio Nazionale degli Interventi per la Difesa del Suolo' (ReNDiS), provided by the Italian Institute for Environmental Protection and Research (ISPRA). It is a database entirely dedicated to public investments in soil defense. Currently, the ReNDiS presents data on the total amount of financial resources allocated to each project, the category of intervention, the project code, and key dates in the public works process based on the different phases reported by the Open BDAP repository. The primary goal of the ReNDiS platform is to create a comprehensive, systematically updated overview of the works and resources engaged in soil defense. While the MEF provides a more detailed and comprehensive database as the primary source, it's crucial to highlight that the implementation of the ReNDiS database is an ongoing process. Currently, it features information sourced from the Ministry of the Environment, and efforts are underway to incorporate details regarding interventions financed through additional channels, including Regional Laws, ordinances, and other funding sources. Furthermore, data from the three different sources are collected using two different methods: for the OpenCoesione and the OpenBDAP Application Programming Interface (API) have been used to process and download the data, while the ReNDiS database is shared in a LinkedOpenData platform, thus it can be retrieved through SPARQL queries. Particularly, in the context of data retrieval, APIs provide structured access to external databases, enabling the extraction of specific data. Specific API requests were crafted to fetch relevant data [De, 2023]. These requests included parameters to filter the data regarding public investments for each region. On the other hand, SPARQL is a query language and protocol used for accessing and manipulating data stored in Resource Description Framework (RDF) format [Prudhommeaux, 2008]. SPARQL is particularly useful for querying Linked Open Data (LOD), where data entities are interlinked and accessible on the web. Indeed, ISPRA is implementing its LOD platform sharing all the data it collects. In this particular case, the ReNDiS database is accessible via a SPARQL endpoint (`https://dati.isprambiente.it/sparql`), which acts as an interface for executing SPARQL queries. The two retrieving methods were implemented in the RStudio IDE for the R software environment [R Core Team et al., 2020], where also the data processing was carried out. Once the three databases are collected the main part is to merge them into a unique database without any duplication. For this purpose, each project is uniquely identified by a project code known as the Codice Unico di Progetto (CUP). The CUP system has played a pivotal role in monitoring and consolidating information related to public investments, as mandated by the Sistema di Monitoraggio degli Investimenti Pubblici (MIP), established under Law 144 in 1999. Furthermore, the adoption of the CUP system has been in effect since January 16, 2003, through Law No. 3. The utilization of the CUP has served as a mechanism to test data duplication across the three datasets, ensuring the absence of redundant information. This systematic approach guarantees the integrity and accuracy of the dataset by confirming that no duplicate entries exist, thereby enhancing the reliability of the collected project data. As introduced before, the three datasets also report key dates regarding the legislative process of public works that in Italy follows four different phases: Planning, Implementation, Conclusion, Operability. The Open Coesione database provides a more granular breakdown of project phases, furnishing both effective and expected dates and categorizing them into six distinct sub-phases. These sub-phases include the feasibility study, preliminary design, definitive design, executive design, execution of works, and testing. Each sub-phase is accompanied by its corresponding start and end dates. In contrast, the OpenBDAP platform consolidates these phases into the four macro phases, consistently presenting both expected and effective dates, along with their respective start and end dates. The ReNDiS database draws its information from OpenBDAP, resulting in identical date records. To streamline the dataset and adopt a pragmatic approach, only effective dates are reported. The OpenBDAP's metadata file elucidates the derivation of these dates, facilitating the merging of datasets, particularly it helps reduce the OpenCoesione dates to the four phases present in both OpenBDAP and ReNDiS. Specifically, the effective start date for the design phase is determined as the minimum effective date among the start dates of the feasibility study, preliminary design, definitive design, and executive design. Conversely, the end date aligns with the maximum effective date among the end dates of these aforementioned phases. For the execution phase, the effective start date is derived as the minimum among the start date of works execution, while the end date considers the maximum end date of the phase. The start date of the conclusion phase is determined by the minimum start date among the testing and closing intervention phases, applying a similar approach to establish the end date for this phase. Lastly, the start date of the operability phase corresponds to the minimum effective start date of the operability phase. Consequently, the OpenCoesione database's date information has been condensed by identifying the minimum or maximum of start or end dates, ensuring alignment with the dates reported in the OpenBDAP and ReNDiS databases. Moreover, the financial information from the three datasets is different. OpenCoesione shows a very detailed partitioning among European and Italian resources, showing different funds for both levels. European funds refer to:

- European Regional Development Fund (ERDF)
- European Social Fund (ESF)
- European Agricultural Fund for Rural Development (EAFRD)
- European Maritime and Fisheries Fund (EMFF)
- Youth Employment Initiative (YEI)

On the other hand, the Italian funds are:

- Fondo di Rotazione nazionale
- Fondo Sviluppo e Coesione (FSC)
- Piano di Azione per la Coesione (PAC)

Additionally, the dataset also differentiates between investments originating from the Italian Government, Regions, Provinces, and Municipalities. The OpenBDAP dataset also reports different fund allocations but in a more aggregated way. It distinguishes the European Resources and National Resources, which are given respectively by the sum of the European funds and National funds. Also, it presents the category of the Local authorities' funds, given by the sum of the Regions, Provinces, and Municipalities financing. In contrast, the ReNDiS database does not have any differentiation of funds implemented, but it reports only the total amount financed for each intervention. For this reason, the database presents only one financial variable representing the total amount of financing per intervention. In the case of the OpenBDAP and the OpenCoesione datasets this information is derived as the sum of all the financial values reported in each fund's resources. To validate the accuracy of this sum, we meticulously scrutinized duplicated interventions and confirmed that the total amount reported by the ReNDiS database aligns precisely with the combined financial resources reported in the other two datasets. A key and innovative point of the creation of this dataset is that the three databases have been enriched with geographical information. The OpenCoesione and the ReNDiS datasets present region, province and municipality codes and denominations derived from the official source of the Italian National Statistics Institute (ISTAT). On the contrary, the OpenBDAP dataset does not present this information in the financial dataset, but the same data platform provides these codes and denominations associated to the CUP of each project in a different database. The latter has been downloaded and the geographical information added to the financial dataset. Thanks to this transformation for the three datasets two different geographical reference systems have been implemented: the coordinates of the centroids of the municipality of each intervention and the areal data outlining the polygonal shape of each municipality. These data options are sourced from the ISPRA database through queries, providing flexibility in accessing the geographical information of municipalities. In the ReNDiS database, georeferencing is seamlessly integrated during the data download via queries, leveraging the existing geo-references within the database. In the Open Coesione and OpenBDAP datasets, centroids or polygons are linked through province and municipal codes.

5.6 Limitations

The dataset provides insights into soil defence investments in Italy, although it presents certain limitations. Firstly, it may not capture all soil defence initiatives carried out in Italy, leading to gaps in information and under-representation of certain regions or types of projects. Indeed, its reliance on publicly available sources may introduce biases in the data, particularly if certain projects or regions are under-represented or omitted from the dataset. This is particularly relevant for the initial phase of the data retrieval, lasting up to around the early 2000's. The dataset's temporal coverage restrains the analysis of long-term trends and the assessment of the sustainability of soil defence efforts. Furthermore, it may lack detailed information on contextual factors such as socio-economic, or environmental influences, which could impact the depth of analysis and understanding of soil defence dynamics. Additionally, while efforts were made to ensure data accuracy and completeness, reliance on administrative records and self-reporting mechanisms may introduce errors or inconsistencies. Therefore, users should exercise caution when interpreting the findings and consider conducting sensitivity analyses to assess the robustness of the results. Despite limitations, the dataset remains a valuable resource for researchers and policymakers, but cautious interpretation and additional contextual research are advised.

Chapter 6

Italian investments for soil defence: retrieving and visualizing data by the PublicWorksFinanceIT R Package[1]

6.1 Paper Summary

The paper presents the development and application of the `PublicWorksFinanceIT R` package, designed to retrieve, harmonise, visualise, and pre-process data on public investments in soil-defence in Italy. Public investments in soil defence infrastructure constitute a key component of the human dimension of coastal and marine ecosystems. They reflect how societies respond to environmental risks—with funding, planning and construction of works designed to mitigate hazards such as floods, erosion and sea-level rise. While such investments are critical, the availability of harmonised, georeferenced, multi-source financial data is often limited. Existing open data repositories are heterogeneous in format, geospatial referencing and accessibility. In this context, the paper Ricciotti and Pollice [2024a] contributes by creating

[1]Ricciotti, L., & Pollice, A. (2025). Italian investments for soil defence: retrieving and visualizing data by the PublicWorksFinanceIT R Package. Annals of Operations Research, 1-28.

a tool that not only integrates data from multiple open platforms but provides user-friendly functions for retrieval, merging and mapping, thus lowering the barrier to statistical learning and spatial analysis of infrastructure investments. The package implements functions allowing the user to: download data from each source, optionally filtering by different levels of geographical disaggregation, date, geo-reference, and thematic tag; merge the datasets across the three sources using the unique project code (CUP), avoiding duplication; enrich the data with municipality-level geographic references: centroids or polygon geometries, region/province/municipality codes and names; visualise the data via mapping functions and exploratory tools built in R. Through the package researchers and policymakers can gain access to standardized, up-to-date, spatially referenced investment data without deep custom coding or data-wrangling from multiple platforms. The package is specifically designed for soil-defence investments (though it can handle broader public works). The spatial granularity at municipality level allows for detailed geographic analysis of investment flows and patterns of funding across Italy's territories (regions, provinces, municipalities). The package and dataset facilitate the subsequent analytical chapters: the ability to retrieve and visualise investments allows to integrate investment data with environmental drivers, perform spatial/time-series modelling of fund allocation, and explore associations between environmental risk and spending.

More in general, this work lowers the barrier for researchers and policymakers to access and analyse infrastructure investment data. The integration of spatial data with financial information supports spatial-statistical modelling of investment patterns, allows for linkage with hazard/exposure data, and supports policy evaluation. In the coastal and marine ecosystem context, the dataset enables evaluation of how mitigation investments are distributed across territories exposed to sea-level rise, storm surge, erosion and other coastal hazards. In summary, this paper presents a practical contribution in the context of the thesis. By packaging multi-source investment data in a georeferenced, accessible, manipulable format, it lays the groundwork for the remainder of the research on coastal defence data analysis.

6.2 Introduction

In recent years, there has been an exponential growth in the abundance and diversity of public data accessible on the web. This growth can be attributed to several factors, primarily the widespread digitization of information, coupled with a growing demand for open data. Also, there is an escalating expectation for both public and private entities to make their data available to a wider audience. Additionally, public institutions are increasingly dedicated to the dissemination and coherence of public data. For instance, the European Union introduced the *Digital Decade Policy Programme* (DDPP), which came into effect on January 9, 2023 [Commission, 2003]. The DDPP, in addition to advancing digital services and bolstering cybersecurity, aims to foster the utilization of artificial intelligence, cloud computing, and big data by businesses. However, it is common to find that data pertaining to the same topic are not consistently organized. Data sources may originate from different platforms, presenting information in varying formats, with distinct variables, or organized in disparate structures. This can potentially impede the progress of researchers, who may need to invest considerable time in retrieving and harmonizing the data before commencing their research. For this reason, as stated also by Lahti [2018], there is the need to create tools to facilitate the access to harmonized data through open source software, thus enhancing accessibility and usability for a wider range of users.

The availability of open data spans across a wide array of research fields, encompassing ecology, finance, health, environment, and more. However, this accessibility does not always translate into positive outcomes, as underscored by Zuiderwijk et al. [2012], who identified "118 socio-technical impediments for the use of open data". These impediments include the limited attention given to users and the potential value that can be derived from these data sources. Furthermore, Reichman et al. [2011] shed light on another issue pertaining specifically to ecological data: its dispersion. They emphasized that ecological data is challenging to discover and preserve due to relatively small datasets often handled by single researchers. An illustrative instance of streamlining the data cleaning process through data retrieval is provided by the `webchem` R package [Szöcs et al., 2020]. This package proves invaluable for researchers working with chemical datasets, especially in tasks involving the association of names or codes with individual chemical elements. Another example of a useful package in the environmental field is the `ARPALData` R package [Maranzano and Algieri, 2023, 2024] which has been developed to retrieve air quality and weather data collected in the Lombardy region (Italy) by the local environmental protection agency (ARPA Lombardia). The package helps users to obtain and visualize clean data and to address some problems such as the identification of missing values patterns. See also Tassan Mazzocco et al. [2023], Carslaw and Ropkins [2012], Grange [2019] for more instances. Also, the `eurostat` R package [Lahti et al., 2017] allows to retrieve data from the Eurostat database and helps users to download, manipulate and visualize this data. The latter's functionalities have been lately improved thanks to the `iotables` R package [Antal, 2025] which helps users in organizing extensive data dividing them into different tables. Taking a financial standpoint, Burdick et al. [2015] devised a novel system, Midas, capable of handling intricate data processing to extract clear and reliable financial information from the US public sources. They also encountered difficulties in developing their system due to the heterogeneity present in the open data landscape. Moreover, the R package `bcdata` [Teucher et al., 2021] enhances the efficient retrieval of geospatial data and metadata from the British Columbia Data Catalogue's public platform. This package creates a smooth linkage between the publicly accessible metadata and datasets in the British Columbia Data Catalogue, integrating with the mapping, modeling, and data processing capabilities inherent in the R software. The R package `PublicWorksFinanceIT` [Ricciotti and Pollice, 2024a] is intended to provide users with a user-friendly tool to retrieve financial data from different sources that collect data regarding Italian expenditures on public works founded by both national and European funds. The data is provided at the municipal level, meaning it offers very detailed, granular spatial information. It can be filtered by region, province, or municipality. To clarify, based on the Nomenclature of Territorial Units for Statistics (NUTS), provided by Eurostat, regions are classified as NUTS 2, and provinces as NUTS 3. Municipalities represent the individual administrative units within each province. This package, available for the `R` software environment, was specifically developed to collect public projects related to soil defence, with the primary goal of monitoring the safeguard against extreme environmental events. Financial data regarding environmental issues have been widely used to assess weather policies and actions taken have resulted in the desired effects or not. For example, Seidl et al. [2021] indicate a favorable trend in national public biodiversity investments encompassing 30 countries, emphasizing correlations between biodiversity investments and positive indicators such as threatened species, and protected areas. Moreover, Prahl et al. [2018] focus on the economic evaluation of storm surges and sea-level rise impacts in coastal cities taking into consideration the costs of dike protection. Notably, the study underscores the importance of striking a balance between spatial detail and universality in data. Jonkman et al. [2013] provide an updated means to evaluate the costs associated with coastal defences, encompassing structures like dikes and nourishing. This underscores the perpetual necessity for enhancing and revising financial information related to the funding and expenses associated with these essential public works. In this context, the `PublicWorksFinanceIT` package becomes particularly relevant as it is specifically designed to gather information on the costs associated with public works dedicated to soil safeguarding with geographic reference for all the interventions specific to each municipality.

The paper is structured as follows: Section 6.3 introduces and elucidates public works and soil defense projects. In Section 6.4, we delineate the three distinct sources from which data are gathered. Moving forward to Section 6.5, a comprehensive overview of the package and its functionalities is provided, together with illustrative examples for application. Section 6.6 showcases two compelling case studies, offering empirical demonstrations of the package's efficacy. Concluding reflections are presented in the final section.

6.3 Public Works

Public works are defined as infrastructure projects and services that are typically financed and undertaken by government authorities to serve the collective needs of the society. These include a diverse range of facilities, such as transportation systems, water and sewage treatment facilities, energy infrastructure, public buildings, and other essential structures. These endeavors are designed to enhance the overall well-being of the community by fostering economic development, ensuring public safety, and improving quality of life. Particularly, in Italy, public projects follow a specific legislative process which can be summarized in four phases: planning, implementation, conclusion, and operability. These stages are mandated to be publicly disclosed and accessible to all, following Article 38 of Law No. 33 in 2013. The are three main sources where data regarding Italian Public Administration are stored: the OpenCoesione, the OpenBDAP, and the ReNDiS websites. These three sources are used to retrieve the data about public works in Italy. Data are collected at the municipal level and reveal the various resilience measures undertaken by each municipality, along with the corresponding allocated budgets.

The three databases have been enriched with geographical references, a functionality integrated into the package. Users can choose to retrieve either the coordinates of the centroids of each municipality or the areal data outlining the polygon of the boundaries of each municipality. These data options are sourced from the ISPRA database through queries, providing flexibility and ensuring constant updating in accessing the geographical information on Italian municipalities.

6.3.1 Public Works for Soil defence

Recent discussions and analyses focus on the impacts of global warming and climate change on the environment and humanity. While addressing these issues is vital, less attention has been given to the financial consequences of this global crisis. International and national agreements aim to allocate sufficient financial resources to tackle the problem. One of the first efforts to address climate change funding was the *Investment and Financial Flows to Address Climate Change* by the United Nations, which analyzed investments and future projections [United Nations Framework Convention on Climate Change (UNFCCC), 2007]. A report at COP28 in 2022 revealed a worrying trend: while 200 billion was invested in environmental protection, funding for harmful activities was over 30 times higher [Programme, 2023]. Soil defense and public works are key to sustainable development, focusing on strengthening infrastructure and ecosystems against erosion, landslides, and flooding. For example, coastal regions are notably susceptible to the impacts of climate change and anthropogic pressures. Nicholls et al. [2007], highlight that according to the Intergovernmental Panel on Climate Change (IPCC) the hazards associated with human development in coastal regions are considerable, indeed approximately 120 million people face risks from tropical cyclones annually.

In addition, climate change is causing rising sea levels and more frequent coastal storms, threatening coastal areas. As a result, its impact extends to the structural integrity of coastal defense systems, leading to increased costs for building new structures or upgrading existing ones. Moreover, according to Masselink and Russell [2013] two main consequences of climate change have an impact on coasts: the sea-level rise and the wave climate. The former does not have impacts on the coasts immediately but implies changes over time. Additionally, human activities are highly concentrated along coastlines, exerting a significant influence on the configuration of seaside areas. For instance, according to the ISPRA 2023 annual report [Buscemi et al., 2024], approximately 30% of the Italian population is concentrated along the coastline, encompassing an area of 43,000 km^2, equivalent to about 13% of the national territory. Mitigating coastal erosion involves implementing various coastal protection measures, such as groins, barriers, beach nourishments, sand ripping, and reconstruction of artificial dunes. However, these works imply large investments as well as maintenance costs. Indeed, projects related to soil defence often involve engineering solutions, such as retaining walls, drainage systems, and vegetation management, to safeguard against the adverse effects of extreme events. Specifically in Italy, the issue of hydro-geological instability is particularly relevant as it affects a significant portion of the peninsula, causing impacts on the population, and the economy. This is due to the characteristics of the morphological conformation of the Italian territory. Italy, moreover, is a highly anthropized country with nearly 8000 municipalities, and a population density of around 200 inhabitants/km^2 [Gallozzi et al., 2020] where a significant amount of urbanized area has been built often in absence of a proper territorial planning, consequently leaving villages and cities vulnerable to heightened risks of landslides and floods. In addition, infrastructure is a key element of the United Nations Sustainable Development Goal (SDG) 9, but its impact on sustainable development extends across all SDGs. According to Thacker et al. [2019], infrastructure affects 72% of SDG targets, influencing 121 out of 169. Notably, for five SDGs (3, 6, 7, 9, and 11), every target is shaped by infrastructure, with water and energy sectors having the most significant impact on SDG 6 (clean water and sanitation) and SDG 7 (affordable and clean energy).

6.4 Available Data

As mentioned before the data are retrieved from three different sources which are the OpenCoesione, the OpenBDAP, and the ReNDiS platforms. In general, while the ReNDiS database provides only information on soil defence financing, the OpenCoesione and the OpenBDAP databases report the same information regarding all public works and differ for a few variables.

Both the OpenCoesione and the OpenBDAP datasets present the following information:

1. Project Local Code,
2. Project Unique Code (CUP),
3. Project description,
4. Nature of the Intervention: Code and Description,
5. Intervention Type: Code and Description,
6. Intervention Sector: Code and Description,
7. Intervention Sub-sector: Code and Description,
8. Intervention Category: Code and Description,
9. Effective and Expected Date of the project's phases,
10. European Funds,
11. Total budgetary savings.

The codes of the nature, type, sector, sub-sector, and category of the intervention are chosen during the procedure of releasing the CUP code and they can be used as filters, allowing the users to narrow the dataset based on their specific area of interest. More specifically, below are described all the types of interventions that can be requested from the two datasets.
The first step when requesting the CUP code is to identify and declare the type of operation which is classified by two layers: the nature and the type of intervention. The first level concerns the nature of the intervention, it presents eight levels (Table 6.1):

The second layer concerns the type of intervention and it is based on the nature code (Table C.1). First, the investment's nature must be identified among the eight different types. Later, the type of intervention must be identified. This layer is interconnected with the first step. For instance, if a project is identified as

Table 6.1: Nature of Intervention Codes

Code	**Nature of Intervention**
01	Purchase of Goods
02	Realization or Purchase of Services
03	**Realization of Public Works**
06	Granting of Aids Entities Other Than Productive Units
07	Granting of Incentives to Production Units
08	Purchase of Equity Investments and Capital Injections

'Purchase of Goods' (code 01), later one must choose among the three different types related to the chosen nature (ichosen .e. new furniture (code 00), extraordinary maintenance (08), other (99)).

Furthermore, projects can be assigned to a category which is defined by three layers: the sector, the sub-sector, and the category of the intervention. This categorization is independent of the previous one regarding the project's nature. Each project must be assigned to a specific sector. The first step is to identify the relevant sector. Once the sector is selected, one can then choose a corresponding sub-sector, followed by a specific category within that sub-sector. For instance, a project may be assigned to the "Transport Infrastructures" sector (code 01). From there, a sub-sector such as "Road Infrastructure" (code 01) can be selected, and finally, a category like "highways" (code 011) can be chosen. To display the codes associated with each layer, two tables are provided in the Appendix: the first table illustrates the sectors along with their respective sub-sectors (Table C.2).

The OpenCoesione and the OpenBDAP datasets present other variables regarding both some descriptive features and some financial indicators of the project. Descriptive variables may encompass details related to the procedural aspects of the public works, the domain of its implementation, the entities entrusted with its execution, and the location where the project is undertaken. Conversely, financial variables delineate the monetary resources required or expended in the actualization of the project.

6.4.1 OpenBDAP dataset

The OpenBDAP database is maintained by the Italian Ministry of Economy and Finance, specifically by the *Ragioneria Generale dello Stato*. While it encompasses data related to the whole Italian Public Administration, for the purposes of this paper and package we focus on the data on investments made for public works. In particular, it offers 48 variables with details on the use of public funds in Italy across various sectors. Data regarding each project are identified by a unique project code known as the *Codice Unico di Progetto* (CUP). The CUP system has played a crucial role in monitoring and aggregating information pertaining to public investments in Italy, as mandated by the *Sistema di Monitoraggio degli Investimenti Pubblici* (MIP), established under Law no. 144 in 1999. Moreover, the adoption of the CUP system has been in effect since January 16, 2003, through Law No. 3. The OpenBDAP database encompasses a range of information about the interventions, including their nature, investment sectors, descriptions of the interventions, implementation timelines, sub-sector details, and other pertinent attributes. From a financial standpoint, it provides details on the projected and effective costs of interventions and allocates the financial flows among national, and local authorities, and private entities. Finally, the dataset also includes information on budgetary savings, which are defined as the difference between the projected costs of interventions and the actual expenditures incurred for their implementation. Apart from the 11 variables presented above which are in common with the OpenCoesione dataset, the OpenBDAP data has 37 more variables describing each public intervention, including:

1. CUP Status: Code and Description (can be "A" for "open" or "C" for "closed"),
2. CUP validity: Starting and Ending Date,
3. Holder Description,
4. Holder Fiscal Code,
5. Institution: Code and Description,
6. Estimated and Effective Costs,
7. Expected and Effective amounts available,
8. Expected and Effective Burdens on Investments,
9. Governmental Funds,
10. Local Authorities Funds,
11. Private Funds,
12. Other Source of Funds,
13. Raising Funds.

Originally, this dataset does not have any geographical reference, but georeferencing has been added thanks to a dataset available in the OpenBDAP platform which reports the Unique Code of each project associated with each region, province, and municipality. Thanks to that it has been possible to associate each project with the coordinates of the centroids and/or the areal data regarding the polygonal boundaries of each municipality.

6.4.2 OpenCoesione dataset

The OpenCoesione website draws its information from Cohesion Policy data, in which EU member states commit to delivering more cohesive programs and data at both national and European levels. This resource is updated bi-monthly, offering a broader view of public investments sourced from both Italian and European funding channels. This database offers a more detailed view of the economic investments regarding public infrastructures: it collects data from both national and European programs and provides a comprehensive breakdown of the financing allocation between national and European resources. Specifically, it delineates the total amount of financing and it specifies the respective contributions from both national and European funds. Additionally, these datasets also differentiate between investments originating from the Italian Government, Regions, Provinces, and Municipalities.

The OpenCoesione dataset presents 199 variables in total. Besides the 11 variables in common with the OpenBDAP dataset already mentioned, it presents other variables that are summarized in Table 6.2:

Table 6.2: Project Variable Categories and Descriptions

Category	**Description**
Programming	Variables that refer to the programming of the interventions.
Territory	Variables related to the region, province, and municipality of each project.
Finance	Detailed features regarding the financial aspects of the project.
Timeline	Additional time references, mainly concerning the bid's procedure.
Progress	Indicators concerning the progress of the intervention from both financial and procedural perspectives.
Subjects	Variables that qualify the subjects involved in the project.
Indicators	Refer to the progress of the project.
Flag	Variables related to the affiliation of the project.

The OpenCoesione dataset is also divided according to thematic projects. Specifically thematic projects refer to 11 different sectors of investments:

1. Research and Innovation,
2. Digital Services Networks,
3. Business Competitiveness,
4. Energy,
5. Environment,
6. Culture and Tourism,
7. Transports,
8. Employment,
9. Social Inclusion and Health,
10. Education and Training,
11. Administrative Capacity.

The OpenCoesione dataset is more complex and complete than the OpenBDAB dataset, but the two datasets do not report the same interventions. Merging both datasets proves useful to obtain more informative data on public works in Italy.

6.4.3 ReNDiS dataset

The *Repertorio Nazionale degli Interventi per la Difesa del Suolo* (National Repertory of Interventions in Soil Defence), ReNDiS provided by ISPRA (the Italian Institute for Environmental Protection and Research), serves as a database specifically dedicated to public investments in soil defence. Currently, the ReNDiS database presents data on the total amount of resources allocated to each project, the category of intervention, and the CUP (Codice Unico Progetto). The primary goal of the ReNDiS platform is to create a comprehensive, systematically updated overview of the works and resources engaged in soil defence. This collective information is intended for sharing among all administrations involved in planning and executing interventions. In this context, ReNDiS aspires to serve as an informative tool, enhancing coordination and, consequently, optimizing national expenditures for soil defence. Furthermore, it aims to promote transparency and facilitate citizens' access to relevant information. In 2015, the ReNDiS platform underwent integration with a specific area designated for the Regions. This area serves to local authorities as the platform for submitting funding requests related to hydro-geological risk mitigation interventions supported by the Ministry of the Environment and the Energy Security. While the Ministry of Economy and Finance (MEF) provides a more detailed and comprehensive database as the primary source, it's crucial to highlight that the implementation of the ReNDiS database is an ongoing process. Currently, it features information sourced from the Ministry of the Environment, and efforts are underway to incorporate details regarding interventions financed through additional channels, including Regional Laws, ordinances, and other funding sources. The Repertorio Nazionale degli Interventi per la Difesa del Suolo (ReNDiS) database, provided by ISPRA, assumes a pivotal role in systematically cataloging public investments directed towards soil defence. By sourcing its information from the OpenBDAP (MEF) database, ReNDiS furnishes users with valuable insights, encompassing details on the total resources allocated to each project, the specific category of intervention, and the corresponding CUP. While drawing upon the comprehensive MEF database as its primary source, ReNDiS distinguishes itself through its continuous implementation, underscoring a commitment to perpetual improvement and the seamless integration of the latest data.

Notably, the categories of interventions within ReNDiS are thoughtfully classified as showed in Table 6.3.

Table 6.3: Intervention's categories

Intervention's categories
Landslide
Not Defined
Flood
Mixed
Avalanche
Fire
Coastal

These categories provide a detailed framework for understanding and categorizing diverse interventions related to soil defence. It's important to note that the ReNDiS database is in continuous implementation, reflecting a commitment to ongoing enhancement and the incorporation of the latest data.

6.5 The PublicWorksFinanceIT package

The `PublicWorksFinanceIT` package contains 11 functions which can be classified as:

- functions for data/information retrieval,
- functions for data visualization.

The download functions facilitate users in obtaining data from all three sources. There are three distinct functions, each tailored to a specific source. For the OpenCoesione and the OpenBDAP platforms APIs have been used to download data, while for the ReNDiS database SPARQL queries have been built. Additionally, two functions are provided to acquire data on the more detailed OpenCoesione data, allowing users to retrieve data based on one or more regions or on specific project themes of interest. Visualization functions are for both areal and point data. Table C.3 gives an overview of the available functions.

6.5.1 Data Retrieving functions

The function `get_data_OBDAP()` is used to download data from the OpenBDAB platform. This function has seven arguments that can be used to filter the data: `cod_reg`, `cod_prov`, `cod_mun`, `start`, `end`, `geo_ref`, `soil_defence`, where "reg" stands for region, "prov" for province, and "mun" for municipality. It is possible to retrieve data for one or more regions listing the region codes provided by the Italian National Statistical Institute (ISTAT) using the argument `cod_reg`. For example, to retrieve data for both Valle d'Aosta and Basilicata regions, which respectively have the codes 02 and 17, the function to call is
`get_data_OBDAP(cod_reg = c("02","17"))` and a unique dataset will be saved in the `R` environment collecting the observations for both regions. Similarly, users can achieve the same outcome using the `cod_prov` or `cod_mun` arguments to specify codes for provinces or municipalities of interest. To help users get the codes of Italian regions, provinces and municipalities the `get_codes()` function displays the names and the codes of each of them. The function has the argument `type` used to specify which information is needed: it can be set equal to `"region", "province"` or `"municipality"` according to the information needed. In addition, data can

be filtered based on the Effective starting and/or ending date of the design of the project through the use of the arguments `start` and `end`. The Effective starting/ending date of the design of the project is taken as a reference time variable as it coincides with the starting/ending date of one of the first phases of the project. Moreover, the `geo_ref` parameter can take one of two values: `"C"` or `"A"`. When using the `"C"` argument, the user requests data with the coordinates of the centroids of municipalities. Conversely, selecting the `"A"` argument retrieves areal data, where each municipality is represented by its polygonal boundary. The data on centroids and polygons for each municipality are sourced from the ISPRA database, directly linked to ISTAT. This ensures that information about each municipality is consistently updated. Finally, the `soil_defence` argument is set by default to `FALSE`, implying that the requested data encompasses all fields and sectors. However, by changing the argument to `TRUE`, users have the option to specifically request data related to soil defence public works, thereby reducing the time required for downloading the data.
To retrieve data from the OpenCoesione website, the users can utilize either the `get_data_region_OC()` or the `get_data_theme_OC()` functions, depending on whether they intend to download data by region or based on thematic projects.
These functions vary by just one argument: for regional data, the users need to specify the regional code using the `cod_reg` parameter, whereas, for retrieving data based on project themes, they should specify the code for the desired theme using the `themes` parameter. To access information on the codes for both regions and themes, a supplementary function is available. It presents the codes associated with each region or theme the users intend to retrieve. To gather regional codes, utilize the `get_info_OC("region")` function, and for theme codes, employ the `get_info_OC("theme")` function. These codes become instrumental when utilizing the functions to retrieve data, whether by region or project theme. For example, to acquire data concerning the Lazio region, the user would invoke `get_data_region_OC(cod_reg = "LAZ")`. Conversely, if the users aim to filter data based on themes, the `get_data_theme_OC` function is employed. In this case, the `themes` argument specifies the theme of interest. For instance, to retrieve data exclusively for projects related to the research and innovation theme, use `get_data_theme_OC(themes = "RICERCA_INNOVAZIONE")`. Both functions share

common arguments: `cod_reg`, `cod_prov`, `cod_mun`, `start`, `end`, `geo_ref`, `soil_defence`. The `start` and `end` parameters allow the users to specify the effective starting and/or ending date of the project, enabling them to filter projects within a desired period. For instance, if the users call
`get_data_region_OC("VDA", start = "20220101", end = "20230331")` they are retrieving data about public works for the Valle d'Aosta region with a starting date on the 1^{st} of January 2022 and the ending date on the 31^{st} of March 2023. Furthermore, the parameters, `cod_prov` and `cod_mun` enable data filtering based on province and municipality codes.
An example is: `get_data_region_OC("VDA", cod_mun = "007002")`, where only data regarding the Antey-Saint-André (007002) municipality are retrieved. Additionally, as for the previous function, the `geo_ref` argument can be used to request geo-referenced data. As for the `get_data_OBDAP` function the `soil_defence` parameter is initially set to `FALSE` by default. However, should the user choose to set it to `TRUE`, the function will retrieve data specifically related to soil defence.

Last, the function `get_data_RENDIS()` allows the user to retrieve data for the ReNDiS database. This function presents seven different arguments: `cod_reg`, `cod_prov`, `cod_mun`, `start`, `end`, `geo_ref` and `type`. The first six arguments are similar to those in the previous functions. The `type` argument allows users to filter data based on different types of interventions in the field of soil defence. To view the available types of interventions, a complementary function, `get_type_RENDIS()`, is provided, displaying the various intervention types. For instance, if a user wishes to obtain data for the Valle d'Aosta region, specifically focusing on interventions related to landslides, the following call can be made: `get_data_RENDIS(cod_reg = "02", type = "Frana", geo_ref= "C")`.

6.5.2 Merging the data

In our research, the integration of diverse datasets of public infrastructure projects is essential for a comprehensive analysis. To facilitate this process, a versatile function has been implemented that enables users to merge the three distinct datasets efficiently. The merge is performed by concatenating the three datasets row-wise, ensuring that all available observations are retained. This approach facilitates a comprehensive dataset that captures the full spectrum of available information.

After the initial merge, any duplicate observations are systematically removed, retaining only those with the most complete information to enhance the dataset's quality and reliability. This function is called `merge_data` and it produces a unified dataset with 32 pivotal variables, including the unique project identifier (`CUP`), a detailed description of the intervention (`Intervention`), the total financial investment (`Finance`), different sources of financial inflows (`State Funding`, `EU Funding`, `Local Authorities Funding`, `Private Funding`, and `Other Funding`), region, province, and municipality names (`DEN_REGION`, `DEN_PROVINCE` and `DEN_MUNICIPALITY`), region, province, and municipality codes (`COD_REGION`, `COD_PROVINCE` and `COD_MUNICIPALITY`), key dates relevant to the legal iter of the project, the geographical geometry (`geom`), and the source (`Source`) of each observation. This amalgamation of key information provides researchers with a consolidated and enriched dataset, facilitating a more nuanced exploration of public infrastructure projects.

6.5.3 Visualization functions

The package provides three versatile functions designed for visualizing both areal and point data: `plot_funds_map()`, `plot_funds_point()`, and `plot_funds_bar()`.

The `plot_funds_map()` function caters to areal datasets and involves specifying two essential arguments: `data` and `var`. These arguments respectively denote the dataset and the variable the user intends to visualize on the map. Relying on `ggplot2` package's functions [Wickham, 2016], this feature-rich tool creates a compelling visualization of areal data on a map. Each municipality's region is filled based on the amount of expenditures it sustains, with the option to transform the amount using a logarithmic scale for enhanced visualization. This function applies to datasets obtained from all three platforms. Below is an illustrative example featuring both the code and the resulting visualization. Figure 6.1 shows the Sardinia's map colored based on the logarithmic scale of the Total Funding variable. The latter represents the total amount of funding received from both public and private investments. It is possible to see that the majority of funds are allocated along the coastline.

The `plot_funds_points()` function is another versatile tool designed for datasets obtained from all three platforms when users request the coordinates of the centroids of each municipality. Similar to the previously mentioned function, it aims to

```
dati_SAR <- get_data_region_OC("SAR", geo_ref = "A")
plot_funds_map(data = dati_SAR, var = "TotalFunding")
```

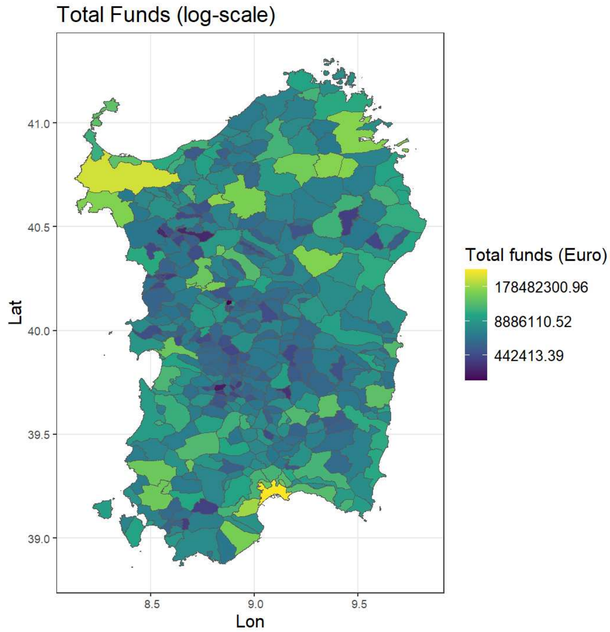


Figure 6.1: Map of Sardinian Municipalities colored by the log-scale of the total amount of public and private fundings. Data displayed are retrieved from the OpenCoesione database.

visualize expenditure data, but it processes point data using the `leaflet` package [Cheng et al., 2019]. This function generates a map displaying circles with varying radii and colors based on the desired expenditure amount in a logarithm scale for enhanced data insight. Additionally, the map includes pop-ups that reveal both the municipality name and the corresponding expenditure amount. Below is an illustrative example featuring both the code and the resulting visualization. Figure 6.2 shows the point map of the Basilicata region. Circles are colored based on the log scale of the variable TotalPublicFunding, which represents the sum of the public funding received. It is clear that the two main cities, Matera and Potenza, received the majority of the funds.

```
dati_BAS <- get_data_region_OC("BAS", geo_ref = "C")
plot_funds_points(data = dati_BAS, var = "TotalPublicFunding")
```

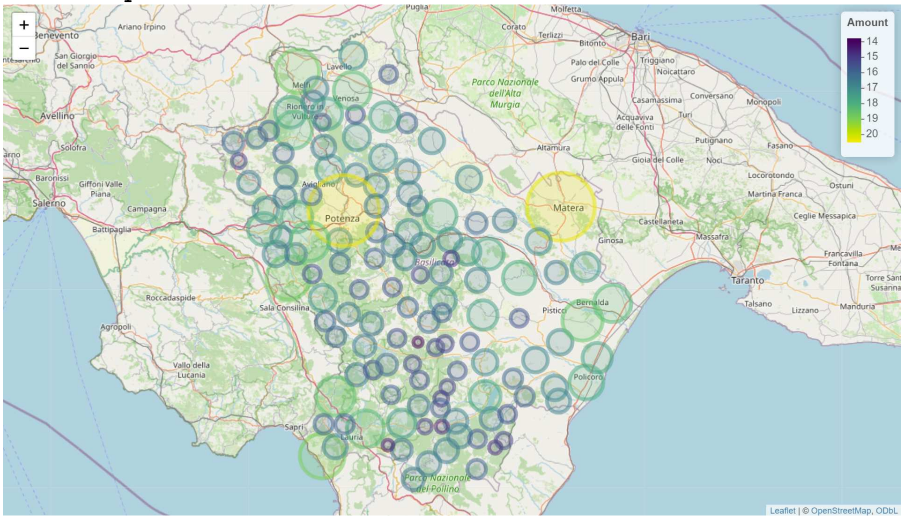


Figure 6.2: Point map of Basilicata: Municipalities' Centroids Color-Coded by the total public funds received.

The `plot_funds_bar()` function serves as the final piece for facilitating visualizations. With this function, users can generate a bar plot illustrating the funding distribution across various regions for different financial resources. It accommodates both areal and point data, offering flexibility for diverse datasets. The function requires two key arguments: `data`, allowing users to specify the dataset, and `var_col`, enabling the selection of columns of interest for visualization. Below two examples of barplots (Figure 6.3 and 6.4).

```
dati_02 <- get_data_OBDAP("02")
plot_funds_bar(data = dati_02, var_col = 11:15)
```

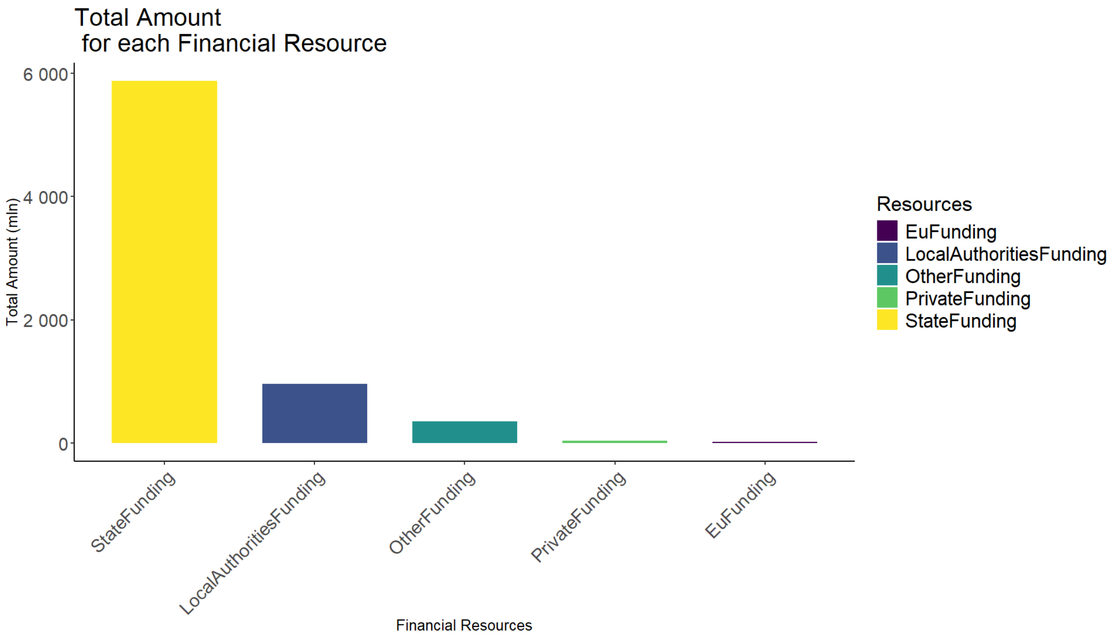


Figure 6.3: Barplot of the partition of financial expenditure for the Valle d'Aosta region using the OpenBDAP dataset. This example shows the allocation among various European funds.

```
dati_VDA <- get_data_region_OC("VDA")
plot_funds_bar(data = dati_VDA, var_col = 10:15)
```

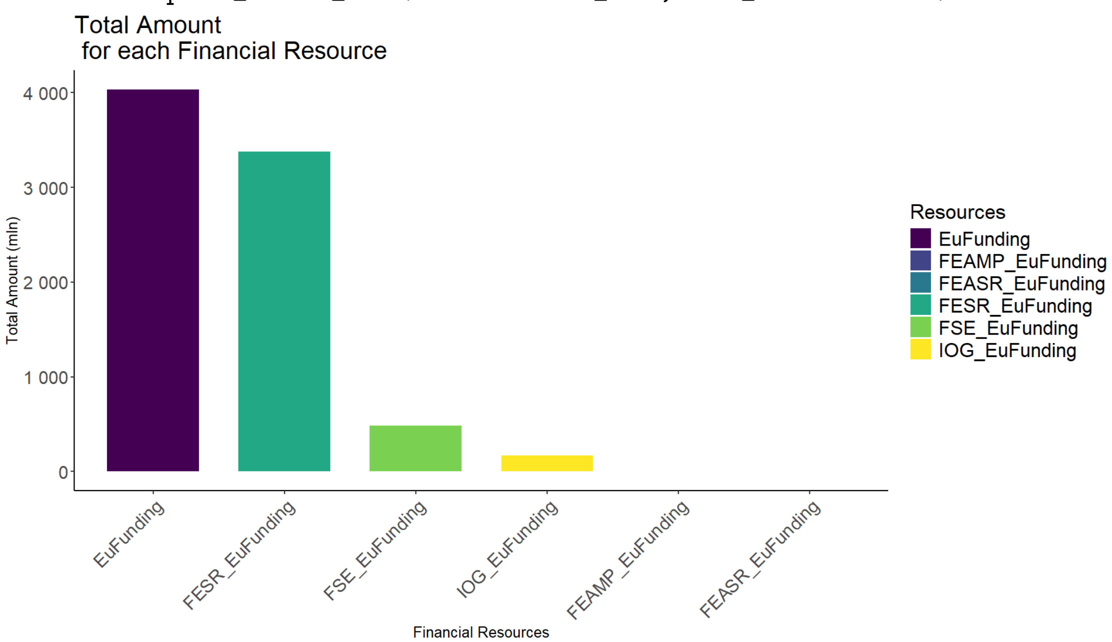


Figure 6.4: Barplot of the partition of financial expenditure for the Valle d'Aosta region using the OpenCoesione dataset. This example shows the allocation among various European funds.

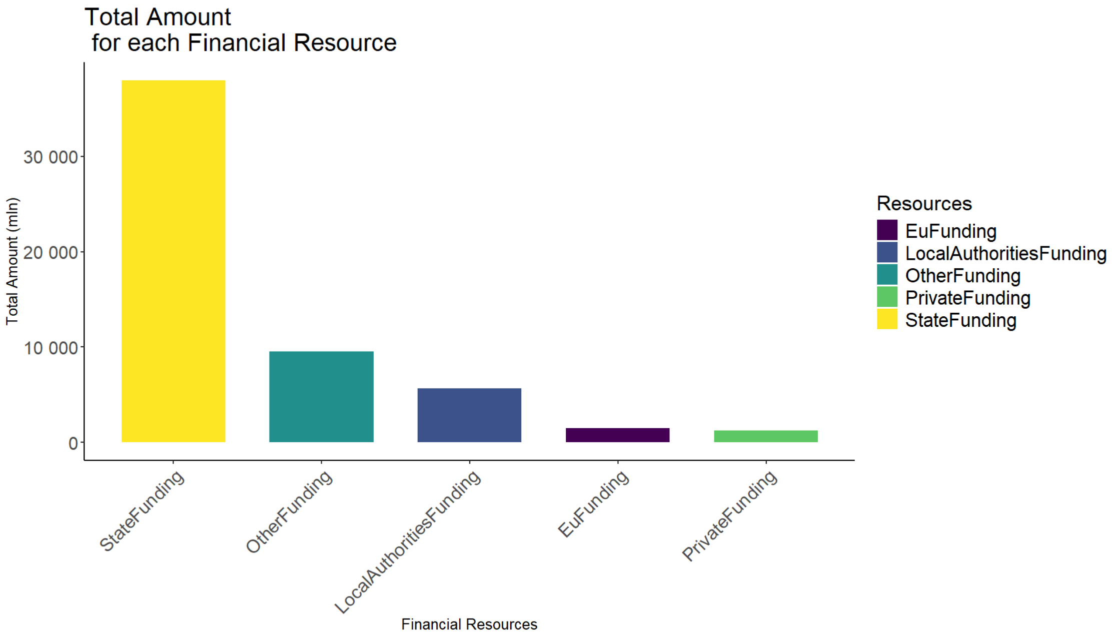
6.6 Case Studies

To provide a more comprehensive understanding of the package, this section introduces two distinct case studies. These examples predominantly center around the exploration of datasets through exploratory data analysis (EDA) tools. The first case study aims to provide a comprehensive overview of financial investments in a particular region, with a focus on various types of public works. The goal of this EDA is to uncover patterns, trends, and insights related to the allocation of funds in the specified area. The second case study narrows its focus to a more detailed examination of soil defence initiatives in multiple regions. This involves a deeper dive into the specific strategies employed by different regions to protect their soil, along with an analysis of the corresponding investments.

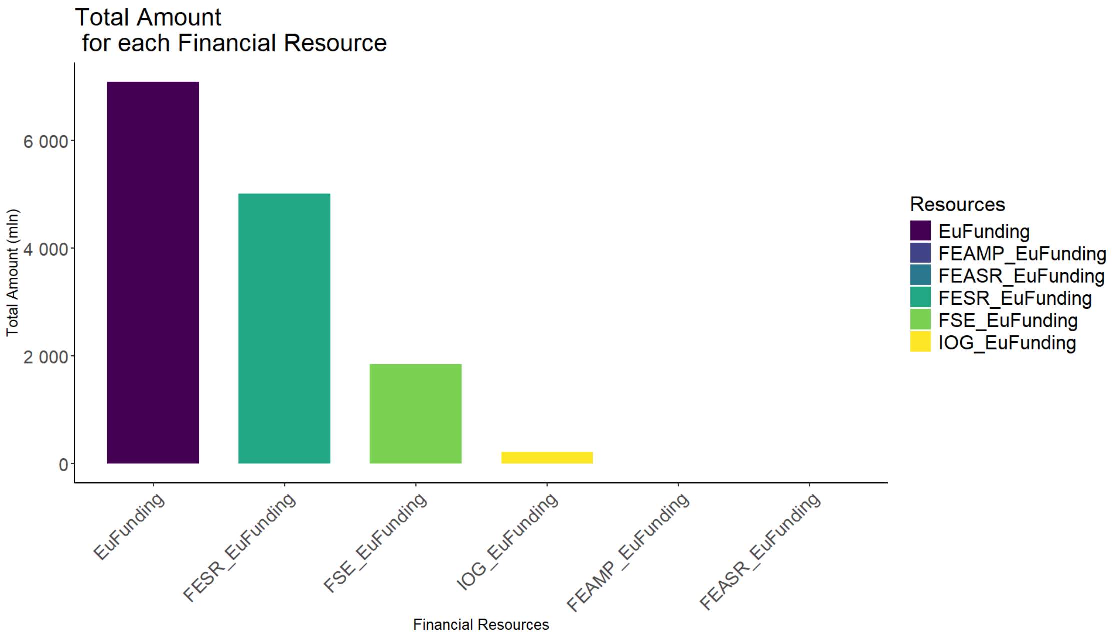
6.6.1 Case Study 1: Financial Allocation Analysis in the Lazio Region

This example explores the financial inflows regarding the Lazio region. Our datasets encompass 74,170 observations from the OpenCoesione dataset, 20,411 from the OpenBDAP dataset, and 1,589 from the ReNDiS platform. Through the merge function (`merge_data()`), we consolidate a unique dataset comprising 76,972 observations, with 20,272 originating from OpenBDAP, 55,776 from the OpenCoesione database, and 924 retrieved from the ReNDiS database. Most observations appeared across all datasets, leading to duplicate entries. As a result, merging the three datasets did not create a database containing unique observations from each. This process is handled within the merging function. This integration provides a comprehensive overview of publicly financed infrastructure projects in the Lazio region, as each dataset contributes valuable insights that complement one another. Visualizing functions have been applied to the datasets. For example the `barplot_funds()` function has been applied to both OpenCoesione and OpenBDAP dataset. As illustrated in Figure 6.5a, the predominant source of funds is identified as the Italian Government, surpassing all other funding channels. This observation underscores a substantial commitment by the government to allocate significant financial resources to the Lazio region. On the other hand, the OpenCoesione database provides insights into the utilization of specific European funds for investments in public works. As depicted in Figure 6.5b, it becomes apparent that the primary European funding sources for the Lazio region include the European Social Fund (FSE) and the European Regional Development Fund (FESR) towards financing projects within the region.

Figure 6.5: Bar-plots showing the varying amounts of financing provided by different sources to the Lazio region.



(a) Bar plot illustrating the various sources of financial inflows extracted from the OpenBDAP dataset for the Lazio region.



(b) Bar-plot illustrating the allocation of funds from European sources within the OpenCoesione dataset for the Lazio region.

A comprehensive understanding of the financial landscape supporting the Lazio region is achievable through the creation of a consolidated bar-plot utilizing the merged dataset. This approach provides a holistic view of the financial scenario, as depicted in Figure 6.6. The Italian government still contributes more to the financing of public works in the Lazio region

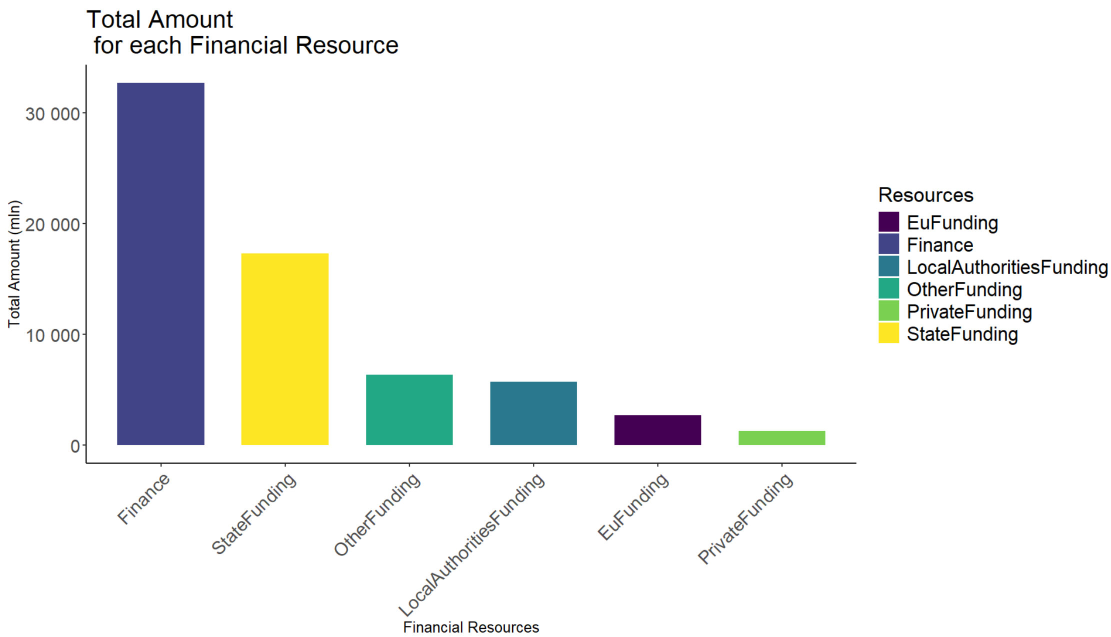


Figure 6.6: Bar-plot visualizing the different sources of financial inflows for the Lazio region, coming from the complete dataset.

Moreover, it is possible to visualize two point maps, which allow us to characterise the disparity among the territory of the region. The map in Figure 6.7 provides a visual representation of the varying financial allocations across the Lazio Region, sourced from the ReNDiS database and thus regarding only soil defence works. The municipality that has received the majority of funds is Rome, as indicated by both the color and size of the circle, with the latter being proportional to the total financing amount for the municipality. Furthermore, considering that the ReNDiS database retrieves data regarding soil defence interventions, the analysis reveals that primary funds are directed towards municipalities along the coastline or nearby, while hinterland municipalities appear to receive comparatively lesser allocations. This observation aligns with the demographic distribution in the Lazio Region, where the provinces boasting higher populations, specifically Rome and Latina [ISTAT, 2020], tend to attract a more substantial share of funding.

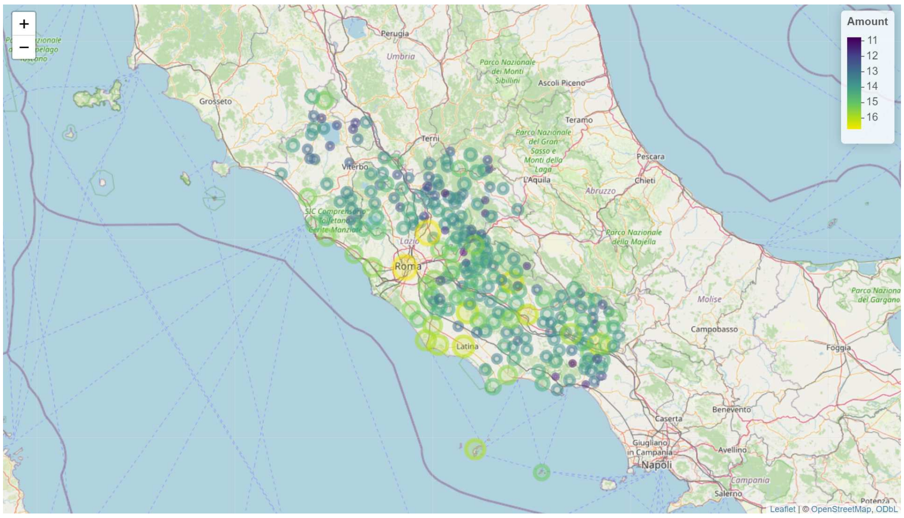


Figure 6.7: Point map illustrating the diverse financial allocations for each municipality in the Lazio region, sourced from the ReNDiS database.

It is possible to extend the visualization of similar maps to the other two data sources, encompassing all categories of interventions. For instance, Figure 6.8 illustrates mapped points representing financial allocations made by Local Authorities in each municipality. Notably, this map displays a more heterogeneous distribution, beyond the municipality of Rome, suggesting that in the Lazio region, Local Authorities invest relatively evenly across their respective territories.

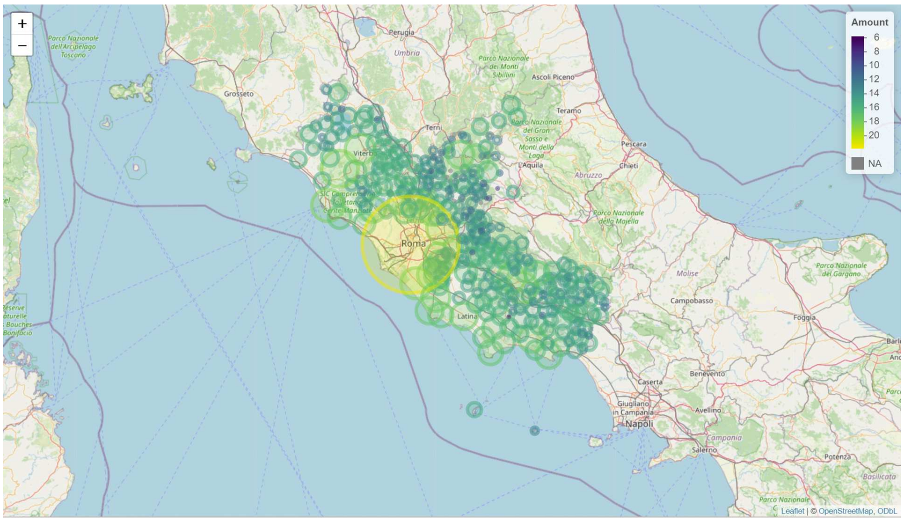


Figure 6.8: Point map depicting financial allocations sourced from Local Authorities within the Lazio region, as extracted from the OpenBDAP database.

6.6.2 Case Study 2: Soil defence Financial Allocations in the Arco Ionico Regions

This case study delves into soil defence public works, with a specific emphasis on the Calabria, Basilicata, and Apulia regions. These three regions together form the "Arco Ionico Region", a name derived from their shared coastline along the Ionian Sea. To analyze the financial aspects of soil defence infrastructures, relevant data can be extracted from the OpenCoesione and OpenBDAP platforms. This can be achieved by specifying the argument `soil_defence = T` within the functions, ensuring that only data related to these specific types of projects is retrieved. It's worth noting that the ReNDiS database exclusively offers information on soil defence public works. The data, acquired through the package functions, comprises 53,044 entries from the Ministry of the Economy and Finance database (OpenBDAP), 187,231 observations from OpenCoesione, and 2,987 from the ReNDiS platform, covering all three regions. By employing the `merge_data()` function, a comprehensive dataset with a total of 223,264 observations is generated. Each of the three datasets contributes valuable information to the overall dataset: 51,492 entries from OpenBDAP, 169,760 from OpenCoesione, and 2,012 observations from ReNDiS.

To comprehensively assess the investment in soil defence infrastructures across the three regions, a visual representation becomes invaluable. For this purpose, areal data is essential, prompting the configuration of the `geo_ref` argument to `"A"`. Figure 6.9, crafted through the utilization of the `plot_funds_map()` function on the merged dataset, not only offers a clear overview of the financial landscape but also facilitates a nuanced understanding of how these regions prioritize and allocate resources for soil defence initiatives. The color-coded map vividly illustrates the varying amounts of total finance received by each municipality, providing insights into the distribution and magnitude of investments in soil defence infrastructure across the regions. White spaces represent municipalities where we do not register any funding. The presence of missing data highlights a broader challenge in the data sources. Missing values can be attributed to incompleteness in the original records. Data regarding public works have recently undergone strict monitoring, which may explain incomplete records. Further, some interventions can be registered by local authorities themselves on platforms such as ReNDiS, although this practice is still not mandatory. These factors must be considered when interpreting the observed trends.

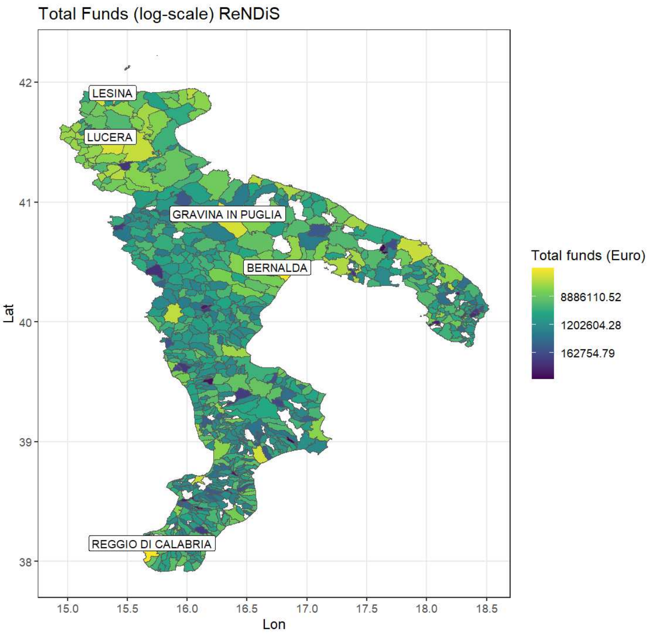


Figure 6.9: Geographic map displaying each municipality in the three regions, color-coded to represent the corresponding amount of total finance received. The five municipalities highlighted received the highest financing amount.

Upon examining the map, a distinct pattern emerges, revealing that a significant portion of the investments is concentrated in the Apulia region, characterized by a prevalent green and yellow color palette. Indeed, three of the five municipalities with the higher financing amount are located in the Apulia region, specifically the municipalities of Lesina, Lucera, and Gravina in Puglia. The former is part of the Gargano National Park, a UNESCO site. A closer inspection further unveils a noteworthy trend: the primary focus of these investments aligns with the larger municipalities, typically serving as the administrative capitals of their respective regions. For instance, the municipality of Reggio di Calabria stands out as the sole municipality in yellow, registering the highest amount of financing not only in the Calabria region but also across all three regions. Additionally, another noteworthy municipality registering significant investment include Bernalda in Basilicata, which presents a long coast line (about 6km). This municipality is extremely exposed to hydro-geological risk, in particular to coastal erosion events. This spatial distribution not only highlights the substantial commitment to soil defence in Apulia but also underscores a strategic approach wherein major urban centers, often representing the regional epicenters, receive a substantial share of the financial allocations. This nuanced observation adds depth to our understanding, shedding light on the regional prioritization and the strategic considerations influencing investment patterns in soil defence infrastructure. Furthermore, employing the `plot_funds_bar()` function allows for a deeper insight into the primary sources contributing to the financial inflows (Figure 6.10). Significantly, upon examining the total financial investments, it becomes evident that the leading investor across these three regions is the category known as "Other Funding." This category encompasses the cumulative funds originating from Free Resources, Foreign State, Other Public, and Other Foreign Subjects. Following closely behind are the funds provided by the Italian Government. However, it is essential to highlight that the ReNDiS database, while providing valuable total investment figures, does not differentiate among the various channels of financing. This caveat underscores the importance of acknowledging the potential nuances in funding sources and understanding that a comprehensive assessment may require additional data or alternative sources to delineate the specific origins of financial support for soil defence initiatives in these regions.

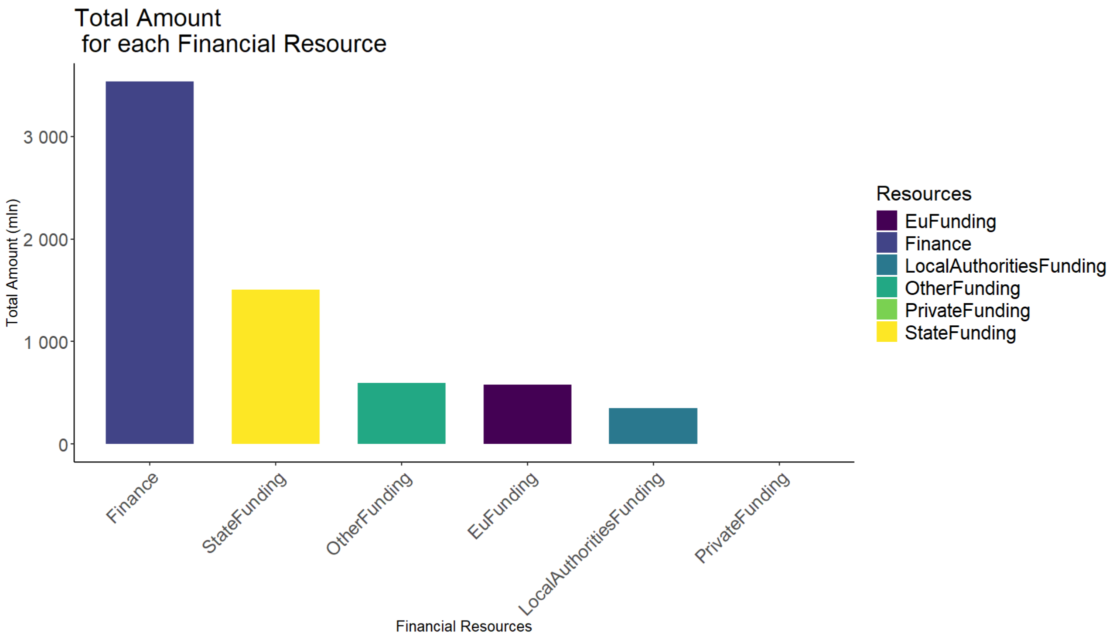


Figure 6.10: Barplot illustrating the sources of financial investments in the three regions, with the Italian Government prominently featured as the primary contributor.

6.7 Conclusion

In conclusion, the `PublicWorksFinanceIT` R package proves to be valuable in enabling users to access a standardized dataset encompassing financial information on public works throughout Italy. Specifically designed to capture and organize data related to soil defence projects, the package plays a crucial role in addressing the environmental challenges associated with soil preservation and safeguarding. However, the analysis also underscores the limitations of the original data sources. Missing values, which can arise from the incompleteness of records in institutional databases, highlight the need for improved reporting practices. Recognizing and addressing these challenges is crucial for ensuring that the dataset remains comprehensive and reliable for future analyses. The significance of the package extends beyond mere data retrieval, as it facilitates a more standardized and coherent approach to handling financial information. By offering a consolidated view of financial data related to public works, especially in the domain of soil defence, the package streamlines the analytical process. This enhancement in data accessibility and coherence sets the stage for more in-depth and comprehensive analyses. The final potential users of the `PublicWorksFinanceIT` R package encompass a diverse range of stakeholders engaged in public works, environmental preservation, and decision-making processes. This includes government agencies, policymakers, environmental researchers, and analysts who can benefit from the standardized dataset to derive meaningful insights. By providing a user-friendly interface and specific functionalities tailored to soil defence projects, the package caters to the needs of professionals seeking a specialized tool for their research and decision-support endeavors. In essence, the `PublicWorksFinanceIT` R package not only contributes to the efficiency of data retrieval and analysis but also serves as a catalyst for informed decision-making in the dynamic landscape of public works. Its user-friendly design, coupled with a focus on soil defence initiatives, positions the package as a valuable asset for those dedicated to enhancing environmental sustainability and making data-driven decisions in the realm of public infrastructure development.

Chapter 7

Cost-to-Coast: Space-time Modeling of Coast Defense Funds and Environmental Drivers in Italy[1]

7.1 Paper Summary

This final chapter represents the culmination of the thesis, where the focus broadens from the physical and biological components of the marine environment to the socio-economic and policy dimensions of coastal resilience. The paper investigates how environmental pressures and social factors influence the allocation of public investments for soil and coastal defense across Italy's coastal municipalities. It offers a novel integration of environmental, demographic, and financial data within a Bayesian spatio-temporal framework, shedding light on the economic geography of environmental adaptation in a country highly exposed to coastal hazards.

Despite the large volume of public spending devoted to soil defense and risk mitigation, a systematic understanding of how such funds are distributed, and whether their allocation aligns with local risk and exposure has remained limited. The study addresses this gap by merging heterogeneous administrative datasets with environmental indicators and modelling their joint spatio-temporal patterns

[1]Ricciotti L., Pollice A., Picone M., Martino S. Cost-to-Coast: Space-time Modeling of Coast Defense Funds and Environmental Drivers in Italy. Environmental and Ecological Statistics. *Accepted for publication.*

using state-of-the-art Bayesian methods.

A suite of model specifications was tested, varying in covariate inclusion and lag structures to reflect possible delays between environmental events and funding decisions. To capture delayed response, environmental covariates were lagged by two years. The spatial-temporal model provides a nuanced picture of how Italy finances its coastal protection. Maps of posterior means and uncertainty reveal higher predicted funding levels in the northwestern and central-western coasts, areas more urbanized and exposed to storm surges, while southern regions receive comparatively less. Despite this unevenness, the overall trend suggests a policy landscape that responds adaptively to environmental stressors. From a policy standpoint, the findings underscore the need for more proactive and risk-sensitive allocation mechanisms. The observed two-year lag between environmental pressure and funding response highlights bureaucratic inertia and suggests opportunities for improving adaptive planning cycles. Moreover, the results call for a more equitable spatial distribution of resources, prioritizing municipalities where environmental hazards intersect with social vulnerability. Such integration aligns with the European Union's Cohesion Policy and the national climate adaptation strategies promoting balanced territorial development and resilience.

Within the overall structure of the thesis, this chapter plays a strategic concluding role. It operationalizes the statistical learning tools developed in previous chapters on a socio-economic domain, thereby linking environmental science with public policy analysis. It extends the data-integration framework introduced in Chapter 5 and leverages the `R` infrastructure described in Chapter 6 to conduct a full-scale applied analysis. Conceptually, this work situates environmental risk not only as a physical phenomenon but as a determinant of financial behaviour and governance. By quantifying the interplay between ecological drivers and fiscal responses, it enriches the thesis's overall narrative: that statistical learning can illuminate the coupled human–environment systems shaping coastal and marine sustainability.

In conclusion, this study demonstrates how spatial and temporal statistical models can bridge the gap between environmental data and policy evaluation. The integration of open administrative databases with high-resolution environmental indicators provides a reproducible blueprint for analysing adaptation finance at national and local levels. Beyond Italy, this approach offers a transferable framework for studying how governments respond financially to environmental risks in coastal zones worldwide. As the final contribution of this thesis, it emphasizes the power of data integration and Bayesian modelling in supporting transparent, adaptive, and scientifically informed environmental governance, thereby linking, in a coherent framework, the progression from sea-state modelling to the quantitative assessment of socio-economic resilience.

7.2 Introduction

In recent years, climate change has increasingly threatened coastal areas worldwide, not only through environmental impacts but also by imposing significant financial and policy challenges. Despite extensive research on environmental impacts, less attention has been paid to how public resources are allocated to mitigate these risks. International and national agreements, such as the United Nations' *Investment and Financial Flows to Address Climate Change* [United Nations Framework Convention on Climate Change (UNFCCC), 2007], aim to allocate sufficient funding for adaptation. However, estimating how these financial flows are distributed locally, and understanding the factors that shape allocation decisions, remain major challenges. This study investigates the spatial and temporal distribution of public funds for soil defense along the Italian coastline, focusing on how environmental pressures and socio-economic factors influence allocation patterns. The novelty of this work lies in combining municipal-level funding data with environmental and social indicators within a Bayesian spatio-temporal modeling framework, providing insights into patterns that have not been analyzed before. Building on this, our analysis focuses on identifying statistical associations between the financial flows allocated for soil defense infrastructures along the Italian coastline and a set of environmental and social factors. Instead, the attempt to establish causal relationships with an econometric perspective goes beyond the scope of the present work. Financial data are retrieved from three different public sources: OpenCoesione, OpenBDAP, and ReNDiS. OpenCoesione is based on Cohesion Policy data, reflecting EU member states' commitment to more cohesive programs and data integration at both national and European levels. It provides a broad perspective on public investments funded by both Italian and European sources. Conversely, OpenBDAP, managed by the Italian Ministry of Economy and Finance, specifically by the Ragioneria Generale dello Stato, covers financial data for the entire Italian Public Administration, including investments in public works. Finally, ReNDiS, maintained by the Italian Institute for Environmental Protection and Research (ISPRA), is a national repository dedicated to public investments for soil defense. For this analysis, we selected data on soil defense interventions for the entire Italian peninsula using the `R` package `PublicWorksFinanceIT` [Ricciotti and Pollice, 2024b]. This package was specifically developed to facilitate the downloading and merging of data from the three repositories while incorporating spatial information through two alternative georeferencing systems.

This study aims to bridge an existing gap by analyzing how public funds for soil defense infrastructures in Italy are distributed across space and time. To this aim, we employ spatiotemporal models within a Bayesian framework, using the Integrated Nested Laplace Approximation (INLA) approach, a computationally efficient method for performing approximate Bayesian inference in latent Gaussian models [Rue et al., 2009, 2017, Bakka et al., 2018]. This method is particularly useful for spatial and spatiotemporal analyses, as it allows for accurate estimation of random effects and precision parameters while avoiding the computational burden of traditional Markov Chain Monte Carlo (MCMC) methods. Indeed, INLA provides faster computations and ease of use, making it useful for large datasets and high-dimensional spatial models. It is well suited for ecological and environmental applications, since it enables flexible modeling of both time and space components. For these reasons, INLA has been extensively used in different research fields such as environmental studies, pollution monitoring, disease mapping, and social and economic research [Bakka et al., 2018, Williams et al., 2019, Belmont et al., 2024, Panunzi et al., 2024].

Within the INLA framework, we propose to model the spatial component using a reparameterization of the Besag-York-Molliè model (BYM) including both a structured and an independent component [Riebler et al., 2016a]. In particular, a mixing parameter is estimated to determine the relative contribution of the structured and unstructured components to the overall spatial variability. To fit the model we used the `inlabru` R package [Bachl et al., 2019].

The remainder of the paper is structured as follows: Section 7.3 places the proposal in the context of the existing recent literature on the topic; Section 7.4 introduces the data on soil defense interventions and possible drivers; Section 7.5 describes the spatiotemporal model, and Section 7.6 discusses the main empirical results; Sections 7.7 and 7.8 conclude with a synthesis of the findings and their implications in the political framework.

7.3 Literature Review

A key manifestation of climate change with profound environmental and financial consequences is sea level rise (SLR). According to Team et al. [2023], the global mean sea level increased by 0.20 m between 1901 and 2018, with human influence identified as the primary driver. This rise in the sea level poses severe risks to coastal infrastructures, ecosystems, and communities, amplifying the urgency of financing adaptation measures. Coastal areas, being highly dynamic and densely populated, are particularly vulnerable to the impacts of sea level rise, which leads to problems like flooding, storm surges, and coastal erosion. When considering coastal erosion specifically, the concept of adaptation, as noted by Masselink and Russell [2013], refers to the adjustment of natural or human systems to mitigate the effects of erosion. Additionally, Nicholls et al. [2007] emphasize that coastal regions face considerable hazards, with approximately 120 million people globally exposed to risks from tropical cyclones every year. Focusing on the Mediterranean, the challenges associated with coastal areas become even more evident. The Mediterranean is characterized by a high concentration of human activities along its coasts, intensifying the pressures on these vulnerable environments. For instance, Ropero et al. [2024] developed a flood-alert system for Andalusian Mediterranean catchments using Bayesian inference, combining weather forecasts and river-level measurements to predict floods hours in advance. In Italy, approximately 30% of the population lives in coastal areas, which cover 43,000 km^2, equivalent to about 13% of the national territory [Buscemi et al., 2024]. The combination of natural and human-induced factors, such as the increasing human pressure along the coastline (e.g. littoralisation) and tourism, makes the Italian coastline especially susceptible to erosion, flooding, and habitat loss, demanding urgent attention to adaptation and mitigation strategies tailored to this unique context. Thus, analyzing the Italian coastline is important since it reflects the broader challenges faced by densely populated coastal areas in the Mediterranean, showing the need for focused adaptation and mitigation strategies.

From a political and financial perspective, research on the effectiveness of public policies and the allocation of public funds in relation to climate change has been extensive. Ronchi et al. [2019] analyzed European policies for soil protection among the European Union (EU) member states, highlighting how the lack of a common framework reduces the effectiveness of the instruments in place. Also Köninger et al. [2022] pointed out the lack of a EU common framework and examined the impact of initiatives on soil biodiversity protection. As of today, the EU has adopted the EU Soil Strategy for 2030, which tries to safeguard soil biodiversity, ensuring sustainable ecosystems, and establishing a common framework for the protection of soils and sustainable management practices. Bisaro et al. [2020] analyzed the complexities of multilevel governance and funding arrangements for public goods, such as coastal flood risk reduction. Their study highlighted diverse fiscal and decision-making frameworks across countries and their implications for future challenges like sea-level rise. To this matter, Panagos et al. [2015] stated that water is the main factor causing soil erosion in Europe, thus they assessed the mean annual rate of soil loss caused by water. Their results showed that the highest rates were observed in the Mediterranean region, particularly Italy showed the highest rate of soil loss (8.46 t/ha). Raclot et al. [2018] also argued that the Mediterranean is a peculiar area, particularly susceptible to soil erosion by water. Still, they also conclude that this issue is mainly driven by anthropogenic pressure than by climate change. Smith et al. [2024] emphasize once again that soil erosion by water is the major factor threatening soil health and suggest best management practices to prevent it. Spatial analyses of public funds and policies have been widely explored using various methods, including Geographic Information System (GIS), regional level spatial analyses, and econometrics models (e.g., Scorza [2013], Antunes et al. [2020]). Cooper et al. [2020] explored coastal defences in the Northern Ireland coast through aerial survey photos geo-referenced in a GIS, discovering that coastal defences are threatening the natural coastal ecosystems and will lead to higher costs for sea defences. However, few studies focus on local administrative units. For instance, Kantamaneni et al. [2022] assessed an erosion coastal vulnerability index (ECVI) for the coastline of Camber in the United Kingdom using data collected from different institutions. Their study revealed that the majority of the coast suffers from erosion due to human intervention and extreme events. Punzo et al. [2022] explore soil consumption in Italy at the municipal level using a spatial econometric framework. To now, limited attention has been given to how changes in social and environmental factors influence the allocation of resources by national authorities and governments. In particular, there is a gap in assessing whether these changes lead to increased funding needs to address environmental and social challenges.

7.4 Data on Italian investments for public soil defense and their potential drivers

Data on soil defense interventions are sourced from three repositories — OpenBDAP, OpenCoesione, and ReNDiS — and aggregated at the municipal spatial level. Each intervention is linked to spatial information based on the polygon geometry of the municipality where it was implemented. The dataset comprises 40.833 soil defense interventions across the Italian peninsula, spanning from 1984 to 2024. For each intervention, the final dataset contains the unique project code (CUP), the total amount financed, the spatial information, and the year of reference [see Ricciotti and Pollice, 2024c, for more details]. A regulatory framework governing the infrastructure implementation process was established in 2009 in Italy by the National Budget reform process. This reform introduced three key concepts: planning, monitoring, and assessment. Specifically, the monitoring process ensures the tracking of resource allocation and the evaluation of actual expenditure outcomes (Legislative Decree 229/2011). Since then, valuable sources of information are available, suggesting to focus the present work on soil defense investments spanning the years 2010 to 2023. This time-period provides a comprehensive and substantial dataset, with a reliable number of monitored public interventions for soil defense. As the focus is on coastal areas, 645 municipalities along the Italian coastline are selected and filtered for the years of interest, resulting into 9030 observations (14 years times 645 municipalities).

We normalize the total funding of each coastal municipality by multiplying it by the municipality's surface area (km^2) and dividing by the length of the coastline (km). Figure 7.1 shows the logarithm of this ratio for the entire study period. Northern Italian municipalities, especially those facing the Tyrrhenian Sea (west coast), have higher funding for soil defense compared to southern ones, aligning with the known Italian North-South divide. To explain the variation in the financial amounts used for soil defense, we considered several possible sets of covariates. Before including them in the model, we acknowledged the potential problem of

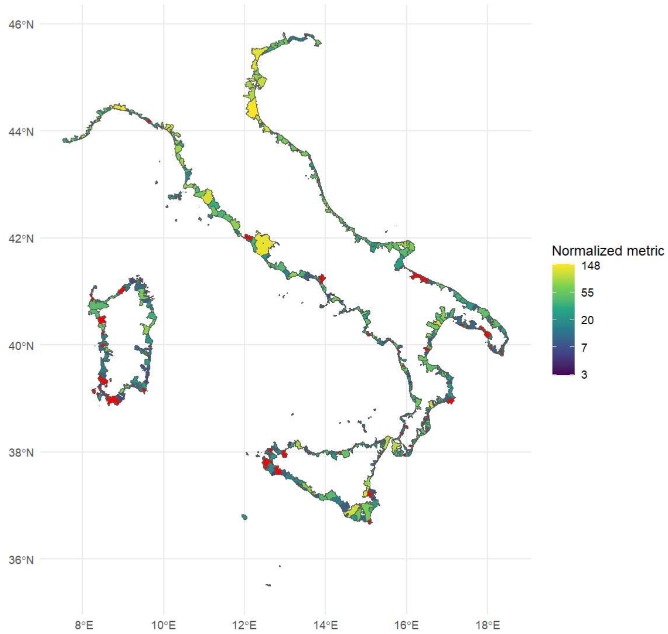


Figure 7.1: Italian coastal municipalities colored by the log-ratio defined in equation (7.1). Red polygons are municipalities with no information.

cointegration between the response and time series covariates [Engle and Granger, 1987], namely soil consumption, days of precipitation ($> 20mm$), and wave energy. To address this, we performed the Engle-Granger cointegration tests available in the `aTSA` R package [Qiu and Qiu, 2015], with no evidence supporting the presence of cointegration in our data. Therefore, the following variables were considered:

- The surface area of each municipality and the total population amount for every year. The resident population on the 1^{st} of January of every year was retrieved from the Italian National Institute of Statistics (ISTAT) website;

- Geo-morphological data related to the configuration and changes of the Italian coastline: total length of the coastline, total length of the natural coast (excluding artificial sections), changes in the coastline (erosion and advancement) of each municipality. Additionally, the coastal configuration can be classified as high coast, low coast, connection to coastal infrastructures, connection to port infrastructures, coastal defense infrastructures, riverbanks, or other (including beaches, piers, and similar structures). These data represent either constant features or changes over time from 2006 to 2020 and are publicly available by ISPRA;

- ISPRA also provides data on soil consumption per year. Data on soil consumption for each municipality between 0 and 300 meters from the coast are here considered as a proxy of the urbanization of coastal areas;

- Coastal exposure was manually assigned based on the direction each municipality's coastline faces (Figure 7.2). This variable is introduced to account for the distinct characteristics of the seas surrounding Italy, particularly the Adriatic Sea, an enclosed basin with unique features (see Ricciotti et al. [2024] for further details);

- Wave energy assessed as $E = 0.5 * H_{m0}^2 \cdot T_e$, where H_{m0} is the significant wave height, and T_e is the wave energy period. Raster data for the last two variables were obtained from the E.U. Copernicus Marine Service Information [Korres et al., 2021] for 2010 to 2023, and the wave energy was calculated. For all grid points the mean wave energy and the $90^{th}, 95^{th}$, and 99^{th} percentiles were obtained for every year. To associate wave energy values to municipalities, a 10 km circular buffer was applied, and the maximum of the data points surrounding each municipality's centroid was used.

- Precipitation data were obtained from the Copernicus Climate Change Service [Copernicus Climate Change Service (C3S), 2021]. Specifically, we extracted the total number of days within a year (from 2010 to 2022) that recorded at least 20 mm of daily precipitation from the Extreme Precipitation risk indicators for Europe and European cities dataset [Mercogliano et al., 2021]. The data is provided on a gridded reference system, with a resolution of $0.1° \times 0.1°$. To assign precipitation values to individual municipalities, we used the centroids of each municipality and matched them to the nearest grid point.

Covariates are summarized in Table 7.1.

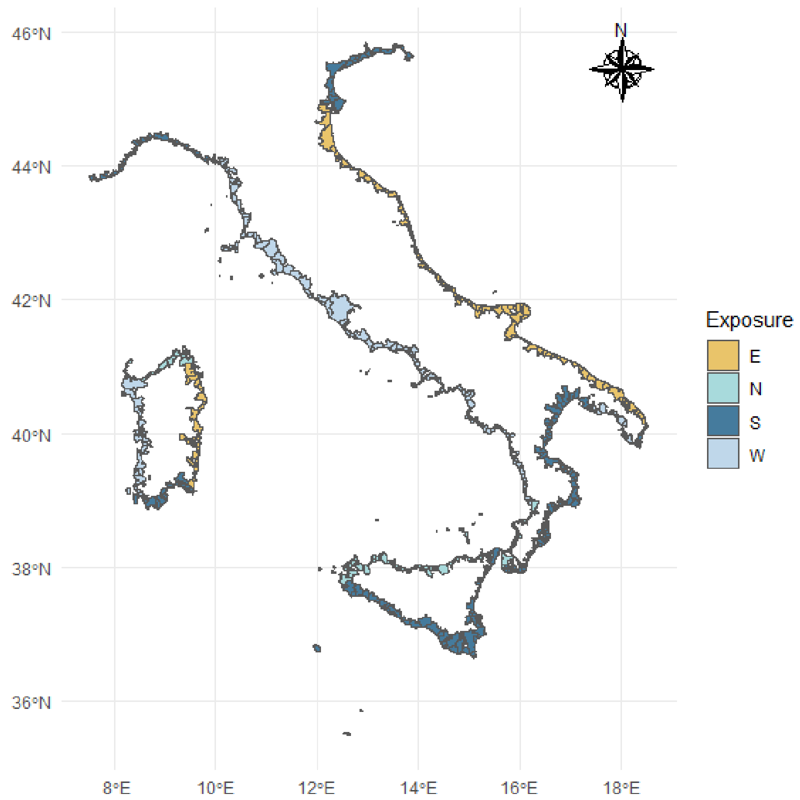


Figure 7.2: Map showing the coastal exposure of Italian municipalities.

Covariate	Measure type	Time varying
Surface area	Continuous	Fixed
Population	Continuous	Varying
Length Coast	Kilometers	Fixed
Natural Coast	Kilometers	Fixed
Coastal Erosion	Kilometers	Fixed
Coastal Advancement	Kilometers	Fixed
Coast Description	Factor	Fixed
Soil Consumption (0-300m)	Kilometers	Varying
Exposure	Factor	Fixed
Wave energy	Continuous	Varying
Precipitation ($\geq 20mm$)	Counts	Varying

Table 7.1: List of covariates used to fit the spatio-temporal model.

7.5 Spatiotemporal modeling framework

Data on fundings for soil defense interventions in Italian coastal municipalities from 2010 to 2023 are analyzed within a spatio-temporal modeling framework. Let

$F_{s,t}$ be the total amount of funding for municipality $s \in \{1,\ldots,645\}$ and year $t \in \{2011,\ldots,2022\}$, consider

$$Y_{s,t} = \log\left(\frac{F_{s,t} \cdot A_s}{C_s}\right) \tag{7.1}$$

where A_s is the municipality's surface area in km^2, and C_s is its coastal length in km. The log-ratio in Equation 7.1 provides a normalized metric that accounts for the geographical scale of each municipality by incorporating both the surface area and the coastal length. Some alternative normalized metrics accounting for e.g. the resident population amount or density were also tested but didn't lead to clear insights. We assume $Y_{s,t}$ to have a Gaussian distribution with variance σ^2 and mean $\mu_{s,t}$ defined as follows:

$$\mu_{s,t} = X_{s,t}\beta + \gamma_t + \delta_s + w_{p(s)}, \tag{7.2}$$

Our approach builds on a rich literature on spatio-temporal modeling of areal data, which provides the theoretical foundation for handling both spatial dependence and temporal dynamics (see, e.g., Sahu [2022], Haining and Li [2020]). In Equation 7.2, $X_{s,t}$ is a vector of covariates (including an intercept) defined for each time point t and municipality s, and β is a vector of unknown parameters.

To account for three distinct sources of variability in our data, we introduce three random effects, as follows:

- a time-specific random effect $\gamma_t \sim \mathcal{N}(0, \sigma_t^2)$ meant to capture year to year variability.

- a spatial-unstructured random effect[2] capturing variation at the municipal level $\delta_s \sim \mathcal{N}(0, \sigma_s^2)$. This effect is meant to capture local municipal variations that are not spatially correlated.

[2]Unstructured spatial effects capture location-specific variability that is not spatially correlated—each area's effect is independent from the others. Structured spatial effects model spatial dependence by borrowing strength from neighboring areas. In the ICAR specification used here, the structured component assumes that geographically adjacent provinces tend to have similar values, capturing the idea of spatial autocorrelation.

- a Besag-York-Molliè [BYM, Riebler et al., 2016b] spatial random effect $w_{p(s)}$ defined as:

$$w_{p(s)} = \frac{1}{\sqrt{\tau_p}}(\sqrt{1-\varphi}v_{p(s)} + \sqrt{\varphi}u_{p(s)}) \tag{7.3}$$

  Here $p(s)$ indicates the Italian province the coastal municipality s belongs to. This effect is meant to capture the spatial variability (both structured and unstructured) at the province level, which was selected as the relevant scale of analysis not only because it adequately reflects territorial heterogeneity, but also because decisions concerning the allocation of funds are typically taken at this level of administration, making it particularly relevant for policy evaluation. The BYM model consists of two elements: an intrinsic conditional autoregressive (ICAR) model $u_{p(s)}$ that accounts for spatial dependence, and an unstructured random effect $v_{p(s)}$ that accounts for the unstructured spatial variability of Italian coastal provinces.

  Notice that while the covariates are measured at the municipal level, ensuring that the fixed effects reflect local determinants, the space-structured random effects account for a broader contextual variation at the province level.

To finalize the model in a Bayesian framework we need to specify prior distributions for all precision and variance parameters. The choice of priors in Bayesian modeling is a topic of active research, particularly for hierarchical models [Gelman, 2006, de Zea Bermudez et al., 2024]. To ease prior elicitation, each of the random effects has been scaled to ensure that the geometric mean of the marginal variances equals one [Sørbye and Rue, 2014]. Moreover, to account for the fact that the vicinity graph for the ICAR model is disconnected (we have in fact three components: the mainland, Sicily and Sardinia) we have used the methodology proposed in Freni-Sterrantino et al. [2018]. In this paper, we adopted Penalized Complexity (PC) priors for the precision parameters of all random effects [Simpson et al., 2017]. These are a popular class of priors that penalize model complexity thus avoiding overfitting. Specifically, we place a PC prior on the BYM precision so that $P(1/\sqrt{\tau_p} > 1) = P(\sigma_p > 1) = 0.01$, and on the mixture parameter φ so that $P(\varphi > 0.5) = 0.5$ as recommended by Riebler et al. [2016b]. We use the same PC priors also for the variances σ_t^2 and σ_s^2. We fit the model using the INLA methodology implemented through the `inlabru` R library [Bachl et al., 2019].

We estimated multiple model specifications by systematically including different combinations of the covariates listed in Table 7.1 in $X_{s,t}$, while keeping the random components fixed. To select the best model, we used three criteria: the Watanabe–Akaike information criterion (WAIC), the Deviance information criterion (DIC) and the marginal log-likelihood (m-lik). The WAIC was introduced by Watanabe [2013] as an extension of the commonly used Akaike Information Criterion (AIC). Unlike AIC, which relies on a single point estimate, WAIC evaluates the entire posterior distribution, making it better suited for assessing Bayesian models. This allows WAIC to provide a more comprehensive measure of model performance and predictive accuracy in the Bayesian framework. The DIC is another generalization of the AIC tailored for Bayesian model selection. Unlike AIC, it incorporates Bayesian principles, making it particularly useful for hierarchical models and models with complex structures [Spiegelhalter et al., 2002, Gelman et al., 1995]. DIC provides a balance between model fit and model complexity. Models with lower DIC values are generally preferred, as they indicate better predictive performance while penalizing overfitting. Last, the marginal likelihood [or model evidence, Gómez-Rubio, 2020] is a key criterion for Bayesian model selection. It is defined as the probability of the observed data under a given model, integrating over all possible parameter values weighted by their prior distribution. Unlike information criteria such as AIC or DIC, the marginal likelihood inherently incorporates both model fit and model complexity, favoring simpler models unless the additional complexity is strongly supported by the data. As we will detail in the next section, the covariates selected by the three criteria capture key social and environmental drivers of coastal management, including the length of the natural coast, coastal description, erosion, precipitation, wave energy, coastal exposure, and soil consumption along the coastline.

7.6 Results

All the covariates listed in Table 7.1 were included in the model selection, as they all contribute to shaping the morphological conformation of the coastline and highlight relevant environmental and weather conditions. We first tested a simple linear model without covariates, followed by models incorporating covariates and various interactions among them. Additionally, we introduced a lag for environmental covariates, considering that funding decisions for protective infrastructures are typically made in response to extreme events. As discussed in Bisaro et al. [2024], European countries are facing long-term risks acting in a relatively short time, often responding with adaptation measures that exceed projected needs, such as constructing sea walls that are larger than current forecasts would warrant. In our case study, we therefore assumed a two-year lag for extreme precipitation and wave energy to reflect the temporal disconnect between environmental pressures and policy responses in public investment. Soil defense funding and implementation decisions often do not respond immediately, but after delays due to planning, budget allocation, and bureaucratic processing. While other lag durations were also tested, the results indicated that a two-year lag provided the best evidence. Here we present the three models that yielded the best performances. These were chosen to highlight the main factors contributing to public expenditures for soil defense in Italian coastal municipalities. Table 7.2 summarizes the covariates included in each of the selected models.

Table 7.2: Covariates, interaction terms, and model selection criteria across three model specifications.

Category	**Variable**	**Model 1**	**Model 2**	**Model 3**
Environmental	Natural Coast	✓	✓	✓
	Coastal Description	✓	✓	
	Erosion	✓	✓	✓
	Precipitation (2-year lag)			✓
	Wave Energy (99^{th} percentile, 2-year lag)			✓
Social	Soil Consumption	✓	✓	✓
Interactions	Wave Energy × Exposure		✓	
Model Fit	WAIC	10242.64	10240.33	**10182.68**
	DIC	10240.20	10237.31	**10182.30**
	Marginal log-likelihood	-5278.212	-5278.898	**-5282.849**

Table 7.2 also presents the results for each of the three model selection criteria, previously outlined in Section 7.5, across the three models. Model 3 exhibits the best performance in terms of model fitting. It incorporates the length of the natural coast, soil consumption along the coastline, erosion, wave energy, and precipitation with a 2 years lag.

Table 7.3 provides parameter estimates for Model 3, including covariates and spatial and temporal precision parameters. The length of the natural coast has a

Table 7.3: Estimated parameters for covariates and random effects for Model 3, including the mean, standard deviation, and 95% credible interval.

Covariates	Mean	Sd	$q_{0.025}$	$q_{0.975}$
Length Natural Coast	-0.33	0.08	-0.49	-0.17
Soil Consumption	0.18	0.05	0.08	0.27
Coastal Erosion	0.29	0.07	0.16	0.43
Wave Energy (99^{th} percentile)	0.29	0.14	0.03	0.57
Precipitation ($> 20mm$)	0.02	0.08	-0.14	0.18
	Mean	**Sd**	**$q_{0.025}$**	**$q_{0.975}$**
τ_y	0.27	0.01	0.25	0.28
τ_t	0.05	0.01	0.03	0.08
τ_{p_s}	2.45	0.79	1.29	4.36
φ_{p_s}	0.18	0.18	0.01	0.67
τ_s	0.93	0.11	0.73	1.18

negative effect on the response variable, which aligns with expectations: the greater the extent of natural coastline, the lower the level of artificial surface, reducing the need for soil defense funding. Soil consumption within 0-300 meters of the coastline exhibits a positive effect, indicating that higher levels of urbanization near the coast correspond to greater financial investments. This finding is intuitive, as protecting developed areas is a priority to safeguard infrastructure and populations. Coastal erosion also has a positive coefficient, suggesting that areas experiencing greater erosion receive increased funding, likely as a response to the need for mitigation efforts. The final two covariates yield particularly insightful results. The wave energy (99th percentile) shows positive effects. This implies that past extreme wave energy events contribute to current funding decisions. In contrast, the number of days of extreme precipitation (two-years lag, days exceeding 20 mm) does not have a significant effect, indicating that recent extreme rainfall days may not directly associate with funding decisions at the municipal or provincial level. Figure 7.3 shows the estimates of the unstructured year effects (τ_t). It exhibits a wide confidence band, indicating considerable temporal heterogeneity. This suggests that external factors influencing funding allocation may vary significantly from year to year, potentially due to economic cycles, policy shifts, or sudden environmental disasters.

At the spatial level, the precision of provincial effects (τ_{p_s}) suggests moderate

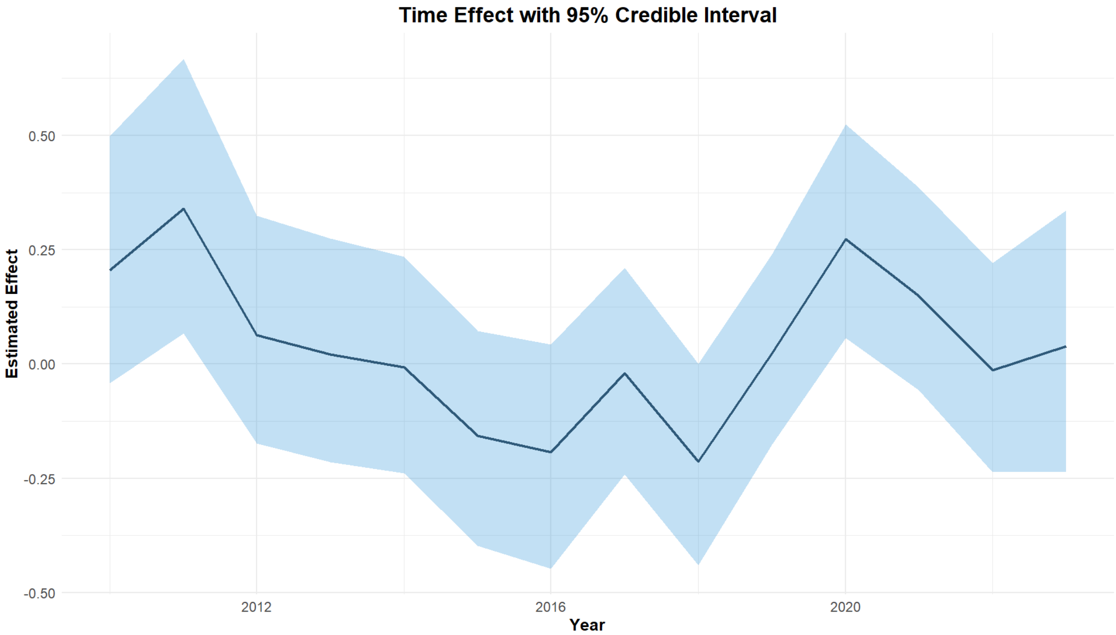
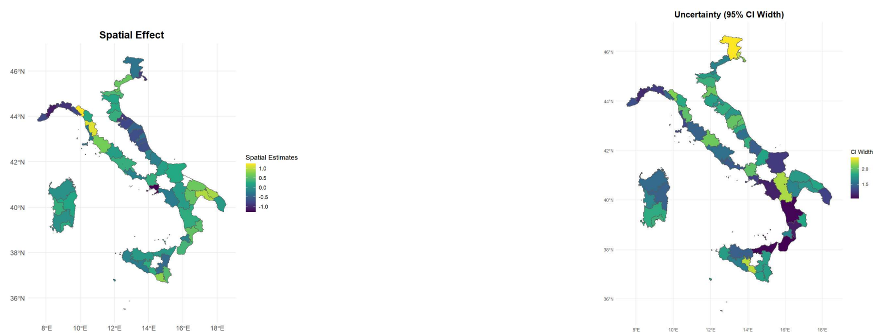


Figure 7.3: Estimated time effect on public expenditure for coastal soil defense, with 95% credible intervals. The solid line represents the posterior mean estimate over time, while the shaded ribbon shows the uncertainty range.

spatial variability in funding allocation (Figure 7.4(a)).



(a) Estimated spatial effect (mean).

(b) Uncertainty (95% CI width).

Figure 7.4: Estimated spatial effect on public expenditure for coastal soil defense and associated uncertainty for each province.

The map displays the estimated spatial effects for Italian provinces using the BYM2 model. The province of Barletta-Andria-Trani, shown as a white polygon in the southeast (Apulia region), was excluded due to the absence of funding data across its municipalities. Estimates for some provinces, particularly in parts of

Liguria, Campania, and Apulia, have narrower intervals (Figure 7.4(b)), indicating greater robustness and statistical reliability. This likely reflects higher data availability and consistency, as well as stronger spatial structure captured by the model. More generally, estimates with central values, represented by neutral colors on the map, tend to have wider credible intervals, indicating greater uncertainty. In contrast, provinces with more extreme spatial effect estimates often show narrower intervals, suggesting higher precision. However, the low value of the mixing parameter ($\varphi_{p_s i} = 0.18$) indicates that most of this variability is due to unstructured noise rather than to a well-defined spatial pattern. This may suggest that decision-making at the provincial level is influenced by localized factors, such as political priorities, rather than purely geographic trends in environmental risks. Finally, the precision for municipal effects (τ_s) suggests relatively low heterogeneity at the municipal level, implying that differences in funding allocation within municipalities are minor compared to those at broader spatial scales. This may reflect centralized decision-making processes that allocate resources at a regional or provincial level rather than in response to hyper-local variations in environmental risks. Overall, these precision estimates indicate that while spatial and temporal effects contribute to the variability in soil defense funding, decision-making processes likely involve additional socioeconomic and political considerations beyond purely environmental factors.

To further understand the overall effect, predictions from the model provide valuable insights into the geographical distribution of the response variable.

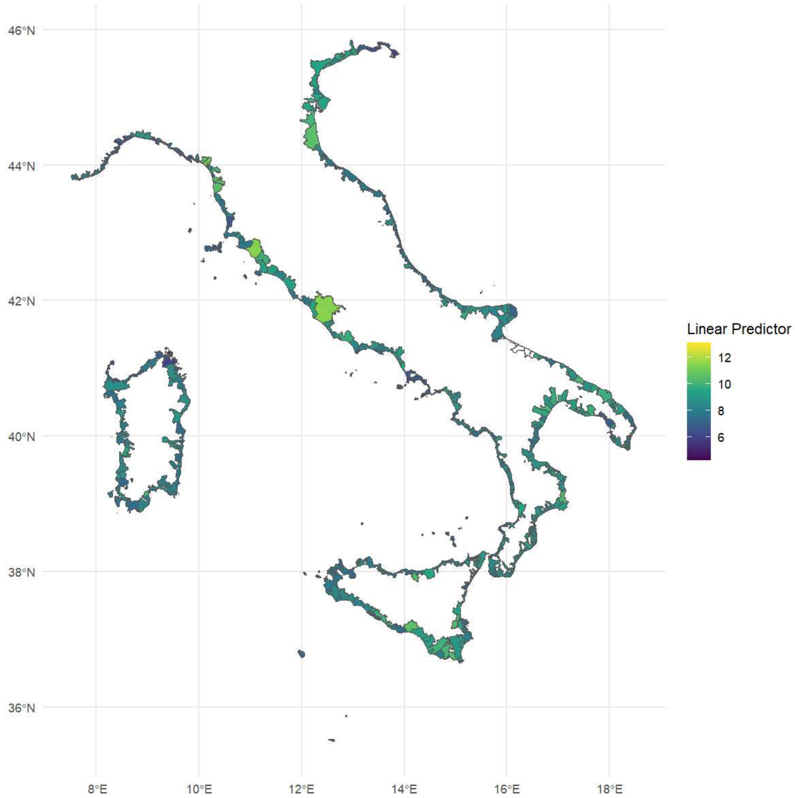


Figure 7.5: Estimated linear predictor effect ($\mu_{s,t}$).

Figure 7.5 shows the estimated linear predictor, $\mu_{s,t}$, incorporating the fixed effects of covariates along with spatial and temporal adjustments, indicating the combined effect of environmental and socio-economic drivers. The spatial distribution has more pronounced values in specific coastal areas, particularly along the northwestern and central-western coasts. The results confirm that areas with higher wave energy and extreme precipitation events tend to receive more funding. Interestingly, some coastal municipalities with limited direct exposure exhibit relatively lower predicted values, reinforcing the idea that the degree of coastal exposure plays a crucial role in fund allocation.

7.7 Policy Implications

Our research sheds light on the design and implementation of soil defense policies along Italy's coastline. First, the evidence that environmental pressures and socioeconomic conditions jointly shape funding allocation suggests the need for a more transparent and risk-sensitive approach to resource distribution. Currently, funding patterns appear to respond unevenly across municipalities and provinces, raising concerns about equity and efficiency in addressing climate and environmental risks.

Further, lagged environmental indicators underscore that public investments are not immediate but occur with temporal delays. Policymakers should account for such delays when planning interventions, as they may reduce the effectiveness of adaptation strategies in areas experiencing accelerating erosion or extreme weather events. Preventing and facilitating administrative procedures could help shorten the gap between hazard recognition and funding planning.

In addition, our study highlights the importance of harmonized data infrastructures. Linking financial, environmental, and social information across different sources remains a challenge, yet it is a prerequisite for transparent monitoring and effective policy evaluation. Improving interoperability between administrative datasets would strengthen the evidence base for decision-making and facilitate cross-regional comparisons. By providing empirical evidence on how ecological and social factors influence soil defense investments, this work supports the broader policy agenda of climate adaptation and disaster risk reduction. In line with EU cohesion policy and national climate strategies, the results advocate for prioritizing areas where both environmental hazards and social vulnerabilities converge, thereby enhancing resilience and ensuring more equitable protection for coastal communities.

7.8 Conclusion

This study aims to explore the relationship between financial and socio-environmental factors related to the defense of Italian coastal areas. Specifically, we sought to understand how changes in morphological features and environmental phenomena influence the allocation of funds for soil protection. Our findings suggest a clear connection between environmental and social factors and financial resource allocation. Increased funding is required following episodes of erosion and extreme weather events, highlighting a reactive approach to soil defense. Furthermore, we investigated spatial and temporal structures in the data, which, although subtle, revealed some patterns. The spatial component, albeit modest, confirms that the highest expenditures occur in the northern regions of the peninsula, particularly along the western coast, where extreme sea events are more frequent. We note that, by capturing spatial dependence via structured random effects at the province level with the BYM2 prior. It is important to acknowledge, however, that the decision to allocate funds is a complex process influenced by multiple factors beyond environmental changes. Political priorities, economic constraints, bureaucracy, and regional development strategies all play a role in shaping how resources are distributed. While our analysis provides valuable insights into the link between extreme environmental events and funding decisions, a more comprehensive assessment incorporating these additional determinants would be necessary for a full understanding of the allocation dynamics. More complete data regarding coastal defense infrastructures are needed to conduct a more in-depth and specific analysis. Additionally, incorporating further environmental variables—such as wave height, surge frequency, and damage indices—could refine our understanding of funding patterns. Building on our findings, future work could explicitly address structural heterogeneity, including macro-regional differences such as the North–South divide, which may influence resource allocation beyond local environmental and social conditions. Future research could also explore alternative modeling approaches, such as other Bayesian hierarchical models or machine learning techniques, to better capture the complexity of financial decision-making in response to coastal risks. Further, in a panel data perspective, spatial dependence could be estimated endogenously from the data using the method of moments (see for example Beenstock and Felsenstein [2012], Bhattacharjee and Jensen-Butler [2013]) and a spatial model such as a SAR model. This would allow for articulating the spatial effects as spillovers, copycat neighborhood effects, and so on. These insights underscore the need for a more proactive and strategic approach to resource distribution, emphasizing preventive measures rather than reactive spending in response to environmental disasters.

Chapter 8

Conclusions and Future Developments

This thesis has addressed multiple aspects of marine and coastal ecosystem analysis through the application of advanced statistical and computational methodologies. Across the different case studies, a common objective has been to deepen the understanding of the interactions between environmental processes, human activities, and decision-making mechanisms in the context of marine and coastal systems. By adopting flexible modeling frameworks such as GAMLSS, hidden semi-Markov models, and hierarchical Bayesian models, this research has demonstrated the effectiveness of modern statistical learning techniques in capturing the inherent complexity, uncertainty, and spatio-temporal variability that characterize environmental and socio-economic systems.

The first study, presented in Chapter 2, focused on sea state behaviour in the Adriatic Sea, offering a data-driven framework to identify and classify distinct marine conditions through the use of the GAMLSS approach. The analysis revealed that wind-related factors—particularly wind speed, direction, and fetch—play a dominant role in determining sea surface dynamics. The incorporation of these meteorological components within the modelling framework enabled the identification of specific regimes associated with both normal and surge conditions. The findings underscore the predictive value of such variables and provide a valuable foundation for sea state forecasting and for assessing the potential impacts of climate change on coastal regions.

Chapter 3 presents a methodological advancement by extending hidden semi-Markov models (HSMMs) to include concomitant variables in the analysis of *acqua alta* events in the Venice Lagoon. This framework proved effective in classifying risk periods and quantifying the influence of environmental and infrastructural factors, that is, the operation of the MOSE barrier system. The model captured both the transition dynamics among latent states and the duration of those states, revealing the system's stabilizing effect in mitigating high-water episodes. While the proposed model was specifically tailored to marine applications, its flexibility makes it suitable for a broad range of environmental and socio-economic systems where latent processes and covariate-dependent state durations are of interest. The discussion also highlighted possible non-parametric extensions and the integration of penalized likelihood methods to further enhance model flexibility.

A third line of research, presented in Chapter 4, applied hierarchical Bayesian spatio-temporal modelling to investigate the dynamics of sardine fisheries in the Mediterranean Sea. The results demonstrated that sardine catches are shaped by a complex combination of seasonal, environmental, and effort-related factors. The fitted models effectively standardized catch-per-unit-effort (CPUE) data and revealed clear spatial heterogeneity in abundance, as well as distinct temporal trends across regions. The findings provide valuable information for fisheries management, emphasizing the need for spatially and temporally adaptive strategies, including area-specific regulations and seasonal closures during reproductive periods. Beyond the specific case study, this work illustrates the potential of Bayesian hierarchical frameworks to disentangle multi-level sources of variability in ecological and economic data.

Last, Chapters 5, 6 and 7 shifted the focus from environmental processes to financial and policy dimensions of environmental management. First, the built of a comprehensive dataset, then the development of the `PublicWorksFinanceIT` R package provided a standardized, accessible, and reproducible framework for collecting and analysing data on soil defence projects in Italy. This tool enhances data transparency and analytical reproducibility, enabling policymakers and researchers to explore financial patterns and investment priorities more effectively. The final empirical study extended this work by examining the determinants of coastal defence expenditures, identifying strong links between environmental factors—such as erosion and extreme weather events—and subsequent increases in public funding. The analysis also revealed geographical disparities, with higher expenditures in northern coastal regions. Nonetheless, the results highlighted the reactive nature of funding decisions and called for a more strategic, preventive approach to resource allocation.

Taken together, these studies contribute to the growing literature on statistical modelling for environmental and socio-economic systems by illustrating how flexible, data-driven approaches can integrate diverse types of information—physical, ecological, and financial—into coherent inferential frameworks. The thesis advances both methodological and applied understanding: methodologically, it emphasizes the value of latent-variable modelling, spatio-temporal inference, and hierarchical structures; substantively, it demonstrates how such tools can inform sustainable management of marine and coastal resources.

Future research should continue to build on these results in several directions. From a methodological perspective, efforts could focus on extending current models to handle longer time series, richer sets of covariates, and non-parametric structures to reduce model misspecification risks. Integrating remote-sensing data and high-frequency observations could further enhance predictive accuracy, particularly in the context of sea level and flood forecasting. From a policy standpoint, combining environmental models with socio-economic indicators could support the development of integrated decision-support tools that bridge the gap between scientific evidence and public resource management.

This thesis highlights the essential role of combining statistical innovation with environmental understanding and policy relevance. Throughout this work, I have learned how powerful statistical models can be when used thoughtfully to interpret complex environmental and socio-economic data. The ability to handle data correctly, to select suitable models, and to interpret results with awareness and responsibility is not only a scientific necessity but also a moral one. In a world where data are increasingly abundant and accessible, the challenge lies not in collecting information, but in understanding it — in transforming numbers into knowledge and knowledge into action.

Working across multiple case studies has shown me how flexible statistical approaches can bridge the gap between theory and practice, linking rigorous quantitative analysis to real-world problems. These models are not just tools for academic exploration; they have the potential to guide decisions that affect ecosystems, economies, and communities. This awareness has reinforced my belief that data science and environmental science must evolve together, as part of an integrated effort to address the urgent challenges of climate change and sustainable resource management.

Ultimately, this research journey has deepened my appreciation for the delicate balance between methodological precision and societal impact. I have come to realize that data, when properly analysed and interpreted, can become a bridge between science and policy — a bridge that allows evidence-based decision-making to lead to more effective, equitable, and sustainable outcomes. As environmental and economic pressures continue to grow, the responsible use of data will be fundamental to designing policies that protect both our planet and the people who depend on it. I hope that the work presented in this thesis contributes, even in a small way, to this ongoing effort to align statistical innovation with the broader goal of understanding and preserving our natural world.

Appendices

Appendix A

Empirical Results for Hidden semi-Markov model with K=2

In this section we briefly report results from the Zero-Inflated Poisson Hidden semi-Markov model with K=2 states. Table A.1 reports parameter estimates alongside standard errors. Here two different states are identified characterized by different weather conditions influencing the *acqua alta*: namely, the steady-state and a more risky-state (Figure A.1), splitting the moderate-risk state discussed in Section 3.5 between the steady-state and the risky states, making the interpretation of the results fuzzier.

The probability of zeros (δ) significantly depends on most of the meteorological covariates except for the water temperature and the cosine of the wind direction. Still the covariates' effects resulted are the same described for the case with K=3 (Section 3.5). What differentiates the two results is the significance of the sine of the angle of the wind direction which has a negative impact on the probability of zero exceedances, indicating the wind directions which lead to high water events. Years 2023 and 2024 show negative values with respect to year 2020, meaning that the number of exceedances in those two years is higher than in 2020. The steady-state is characterized by two environmental covariates. Water temperature has a negative impact, meaning that lower water temperatures lead to more exceedances in State 1, emphasizing seasonality. Atmospheric pressure has a negative impact on the counts of exceedances, thus lower values of the air pressure will increase the number of the exceedances. In this case the MOSE activation indicates a negative

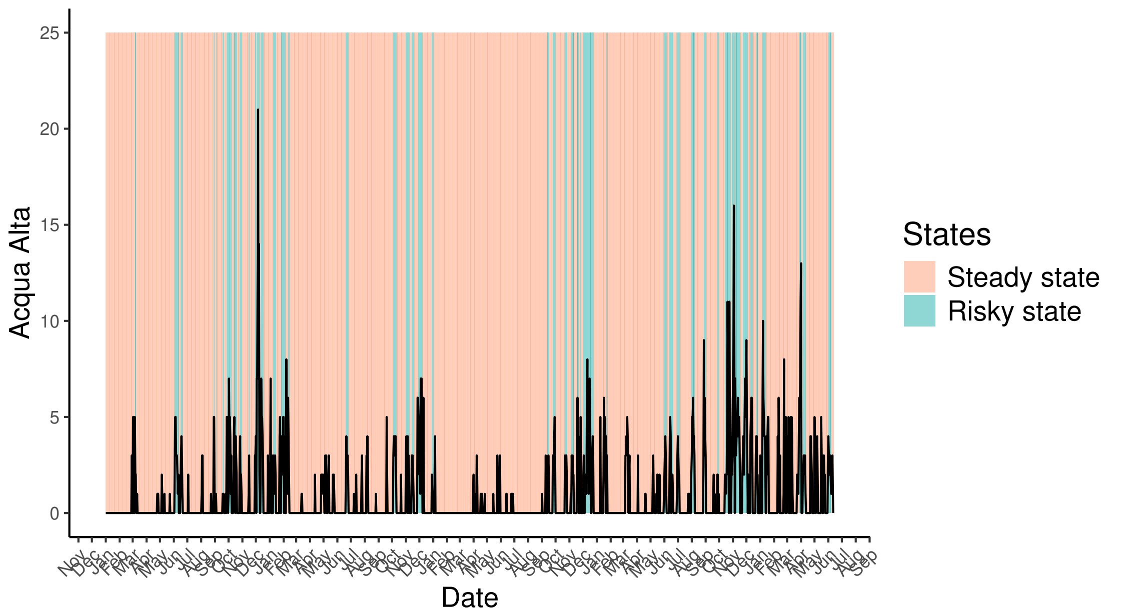
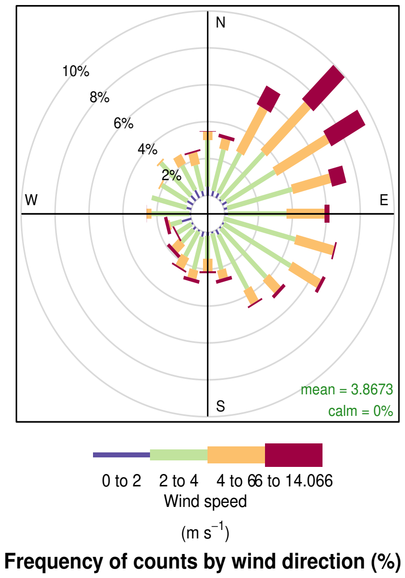


Figure A.1: Observed counts and inferred classification obtained by the Viterbi algorithm.

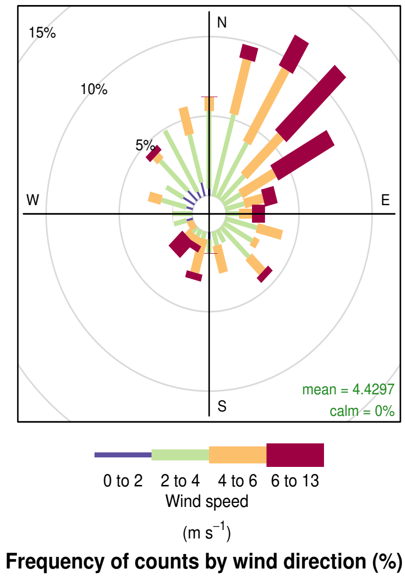
Table A.1: Estimated coefficients for the Zero-Inflated Hidden semi-Markov with $K = 2$ and endogeneity.

Variable	$\boldsymbol{\delta}$		$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$	
	Coefficients	Standard Errors	Coefficients	Standard Errors	Coefficients	Standard Errors
Intercept	**4.256**	0.702	0.101	0.359	**0.890**	0.084
Wind Speed	**-0.418**	0.118	-0.024	0.022	**0.092**	0.021
Temperature	-0.028	0.040	**-0.047**	0.022	**-0.031**	0.009
Humidity	**-0.062**	0.030	0.008	0.012	**0.013**	0.003
Pressure	**0.231**	0.036	**-0.044**	0.012	**-0.049**	0.005
Sin(Wind Direction)	**-0.450**	0.229	-0.118	0.112	-0.002	0.067
Cos(Wind Direction)	0.113	0.274	0.046	0.104	**0.086**	0.034
Year = 2020			Benchmark			
Year = 2021	-1.370	0.695	0.346	0.301	-0.229	0.132
Year = 2022	-0.701	0.894	-0.401	0.434	**-0.297**	0.124
Year = 2023	**-1.894**	0.714	-0.012	0.317	0.188	0.100
Year = 2024	**-2.682**	0.747	0.455	0.302	0.266	0.136
MOSE	-0.843	0.942	**-0.590**	0.206	-0.189	0.110
			$(\tilde{\zeta}_{01}, \tilde{\zeta}_{11})'$		$(\tilde{\zeta}_{02}, \tilde{\zeta}_{12})'$	
			Coefficients	Standard Errors	Coefficients	Standard Errors
Intercept			**19.438**	3.116	**4.453**	0.589
MOSE			**-17.102**	3.335	-0.798	1.392

impact on the counts of exceedances. MOSE activation has also a negative impact on the sojourn parameters meaning that its activation reduces the time spent in the steady-state. In this case the transition probability matrix was not modeled conditioned to the MOSE variable, since it is a $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ matrix. The risky-state comprehends higher values of counts of exceedances. The wind speed has a positive effect on the number of counts, while the water temperature still has a negative impact. An increase in the relative humidity increases the number of counts in the state, while an increase in the atmospheric pressure reduces the number of exceedances. In this case the cosine of the angle of the wind direction has a positive impact, highlighting the positive impact of the wind direction on the counts. In this case the wind direction identified in this state is the Bora wind (Figure A.2b). Year 2022 shows lower counts in this state compared to 2020. The MOSE does not show any impact on both the counts and the sojourn distribution parameters, since we are already in a state where high values of exceedances are registered.



(a)



(b)

Figure A.2: Wind Rose displaying prevalent wind directions and wind speed in the steady-state (Figure A.2a). Figure A.2b shows the wind rose displaying prevalent wind directions and wind speed in the risky-state.

Appendix B

Hyper-parameters estimates of the sardines landings models

Tables showing the hyper-parameter estimates for all the fitted models.

Parameter	**Model 1**				**Model 2**			
	Mean	SD	q0.025	q0.975	Mean	SD	q0.025	q0.975
τ_y	0.343	0.008	0.327	0.360	0.346	0.008	0.330	0.363
ρ_ω	1.069	0.152	0.799	1.397	1.031	0.144	0.769	1.338
σ_ω	2.056	0.129	1.807	2.318	2.046	0.126	1.807	2.306
τ_γ	–	–	–	–	161.936	277.554	10.825	812.913
τ_δ	264.956	176.742	38.156	693.612	288.209	192.046	60.958	782.619

Table B.1: Hyper-parameter estimates for the random effects in Model 1 and Model 2.

Parameter	**Model 3**				**Model 4**			
	Mean	SD	q0.025	q0.975	Mean	SD	q0.025	q0.975
τ_y	0.354	0.009	0.338	0.371	0.352	0.009	0.335	0.369
ρ_ω	1.181	0.172	0.883	1.559	1.141	0.164	0.858	1.503
σ_ω	2.094	0.142	1.832	2.393	2.084	0.135	1.832	2.364
τ_γ	21.055	16.389	4.758	64.666	59.561	43.464	12.546	173.695
τ_δ	289.042	187.970	60.894	768.990	281.995	208.251	64.547	833.256

Table B.2: Hyper-parameter estimates for the random effects in Model 3 and Model 4.

Appendix C

Reference Tables to interventions and functions

This section contains tables detailing the codes associated with each sub-sector and category of intervention referred to datasets retrieved from both databases.

Table C.1: Intervention Types and Codes

Nature code	Type Code	Type of Intervention
01	00	New Furniture
	08	Extraordinary Maintenance
	99	Other
02	10	Assistance
	11	Studies and Designs
	12	Training Courses
	13	Consulting
	14	Research Project
	99	Other
03	01	New Construction
	02	Demolition
	03	Restoration
	04	Restructuring

Continued on next page

Table C.1 (continued)

Nature code	Type Code	Type of Intervention
	05	Renovation
	06	Ordinary Maintenance
	07	Extraordinary Maintenance
	51	New Construction Completion
	52	Demolition Completion
	53	Restoration Completion
	54	Restructuring Completion
	55	Renovation Completion
	56	Ordinary Maintenance Completion
	57	Extraordinary Maintenance Completion
	58	Expansion
	99	Other
06	01	Purchase Real Services
	02	New Construction
	03	Renovation
	04	Restructuring
	08	Extraordinary Maintenance
	99	Other
07	01	New Construction
	04	Restructuring
	08	Extraordinary Maintenance
	09	Expansion
	11	Studies and Designs
	15	Modernisation
	16	Conversion
	17	Reactivation
	18	Relocation
	19	Purchase Real Services
	20	Research Activities

Continued on next page

Table C.1 (continued)

Nature code	Type Code	Type of Intervention
	21	Work Incentives
	99	Other
08	01	Purchase of Equity Investments
	02	Capital Injections
	99	Other

Table C.2: Main Sectors and Sub-sectors

Main Sector	**Sub-Sector**
01: Transport Infrastructures	01: Road Infrastructures 02: Airports 03: Railways 04: Maritime, Lacual and Fluvial 05: Urban Transport 06: Multimodal Transports and Other Transport Modes
02: Environmental and Water Resources Infrastructures	05: Land defence 10: Disposal of Effluents and Waste 11: Environmental Protection, Enhancement, and Enjoyment Work 12: Redevelopment and Recovery of Urban and Industrial Sites 15: Water Resources
03: Energy Sector Infrastructures	06: Energy Production 16: Energy Distribution
04: Infrastructure for the Equipment of Production Areas	38: Civil Infrastructures for Industrial Areas
05: Social Works and Infrastructures	08: Social and Educational 10: Housing 11: Works for the Recovery, Enhancement, and Use of Cultural Heritage 12: Sports, Entertainment, and Leisure 30: Healthcare 31: Worship 32: Defence

Continued on next page

Table C.2: Main Sectors and Sub-sectors (continued)

Main Sector	Sub-Sector
	33: Administrative and Offices
	34: Judicial and Penitentiary
	35: Public Safety
	99: Other Social Works and Infrastructure
06: Works, Installations, and Equipment for Productive Activities and Research	02: Works, Plants, and Equipment for the Silvo-Forestry Sector
	13: Works, Plants, and Equipment for Agriculture, Livestock Farming, and Silviculture
	14: Plants and Equipment for Fishing
	39: Works, Plants, and Equipment for Industrial and Craft Activities
	40: Works and Infrastructures for Research
	41: Works and Structures for Tourism
	42: Structures and Equipment for Commerce and Services
07: Infrastructure for Telecommunications and IT	17: Telecommunications Infrastructures
	18: Information Technology
08: Research, Technological Development, and Innovation	60: Projects for Public-Private Partnership and Dissemination
	61: Research Projects at Universities and Research Institutes

Continued on next page

Table C.2: Main Sectors and Sub-sectors (continued)

Main Sector	Sub-Sector
	62: Research Projects within Companies
09: Businesses Services	20: Services to Agricultural, Forestry, and Fisheries Enterprises
	21: Services to Industrial Enterprises
	22: Services to Tourism Enterprises
	23: Financial Intermediation Services
	25: Services to Commercial Enterprises
10: Services for Public Administration and for the Community	01: Services and Technologies for Information and Communications
	02: Services and Computer Applications for Citizens and Businesses
	03: Innovative Actions
	30: Services in Support of Development and Qualification of the Employment Services System
	32: Devices and Tools to Support the Qualification of the Training Offer System
	33: Devices and Tools to Support the Qualification of the Education Offer System
	34: Devices and Tools to Support the Integration between Systems
	41: Technical Assistance Services to Public Administration

Continued on next page

Table C.2: Main Sectors and Sub-sectors (continued)

Main Sector	Sub-Sector
	92: Services to Employees of Productive Enterprises
	93: Essential Services for the Rural Population
	99: Other Services for the Community
11: Training and Support for the Labor Market	71: Training for Employment
	72: Other Training and Work Experience Tools
	75: Employment Contributions and Incentives
	80: Other Support for the Labor Market

Table C.3: PublicWorksFinanceIT available functions.

Function	Category	Description
`get_codes`	Auxiliary	Retrieve codes for Italian regions, provinces, and municipalities
`get_data_OBDAP`	Download	Download data from OpenBDAP repository.
`get_data_region_OC`	Download	Download data from Open Coesione repository based on region.
`get_data_RENDIS`	Download	Download data from ReNDiS repository.
`get_data_theme_OC`	Download	Download data from Open Coesione repository based on theme's project.
`get_info_OC`	Auxiliary	Get information on regional and themes codes for the Open Coesione data.
`get_type_RENDIS`	Auxiliary	Get information on types of intervention from the ReNDiS data.
`merge_data`	Merging	Merge the three dataset eliminating duplicated observation.
`plot_funds_bar`	Plotting	Bar plot of funding distribution across for different financial resources.
`plot_funds_map`	Mapping	Areal map of the region of interest color based on the finance amount.
`plot_funds_points`	Mapping	Point map of the region of interest color based on the finance amount.

Bibliography

Timo Adam, Roland Langrock, and Christian H Weiß. Penalized estimation of flexible hidden Markov models for time series of counts. *Metron*, 77(2):87–104, 2019.

Pierre Ailliot, Julie Bessac, Valérie Monbet, and Françoise Pene. Non-homogeneous hidden Markov-switching models for wind time series. *Journal of Statistical Planning and Inference*, 160:75–88, 2015.

Tommaso Alberti, Marco Anzidei, Davide Faranda, Antonio Vecchio, Marco Favaro, and Alvise Papa. Dynamical diagnostic of extreme events in venice lagoon and their mitigation with the mose. *Scientific Reports*, 13(1):10475, 2023.

Marco Alfò and Antonello Maruotti. Two-part regression models for longitudinal zero-inflated count data. *Canadian Journal of Statistics*, 38(2):197–216, 2010.

Marco Alfò, Antonello Maruotti, and Giovanni Trovato. A finite mixture model for multivariate counts under endogenous selectivity. *Statistics and Computing*, 21: 185–202, 2011.

Mumtaz Ali, Ramendra Prasad, Yong Xiang, and Ravinesh C Deo. Near real-time significant wave height forecasting with hybridized multiple linear regression algorithms. *Renewable and Sustainable Energy Reviews*, 132:110003, 2020.

Anne Altmayer. The Mediterranean fishing area: State of play. Technical report, European Parliamentary Research Service, 2024. URL `https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2024)766245`.

Daniel Antal. Reproducible input-output economics analysis, economic and environmental impact assessment with empirical data, September 2025.

Micaela Antunes, Miguel Viegas, Celeste Varum, and Carlos Pinho. The impact of structural funds on regional growth: a panel data spatial analysis. *Intereconomics*, 55:312–319, 2020. doi: 10.1007/s10272-020-0921-1.

Servizio Mareografico & Università degli Studi di Roma Tre APAT. Atlante delle onde nei mari italiani = italian wave atlas. `https://www.ancientportsantiques.com/wp-content/uploads/Documents/ETUDESarchivees/Waves/ItalyWaveStat-Text2004.PDF`, 2004.

Fabian E. Bachl, Finn Lindgren, David L. Borchers, and Janine B. Illian. inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10:760–766, 2019. doi: 10.1111\/2041-210X.13168.

H. Bakka, H. Rue, G.-A. Fuglstad, A. Riebler, D. Bolin, J. Illian, E. Krainski, D. Simpson, and F. Lindgren. Spatial modeling with R-INLA: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(6):e1443, 2018. doi: 10.1002/wics.1443.

Vlad Barbu and Nikolaos Limnios. Maximum likelihood estimation for hidden semi-Markov models. *Comptes Rendus. Mathématique*, 342(3):201–205, 2006.

Francesco Bartolucci and Alessio Farcomeni. Information matrix for hidden Markov models with covariates. *Statistics and Computing*, 25:515–526, 2015.

Francesco Bartolucci and Alessio Farcomeni. A shared-parameter continuous-time hidden markov and survival model for longitudinal data with informative dropout. *Statistics in medicine*, 38(6):1056–1073, 2019.

Francesco Bartolucci, Alessio Farcomeni, and Fulvia Pennoni. *Latent Markov models for longitudinal data.* CRC Press, 2012.

Gualtiero Basilone, Rosalia Ferreri, Salvatore Aronica, Salvatore Mazzola, Angelo Bonanno, Antonella Gargano, Maurizio Pulizzi, Ignazio Fontana, Giovanni Giacalone, Pietro Calandrino, Simona Genovese, and Marco Barra. Reproduction and Sexual Maturity of European Sardine (Sardina pilchardus) in the Central Mediterranean Sea. *Frontiers in Marine Science*, 8, 2021. doi: 10.3389/fmars.2021.715846.

Michael Beenstock and Daniel Felsenstein. Nonparametric estimation of the spatial connectivity matrix using spatial panel data. *Geographical Analysis*, 44(4):386–397, 2012. doi: 10.1111/j.1538-4632.2012.00851.x.

Jafet Belmont, Sara Martino, Janine Illian, and Håvard Rue. Spatio-temporal occupancy models with INLA. *Methods in Ecology and Evolution*, 15(11):2087–2100, 2024. doi: 10.1111/2041-210X.14422.

Geir D Berentsen, Jan Bulla, Antonello Maruotti, and Bård Støve. Modelling clusters of corporate defaults: Regime-switching models significantly reduce the contagion source. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(3):698–722, 2022.

Arnab Bhattacharjee and Chris Jensen-Butler. Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics*, 43 (4):617–634, 2013. doi: 10.1016/j.regsciurbeco.2013.03.005.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725, 2000.

A. Bisaro, G. Galluccio, E. Fiorini Beckhauser, F. Biddau, R. David, F. d'Hont, A. Góngora Zurro, G. Le Cozannet, S. McEvoy, B. Pérez Gómez, C. Romagnoli, E. Sini, and J. Slinger. Sea level rise in europe: Governance context and challenges. *State of the Planet*, pages 1–52, 2024. doi: 10.5194/sp-3-slre1-7-2024.

Alexander Bisaro, Mark de Bel, Jochen Hinkel, Sien Kok, Tim Stojanovic, and Daniel Ware. Multilevel governance of coastal flood risk reduction: A public finance perspective. *Environmental Science & Policy*, 112:203–212, 2020. doi: 10.1016/j.envsci.2020.05.018.

NRRC Booij, Roeland C Ris, and Leo H Holthuijsen. A third-generation wave model for coastal regions: 1. model description and validation. *Journal of geophysical research: Oceans*, 104(C4):7649–7666, 1999.

Jan Bulla and Andreas Berzel. Computational issues in parameter estimation for stationary hidden Markov models. *Computational Statistics*, 23:1–18, 2008.

Jan Bulla, Ingo Bulla, and Oleg Nenadić. hsmm—an r package for analyzing hidden semi-Markov models. *Computational Statistics & Data Analysis*, 54(3):611–619, 2010.

Jan Bulla, Francesco Lagona, Antonello Maruotti, and Marco Picone. A multivariate hidden markov model for the identification of sea regimes from incomplete skewed and circular time series. *Journal of Agricultural, Biological, and Environmental Statistics*, 17:544–567, 2012.

D. Burdick, M. Hernández, H. Ho, G. Koutrika, R. Krishnamurthy, L. C. Popa, I. Stanoi, S. Vaithyanathan, and S. R. Das. Extracting, linking and integrating data from public sources: A financial case study. *Available at SSRN 2666384*, 2015. doi: 10.2139/ssrn.2666384.

Simona Buscemi, Annunziata Centra, Giovanni Finocchiaro, Cristina Frizza, Alessandra Galosi, Elisabetta Giovannini, Daria Gorozhankina, Silvia Iaccarino, Renato Marra Campanale, Raffaele Morelli, Francesca Palomba, Luca Segazzi, Nicolò G. Tria, and Patrizia Valentini. Ambiente in italia: uno sguardo d'insieme. annuario dei dati ambientali 2023. Technical report, Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA), 2024. URL `https://www.isprambiente.gov.it/files2024/pubblicazioni/stato-ambiente/ambiente-in-italia-annuario-2023.pdf`.

Gabriel Calvo, Carmen Armero, Luigi Spezia, and Maria Grazia Pennino. Bayesian joint longitudinal models for assessing the exploitation rates of sardine stock in the mediterranean sea. *Stoch Environ Res Risk Assess*, 38:1635–1646, 2024. doi: 10.1007/s00477-023-02649-0.

Michela Cameletti, Finn Lindgren, Daniel Simpson, and Håvard Rue. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, 97:109–131, 2013. doi: 10.1007/s10182-012-0196-3.

Caterina Canale, Giuseppe Barbaro, Olga Petrucci, Giovanni Besio, Giandomenico Foti, GIUSEPPINA CHIARA BARILLÀ, and PIERFABRIZIO PUNTORIERI. Analysis of the concurrent conditions of floods and sea storms: a case study of crotone, italy. *WIT Transactions on Engineering Sciences*, 129:147–156, 2020. ISSN 1743-3533. doi: 10.2495/RISK20013.

D Canesso, M Cordella, and G Arena. Manuale di mareografia e linee guida per i processi di validazione dei dati mareografici, 2012.

L Carbognin, P Teatini, and L Tosi. The impact of relative sea level rise on the northern adriatic sea coast, italy. *WIT Transactions on Ecology and the Environment*, 127:137–148, 2009.

David C. Carslaw and Karl Ropkins. openair — an r package for air quality data analysis. *Environmental Modelling I& Software*, 27–28(0):52–61, 2012. ISSN 1364-8152. doi: 10.1016/j.envsoft.2011.09.008.

Joaquin Cavieres and Orietta Nicolis. Using a spatio-temporal bayesian approach to estimate the relative abundance index of yellow squat lobster (Cervimunida johni) off Chile. *Fisheries Research*, 208:97 – 104, 2018. doi: 10.1016/j.fishres.2018.07.002.

Gilles Celeux and Gilda Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification*, 13:195–212, 1996.

Cathy WS Chen, Khemmanant Khamthong, and Sangyeol Lee. Markov switching integer-valued generalized auto-regressive conditional heteroscedastic models for dengue counts. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 68(4):963–983, 2019.

Joe Cheng, Bhaskar Karambelkar, Yihui Xie, Hadley Wickham, Kenton Russell, Kent Johnson, Barret Schloerke, and Vladimir Agafonkin. Package 'leaflet'. *R package version*, 2(1), 2019.

Francesco Colloca, Marco Enea, Sergio Ragonese, and Manfredi Di Lorenzo. A century of fishery data documenting the collapse of smooth-hounds (mustelus spp.) in the mediterranean sea. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 27(6):1145–1155, 2017a.

Francesco Colloca, Giuseppe Scarcella, and Simone Libralato. Recent Trends and Impacts of Fisheries Exploitation on Mediterranean Stocks and Ecosystems. *Frontiers in Marine Science*, 4:244, 2017b. doi: 10.3389/fmars.2017.00244.

European Commission. Directive 2003/98/ec of the european parliament and of the council of 17 november 2003 on the re-use of public sector information, 2003. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32003L0098`.

JAG Cooper, MC O'connor, and Sarah McIvor. Coastal defences versus coastal ecosystems: a regional appraisal. *Marine Policy*, 111:102332, 2020. doi: 10.1016/j.marpol.2016.02.021.

Copernicus Climate Change Service (C3S). Extreme precipitation risk indicators for europe and european cities from 1950 to 2019. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2021. URL `https://doi.org/10.24381/cds.3a9c4f89`. Accessed on February 3, 2025.

Eudriano FS Costa, Gustavo M Teixeira, Fúlvio AM Freire, June F Dias, and Adilson Fransozo. Effects of biological and environmental factors on the variability of paralonchurus brasiliensis (sciaenidae) density: An gamlss application. *Journal of Sea Research*, 183:102203, 2022.

C Mitchell Dayton and George B Macready. Concomitant-variable latent-class models. *Journal of the american statistical association*, pages 173–178, 1988.

Brajesh De. Api management. In *API Management: An Architect's Guide to Developing and Managing APIs for Your Organization*, pages 27–47. Springer, 2023. doi: 10.1007/979-8-8688-0054-2_2.

P. de Zea Bermudez, J. Miguel Marín, Håvard Rue, and Helena Veiga. Integrated nested laplace approximations for threshold stochastic volatility models. *Econometrics and Statistics*, 30:15–35, 2024. doi: 10.1016/j.ecosta.2021.08.006.

Partha Deb, Murat K Munkin, and Pravin K Trivedi. Bayesian analysis of the two-part model with endogeneity: application to health care expenditure. *Journal of Applied Econometrics*, 21(7):1081–1099, 2006.

Samar Deen, Verena Jauss, and Patrick J. Sullivan. Estimating effects of ocean environmental conditions on summer flounder (Paralichthys dentatus) distribution. *Environmental and Ecological Statistics*, 32(1):283 – 310, 2025. doi: 10.1007/s10651-025-00647-9.

Laurence L Delina, Chloe Chan, and Timothy Joseph Henares. A dataset on two decades of japan's energy investments in southeast asian countries, 2000–2020. *Data in Brief*, 52:109818, 2024. doi: 10.1016/j.dib.2023.109818.

Stacia M DeSantis and Dipankar Bandyopadhyay. Hidden Markov models for zero-inflated poisson counts with an application to substance use. *Statistics in medicine*, 30(14):1678–1694, 2011.

Roberto Di Mari and Antonello Maruotti. A two-step estimator for generalized linear models for longitudinal data with time-varying measurement error. *Advances in Data Analysis and Classification*, 16:273–300, 2022.

Chuong B Do and Serafim Batzoglou. What is the expectation maximization algorithm? *Nature biotechnology*, 26(8):897–899, 2008.

Theodoros Economou, Trevor C Bailey, and Zoran Kapelan. MCMC implementation for bayesian hidden semi-Markov models with illustrative applications. *Statistics and Computing*, 24:739–752, 2014.

Robert F Engle and Clive WJ Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276, 1987. doi: 10.2307/1913236.

European Commission: Joint Research Centre, Scientific, Technical and Economic Committee for Fisheries (STECF). Evaluation of fisheries dependent information (fdi) for eu fleets (stecf 24-11). Scientific and technical report, Publications Office of the European Union, 2025. URL `https://data.europa.eu/doi/10.2760/3446647`.

European Parliament and Council. Regulation (EU) 2017/1004 of the European Parliament and of the Council of 17 May 2017 on the establishment of a Union framework for the collection, management and use of data in the fisheries sector and support for scientific advice regarding the Common Fisheries Policy.

`https://eur-lex.europa.eu/eli/reg/2017/1004/oj/eng`, May 2017. Official Journal of the European Union, L 157, 20.6.2017, p. 1–21.

Cindy Xin Feng. A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of statistical distributions and applications*, 8 (1):8, 2021.

Andrew Fichera, Rachel King, Jarrod Kath, David Cobon, and Kathryn Reardon-Smith. Spatial modelling of agro-ecologically significant grassland species using the INLA-SPDE approach. *Scientific Reports*, 13(1), 2023. doi: 10.1038/s41598-023-32077-7.

Guido Fioravanti, Sara Martino, Michela Cameletti, and Andrea Toreti. Interpolating climate variables by using INLA and the SPDE approach. *International Journal of Climatology*, 43(14):6866–6886, 2023. doi: 10.1002/joc.8240.

Fabio Fiorentino and Sergio Vitale. How Can We Reduce the Overexploitation of the Mediterranean Resources? *Frontiers in Marine Science*, 8:674633, 2021. doi: 10.3389/fmars.2021.674633.

Grant Foster and Patrick T Brown. Time and tide: analysis of sea level time series. *Climate Dynamics*, 45(1-2):291–308, 2015.

Paolo Franco and Antonio Michelato. Northern adriatic sea: oceanography of the basin proper and of the western coastal zone. In *Marine coastal eutrophication*, pages 35–62. Elsevier, 1992.

Anna Freni-Sterrantino, Massimo Ventrucci, and Håvard Rue. A note on intrinsic conditional autoregressive models for disconnected graphs. *Spatial and spatio-temporal epidemiology*, 26:25–34, 2018. doi: 10.1016/j.sste.2018.04.002.

Geir-Arne Fuglstad, Daniel Simpson, Finn Lindgren, and Håvard Rue. Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, 114(525):445–452, 2019. doi: 10.1080/01621459.2017.1415907.

Eranga K Galappaththi, Vasantha B Susarla, Samantha JT Loutet, Stephanie T Ichien, Amanda A Hyman, and James D Ford. Climate change adaptation in fisheries. *Fish and Fisheries*, 23(1):4–21, 2022. doi: 10.1111/faf.12595.

P. L. Gallozzi, B. Dessì, C. Iadanza, E. M. Guarneri, T. Marasciulo, F. Miscione, D. Spizzichino, I. Rischia, and A. Trigila. Rendis 2020: La difesa del suolo in vent'anni di monitoraggio ispra sugli interventi per la mitigazione del rischio idrogeologico - edizione 2020. Technical Report Rapporti 328/20, ISPRA, 2020.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515 – 534, 2006. doi: 10.1214/06-BA117A.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis chapman & hall*. Chapman and Hall/CRC., 1995. URL `https://doi.org/10.1201/9780429258411`.

General Fisheries Commision for the Mediterranean. *The State of Mediterranean and Black Sea Fisheries 2023 - Special edition*. FAO, 2023. doi: 10.4060/cc8888en.

General Fisheries Commission for the Mediterranean (GFCM). Recommendation gfcm/44/2021/20 on a multiannual management plan for the sustainable exploitation of small pelagic stocks in the adriatic sea. `https://gfcm.sharepoint.com/.../REC.CM_GFCM_44_2021_20-e.pdf` (access restricted), 2021.

General Fisheries Commission for the Mediterranean (GFCM). Recommendation gfcm/47/2024/4 on a long-term fishing regime and the establishment of catch limits in 2025 for small pelagic stocks in the adriatic sea (gfcm gsas 17 & 18), stemming from recommendation gfcm/44/2021/20. Technical report, General Fisheries Commission for the Mediterranean (GFCM), 2024a. Adopted at the 47th GFCM session, October–November 2024; available via GFCM SharePoint (restricted access).

General Fisheries Commission for the Mediterranean (GFCM). Report of the Working Group on Stock Assessment of Small Pelagic Species (WGSASP), May 2024, Rome. Technical report, General Fisheries Commission for the Mediterranean

(GFCM), 2024b. Working Group on Stock Assessment of Small Pelagic Species (WGSASP), 06–08 May 2024.

Maurizio Gibin, Antonella Zanzi, Zeynep Hekim, Maciej Adamowicz, and Maksims Kovsars. Fisheries landings & effort: data by c-square, 2024. URL `http://data.europa.eu/89h/00ae6659-ddde-4314-a9da-717bb2e82582`.

Filippo Giorgi. Climate change hot-spots. *Geophysical research letters*, 33(8), 2006.

Virgilio Gómez-Rubio. *Bayesian inference with INLA*. Chapman and Hall/CRC, 2020. URL `https://doi.org/10.1201/9781315175584`.

Cesc Gordó-Vilaseca, Maria Grazia Pennino, Marta Albo-Puigserver, Matthias Wolff, and Marta Coll. Modelling the spatial distribution of Sardina pilchardus and Engraulis encrasicolus spawning habitat in the NW Mediterranean Sea. *Marine Environmental Research*, 2021. doi: 10.1016/j.marenvres.2021.105381.

Stuart K. Grange. *Technical note: saqgetr R package*, 2019. URL `https://drive.google.com/open?id=1IgDODHqBHewCTKLdAAxRyR7ml8ht6Ods`.

G. Guadagnoli and J. López. Rebuilding western Mediterranean fisheries: has the western Mediterranean multiannual plan delivered? 2019–2024, 2024. Accessed July 2025.

Pierre Guérin and Massimiliano Marcellino. Markov-switching midas models. *Journal of Business & Economic Statistics*, 31(1):45–56, 2013.

Nicolas Guillou. Modelling effects of tidal currents on waves at a tidal stream energy site. *Renewable Energy*, 114:180–190, 2017.

Beniamino Hadj-Amar, Jack Jewson, and Mark Fiecas. Bayesian approximations to hidden semi-Markov models for telemetric monitoring of physical activity. *Bayesian Analysis*, 18(2):547–577, 2023.

Robert P Haining and Guangquan Li. *Regression Modelling Wih Spatial and Spatial-Temporal Data: A Bayesian Approach.* CRC Press, 2020.

Manuel Hidalgo, AE El-Haweet, AC Tsikliras, EM Tirasin, T Fortibuoni, F Ronchi, V Lauria, O Ben Abdallah, E Arneri, L Ceriola, et al. Risks and adaptation options for the Mediterranean fisheries in the face of multiple climate change drivers and impacts. *ICES Journal of Marine Science*, 2022. doi: 10.1093/icesjms/fsac185.

James Honaker, Gary King, and Matthew Blackwell. Amelia ii: A program for missing data. *Journal of Statistical software*, 45(7):1–47, 2011. doi: 10.18637/jss.v045.i07.

Simon D. Hoyle, Robert A. Campbell, Nicholas D. Ducharme-Barth, Arnaud Grüss, Bradley R. Moore, James T. Thorson, Laura Tremblay-Boyer, Henning Winker, Shijie Zhou, and Mark N. Maunder. Catch per unit effort modelling for stock assessment: A summary of good practices. *Fisheries Research*, 269:106860, 2024. doi: 10.1016/j.fishres.2023.106860.

Weinan Huang and Sheng Dong. Probability distribution of wave periods in combined sea states with finite mixture models. *Applied Ocean Research*, 92:101938, 2019.

Weinan Huang and Sheng Dong. Joint distribution of individual wave heights and periods in mixed sea states using finite mixture models. *Coastal Engineering*, 161:103773, 2020.

ISTAT. Dati statistici per il territorio: Regione lazio, may 2020. URL `https://www.istat.it/it/files/2020/05/12_Lazio_Scheda_rev.pdf`.

Philip Jonathan and Kevin Ewans. Statistical modelling of extreme ocean environments for marine design: a review. *Ocean Engineering*, 62:91–109, 2013.

S. N. Jonkman, M. M. Hillen, R.J. Nicholls, W. Kanning, and M. van Ledden. Costs of adapting coastal defences to sea-level rise—new estimates and their implications. *Journal of Coastal Research*, 29(5):1212–1226, 2013. doi: 10.2112/JCOASTRES-D-12-00230.1.

Komali Kantamaneni, Louis Rice, Xiaoping Du, Belqais Allali, and Komali Yenneti. Are current uk coastal defences good enough for tomorrow? an assessment of vulnerability to coastal erosion. *Coastal Management*, 50(2):142–159, 2022. doi: 10.1080/08920753.2022.2022971.

Julia Köninger, Panagiotis Panagos, Arwyn Jones, MJI Briones, and Alberto Orgiazzi. In defence of soil biodiversity: Towards an inclusive protection in the european union. *Biological Conservation*, 268:109475, 2022. doi: 10.1016/j.biocon.2022.109475.

G. Korres, M. Ravdas, D. Denaxa, and M. Sotiropoulou. Mediterranean sea waves reanalysis interim (cmems med-waves, medwam3i system) (version 1) [data set], 2021. URL `https://doi.org/10.25423/CMCC/MEDSEA_MULTIYEAR_WAV_006_012_CLIM`.

Francesco Lagona and Marco Mingione. Nonhomogeneous hidden semi-markov models for toroidal data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, page qlae049, 2024.

Francesco Lagona and Marco Picone. Model-based clustering of multivariate skew data with circular components and missing values. *Journal of Applied Statistics*, 39(5):927–945, 2012.

Francesco Lagona and Zhen Zhang. A missing composite covariate in survival analysis: A case study of the chinese longitudinal health and longevity survey. *Statistics in medicine*, 29(2):248–261, 2010.

Francesco Lagona, Marco Picone, Antonello Maruotti, and Simone Cosoli. A hidden Markov approach to the analysis of space–time environmental data with linear and circular components. *Stochastic Environmental Research and Risk Assessment*, 29:397–409, 2015.

L. Lahti, J. Huovari, M. Kainu, and P. Biecek. Retrieval and analysis of eurostat open data with the eurostat package. *R J.*, 9(1):385, 2017. ISSN 2073-4859. URL `https://journal.r-project.org/archive/2017/RJ-2017-019/RJ-2017-019.pdf`.

Leo Lahti. Open data science. In Wouter Duivesteijn, Arno Siebes, and Antti Ukkonen, editors, *Advances in Intelligent Data Analysis XVII: 17th International Symposium, IDA 2018,'s-Hertogenbosch, The Netherlands, October 24–26, 2018, Proceedings 17*, pages 31–39, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01768-2. doi: 10.1007/978-3-030-01768-2_3.

Roland Langrock and Walter Zucchini. Hidden Markov models with arbitrary state dwell-time distributions. *Computational Statistics & Data Analysis*, 55(1):715–724, 2011.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011. doi: https://doi.org/10.1111/j.1467-9868.2011.00777.x.

Finn Lindgren, Fabian Bachl, Janine Illian, Man Ho Suen, Håvard Rue, and Andrew E Seaton. inlabru: software for fitting latent Gaussian models with non-linear predictors. *arXiv preprint arXiv:2407.00791*, 2024.

Piero Lionello, David Barriopedro, Christian Ferrarin, Robert J Nicholls, Mirko Orlić, Fabio Raicich, Marco Reale, Georg Umgiesser, Michalis Vousdoukas, and Davide Zanchettin. Extreme floods of venice: characteristics, dynamics, past and future evolution. *Natural Hazards and Earth System Sciences*, 21(8):2705–2731, 2021.

Iain L MacDonald. Numerical maximisation of likelihood: A neglected alternative to em? *International Statistical Review*, 82(2):296–308, 2014.

Iain L MacDonald and Feroz Bhamani. A time-series model for underdispersed or overdispersed counts. *The American Statistician*, 2018.

P. Maranzano and A. Algieri. *Retrieving and Analyzing Air Quality and Weather Data from ARPA Lombardia*, 2023. URL `https://CRAN.R-project.org/package=ARPALData`. R package version 1.5.0.

P. Maranzano and A. Algieri. Arpaldata: an r package for retrieving and analyzing air quality and weather data from arpa lombardia (italy). *Environmental and Ecological Statistics*, pages 1–32, 2024. doi: 10.1007/s10651-024-00599-6.

K Marini. Climate and environmental change in the mediterranean–main facts. *MedEC. Erişim*, 1:2019, 2018.

Reza Marsooli and Ning Lin. Numerical modeling of historical storm tides and waves and their interactions along the us east and gulf coasts. *Journal of Geophysical Research: Oceans*, 123(5):3844–3874, 2018.

Paloma Martín, Ana Sabatés, Josep Lloret, and Javier Martin-Vide. Climate modulation of fish populations: the role of the Western Mediterranean Oscillation (WeMO) in sardine (Sardina pilchardus) and anchovy (Engraulis encrasicolus) production in the north-western Mediterranean. *Climatic Change*, 110(3):925–939, 2012. doi: 10.1007/s10584-011-0091-z.

Antonello Maruotti. Handling non-ignorable dropouts in longitudinal data: a conditional model based on a latent markov heterogeneity structure. *Test*, 24(1): 84–109, 2015.

Antonello Maruotti and Pierfrancesco Alaimo Di Loro. Co2 emissions and growth: A bivariate bidimensional mean-variance random effects model. *Environmetrics*, page e2793, 2023.

Antonello Maruotti and Antonio Punzo. Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers. *Computational Statistics & Data Analysis*, 113:475–496, 2017.

Antonello Maruotti and Antonio Punzo. Initialization of hidden Markov and semi-Markov models: A critical evaluation of several strategies. *International Statistical Review*, 89(3):447–480, 2021.

Antonello Maruotti and Roberto Rocci. A mixed non-homogeneous hidden Markov model for categorical data, with application to alcohol consumption. *Statistics in Medicine*, 31(9):871–886, 2012.

Antonello Maruotti, Lea Petrella, and Luca Sposito. Hidden semi-Markov-switching quantile regression for time series. *Computational Statistics & Data Analysis*, 159: 107208, 2021.

G. Masselink and P. Russell. Impacts of climate change on coastal erosion. *MCCIP Science Review*, 2013:71–86, 2013. doi: 10.14465/2013.arc09.071-086.

Mark N. Maunder and André E. Punt. Standardizing catch and effort data: a review of recent approaches. *Fisheries Research*, 70(2):141–159, 2004. doi: 10.1016/j.fishres.2004.08.002.

Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2007.

Geoffrey J McLachlan, Thriyambakam Krishnan, and See Ket Ng. The em algorithm. Technical report, Papers, 2004.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019. doi: 10.1146/annurev-statistics-031017-100325.

P. Mercogliano, G. Rianna, A. Reder, M. Raffa, M. Mancini, M. Stojiljkovic, C. de Valk, and G. van der Schrier. Extreme precipitation risk indicators for europe and european cities from 1950 to 2019. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2021. URL `https://doi.org/10.24381/cds.3a9c4f89`. Accessed on February 3, 2025.

Luca Merlo, Antonello Maruotti, Lea Petrella, and Antonio Punzo. Quantile hidden semi-Markov models for multivariate time series. *Statistics and Computing*, 32 (4):61, 2022.

Ministero dell'Ambiente e della Sicurezza Energetica (MASE). Strategia nazionale di adattamento ai cambiamenti climatici (snac). Technical report, MASE, 2015. URL `https://www.mase.gov.it/portale/documents/d/guest/documento_snac-pdf`. Italian national adaptation strategy to climate change.

Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

Paula Moraga, Christopher Dean, Joshua Inoue, Piotr Morawiecki, Shahzeb Raja Noureen, and Fengpei Wang. Bayesian spatial modelling of geostatistical data using INLA and SPDE methods: A case study predicting malaria risk in Mozambique. *Spatial and Spatio-temporal Epidemiology*, 39:100440, 2021. doi: 10.1016/j.sste.2021.100440.

Sara Morucci, Marco Picone, Gabriele Nardone, and Giovanni Arena. Tides and waves in the central mediterranean sea. *Journal of Operational Oceanography*, 9: s10–s17, 2016. doi: 10.1080/1755876X.2015.1120960.

Fabien Moullec, Nicolas Barrier, Sabrine Drira, François Guilhaumon, Patrick Marsaleix, Samuel Somot, Caroline Ulses, Laure Velez, and Yunne-Jai Shin. An End-to-End Model Reveals Losers and Winners in a Warming Mediterranean Sea. *Frontiers in Marine Science*, 6:345, 2019. doi: 10.3389/fmars.2019.00345.

Simon Newcomb. A generalized theory of the combination of observations so as to obtain the best result. *American journal of Mathematics*, pages 343–366, 1886.

R. J. Nicholls, P. P. Wong, V. R. Burkett, J. O. Codignotto, J. E. Hay, R. F. McLean, S. Ragoonaden, and C. D. Woodroffe. *Coastal systems and low-lying areas*, chapter 6, pages 315–356. Cambridge University Press, 2007.

Jared O'Connell and Søren Højsgaard. Hidden semi Markov models for multiple observation sequences: The mhsmm package for r. *Journal of Statistical Software*, 39:1–22, 2011.

M Orlic, M Gacic, and PE Laviolette. The currents and circulation of the adriatic sea. *Oceanologica acta*, 15(2):109–124, 1992.

Marius Ötting, Roland Langrock, and Antonello Maruotti. A copula-based multivariate hidden markov model for modelling momentum in football. *AStA Advances in Statistical Analysis*, pages 1–19, 2021.

Panos Panagos, Pasquale Borrelli, Jean Poesen, Cristiano Ballabio, Emanuele Lugato, Katrin Meusburger, Luca Montanarella, and Christine Alewell. The new assessment of soil loss by water erosion in europe. *Environmental Science & Policy*, 54:438–447, 2015. doi: https://doi.org/10.1016/j.envsci.2015.08.012.

Greta Panunzi, Stefano Moro, Isa Marques, Sara Martino, Francesco Colloca, Francesco Ferretti, and Giovanna Jona Lasinio. Estimating the spatial distribution of the white shark in the Mediterranean Sea via an integrated species distribution model accounting for physical barriers. *Environmetrics*, page e2876, 2024. doi: 10.1002/env.2876.

Greta Panunzi, Stefano Moro, Isa Marques, Sara Martino, Francesco Colloca, Francesco Ferretti, and Giovanna Jona Lasinio. Estimating the spatial distribution of the white shark in the Mediterranean Sea via an integrated species distribution model accounting for physical barriers. *Environmetrics*, 36:e2876, 2025. doi: 10.1002/env.2876.

Maria Grazia Pennino, David Conesa, Antonio López-Quílez, Facundo Muñoz, Angel Fernández, and Jose Maria Bellido. Fishery-dependent and -independent data lead to consistent estimations of essential habitats. *ICES Journal of Marine Science*, 73(9):2302–2310, 2016. doi: 10.1093/icesjms/fsw062.

Arthur Pewsey, Markus Neuhäuser, and Graeme D Ruxton. *Circular statistics in R*. OUP Oxford, 2013.

Hung Vuong Pham, Maria Katherina Dal Barco, Marco Cadau, Remi Harris, Elisa Furlan, Silvia Torresan, Sara Rubinetti, Davide Zanchettin, Angelo Rubino, Ivan Kuznetsov, et al. Multi-model chain for climate change scenario analysis to support coastal erosion and water quality risk management for the metropolitan city of venice. *Science of The Total Environment*, 904:166310, 2023.

Jennifer Pohle, Roland Langrock, Floris M Van Beest, and Niels Martin Schmidt. Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, 22:270–293, 2017.

Jennifer Pohle, Timo Adam, and Larissa T Beumer. Flexible estimation of the state dwell-time distribution in hidden semi-Markov models. *Computational Statistics & Data Analysis*, 172:107479, 2022.

B. F. Prahl, M. Boettle, L. Costa, J. P. Kropp, and D. Rybski. Damage and protection cost curves for coastal floods within the 600 largest european cities. *Scientific data*, 5(1):1–18, 2018. doi: 10.1038/sdata.2018.34.

United Nations Environment Programme. State of finance for nature: The big nature turnaround – repurposing $7 trillion to combat nature loss. Technical report, United Nations, december 2023.

Eric Prudhommeaux. Sparql query language for rdf. *http://www.w3.org/TR/rdf-sparql-query/*, 2008.

André E Punt. Spatial stock assessment methods: a viewpoint on current issues and assumptions. *Fisheries Research*, 213:132–143, 2019. doi: 10.1016/j.fishres.2019.01.014.

Antonio Punzo, Salvatore Ingrassia, and Antonello Maruotti. Multivariate generalized hidden Markov regression models with random covariates: physical exercise in an elderly population. *Statistics in Medicine*, 37(19):2797–2808, 2018.

Antonio Punzo, Salvatore Ingrassia, and Antonello Maruotti. Multivariate hidden Markov regression models: random covariates and heavy-tailed distributions. *Statistical Papers*, 62(3):1519–1555, 2021.

Gennaro Punzo, Rosalia Castellano, and Emma Bruno. Exploring land use determinants in italian municipalities: Comparison of spatial econometric models. *Environmental and Ecological Statistics*, 29(4):727–753, 2022. doi: 10.1007/s10651-022-00541-8.

Debin Qiu and Maintainer Debin Qiu. Package 'atsa'. *Erişim adresi https://cran.r-project. org/web/packages/aTSA/aTSA. pdf*, 2015. doi: 10.32614/CRAN.package.aTSA.

R R Core Team et al. R core team r: a language and environment for statistical computing. *Foundation for Statistical Computing*, 2020.

Damien Raclot, Yves Le Bissonnais, Mohamed Annabi, Mohamed Sabir, and Anna Smetanova. Main issues for preserving mediterranean soil resources from water erosion under global change. *Land Degradation & Development*, 29(3):789–799, 2018. doi: 10.1002/ldr.2774.

O. J. Reichman, M. B. Jones, and M. P. Schildhauer. Challenges and opportunities of open data in ecology. *Science*, 331(6018):703–705, 2011. doi: 10.1126/science.1197962.

Benny Ren and Ian Barnett. Combining mixed effects hidden Markov models with latent alternating recurrent event processes to model diurnal active-rest cycles. *Biometrics*, 79:3402–3417, 2023.

Lorena Ricciotti. Geo-referenced Harmonized Financial Data on Soil Defense Public Works in Italy [data set], mar 2024.

Lorena Ricciotti and Alessio Pollice. *PublicWorksFinanceIT: Soil Defense Investments in Italy: Data Retrieval, Analysis, Visualization*, 2024a. R package version 0.2.0.

Lorena Ricciotti and Alessio Pollice. *PublicWorksFinanceIT: Soil Defense Investments in Italy: Data Retrieval, Analysis, Visualization.* https://CRAN.R-project.org/package=PublicWorksFinanceIT, 2024b. Version 0.3.0.

Lorena Ricciotti and Alessio Pollice. An integrated georeferenced dataset of public investments for soil defence in italy. *Data in Brief*, 57:110909, 2024c. doi: 10.1016/j.dib.2024.110909.

Lorena Ricciotti, Marco Picone, Alessio Pollice, and Antonello Maruotti. Modelling and clustering sea conditions: Bivariate finitemixtures of generalized additive models for location, shape, and scale applied to the analysis of meteorological tides and wave heights. *Journal of Marine Science and Engineering*, 12(5):740, 2024. doi: 10.3390/jmse12050740.

Lorena Ricciotti, Marco Picone, Alessio Pollice, and Antonello Maruotti. A zero-inflated hidden semi-markov model with covariate-dependent sojourn parameters for analysing marine data in the venice lagoon. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 74(2):506–529, 2025. doi: 10.1093/jrsssc/qlae065.

Andrea Riebler, Sigrunn H Sørbye, Daniel Simpson, and Håvard Rue. An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4):1145–1165, 2016a. doi: 10.1177/0962280216660421.

Andrea Riebler, Sigrunn H Sørbye, Daniel Simpson, and Håvard Rue. An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4):1145–1165, 2016b. doi: 10.1177/0962280216660421.

Robert A Rigby and D Mikis Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005.

Silvia Ronchi, Stefano Salata, Andrea Arcidiacono, Erika Piroli, and Luca Montanarella. Policy instruments for soil protection among the eu member states: A comparative analysis. *Land use policy*, 82:763–780, 2019. doi: 10.1016/j.landusepol.2019.01.017.

Rosa F Ropero, M Julia Flores, and Rafael Rumí. Flash floods in mediterranean catchments: a meta-model decision support system based on bayesian networks. *Environmental and Ecological Statistics*, 31(1):27–56, 2024. doi: 10.1007/s10651-023-00587-2.

Claudio Rubino, Giada Adelfio, Antonino Abbruzzo, Mar Bosch-Belmar, Manfredi Di Lorenzo, Fabio Fiorentino, Vita Gancitano, Francesco Colloca, and Giacomo Milisenda. Exploring the effects of temperature on demersal fish communities in the Central Mediterranean Sea using INLA-SPDE modeling approach. *Environmental and Ecological Statistics*, 31(3):629–647, 2024. doi: 10.1007/s10651-024-00609-7.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2): 319–392, 2009. doi: 10.1111/j.1467-9868.2008.00700.x.

Håvard Rue, Andrea Riebler, Sigrunn H Sørbye, Janine B Illian, Daniel P Simpson, and Finn K Lindgren. Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, 4(1):395–421, 2017. doi: 10.1146/annurev-statistics-060116-054045.

Ana CL Sá, Maria AA Turkman, and José MC Pereira. Exploring fire incidence in portugal using generalized additive models for location, scale and shape (gamlss). *Modeling Earth Systems and Environment*, 4:199–220, 2018.

Sujit Sahu. *Bayesian modeling of spatio-temporal data with R.* Chapman and Hall/CRC, 2022.

Pietro Scala, Giuseppe Cipolla, Dario Treppiedi, and Leonardo Valerio Noto. The use of gamlss framework for a non-stationary frequency analysis of annual runoff data over a mediterranean area. *Water*, 14(18):2848, 2022.

Francesco Scorza. Improving EU Cohesion Policy: The spatial distribution analysis of regional development investments funded by eu structural funds 2007/2013 in italy. In Beniamino Murgante, Sanjay Misra, Maurizio Carlini, Carmelo M. Torre, Hong-Quang Nguyen, David Taniar, Bernady O. Apduhan, and Osvaldo Gervasi, editors, *Computational Science and Its Applications – ICCSA 2013*, pages 582–593. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-39646-5_42.

A. Seidl, K. Mulungu, M. Arlaud, O. van den Heuvel, and M. Riva. The effectiveness of national biodiversity investments to protect the wealth of nature. *Nature Ecology & Evolution*, 5(4):530–539, 2021. doi: 10.1038/s41559-020-01372-1.

Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G. Martins, and Sigrunn H. Sørbye. Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, 32(1):1 – 28, 2017. doi: 10.1214/16-STS576.

Rajesh Singh and Quinn Weninger. Discretion rather than rules in multiple-species fisheries. *Journal of Environmental Economics and Management*, 125:102983, 2024. doi: 10.1016/j.jeem.2024.102983.

Pete Smith, Rosa M Poch, David A Lobb, Ranjan Bhattacharyya, Ghiath Alloush, Gaius D Eudoxie, Lúcia HC Anjos, Michael Castellano, Georges M Ndzana, Claire Chenu, et al. Status of the world's soils. *Annual review of environment and resources*, 49(2024):73–104, 2024. doi: 10.1146/annurev-environ-030323-075629.

M Soto, Lourdes Fernández-Peralta, J Rey, I Czerwisnki, Ramón García-Cancela, Marcos Llope, J Cabrera-Busto, M Liébana, and MG Pennino. Modelling drivers of trawl fisheries discards using Bayesian spatio-temporal models. *Fisheries Research*, 268:106830, 2023. doi: 10.1016/j.fishres.2023.106830.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series b (statistical methodology)*, 64(4):583–639, 2002. doi: 10.1111/1467-9868.00353.

Timo Michael Staeudle, Bram Parmentier, and Jan Jaap Poos. Accounting for spatio-temporal distribution changes in size-structured abundance estimates for a data-limited stock of Raja clavata. *ICES Journal of Marine Science*, 81(8): 1607 – 1619, 2024. doi: 10.1093/icesjms/fsae106.

D. Mikis Stasinopoulos and Robert A. Rigby. Generalized additive models for location scale and shape (GAMLSS) in *R. Journal of Statistical Software*, 23(7), 2007. ISSN 1548-7660. doi: 10.18637/jss.v023.i07. URL `http://www.jstatsoft.org/v23/i07/`.

Mikis D Stasinopoulos, Robert A Rigby, Gillian Z Heller, Vlasios Voudouris, and Fernanda De Bastiani. *Flexible regression and smoothing: using GAMLSS in R*. CRC Press, Taylor & Francis Group, 2017.

Mikis D Stasinopoulos, Robert A Rigby, and Fernanda De Bastiani. Gamlss: A distributional regression approach. *Statistical Modelling*, 18(3-4):248–273, 2018.

A. G. Stephenson. *Harmonic Analysis of Tides Using TideHarmonics*, 2016. URL `https://CRAN.R-project.org/package=TideHarmonics`.

Oliver Stoner and Theo Economou. An advanced hidden Markov model for hourly rainfall time series. *Computational Statistics & Data Analysis*, 152:107045, 2020.

E. Szöcs, T. Stirling, E. R. Scott, A. Scharmüller, and R. B. Schäfer. webchem: an r package to retrieve chemical information from the web. *Journal of Statistical Software*, 93:1–17, 2020. doi: 10.18637/jss.v093.i13.

Sigrunn Holbek Sørbye and Håvard Rue. Scaling intrinsic gaussian markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51, 2014. ISSN 2211-6753. doi: 10.1016/j.spasta.2013.06.004. Spatial Statistics Miami.

Olli Tahvonen, Martin F Quaas, and Rüdiger Voss. Harvesting selectivity and stochastic recruitment in economic models of age-structured fisheries. *Journal of Environmental Economics and Management*, 92:659–676, 2018. doi: 10.1016/j.jeem.2017.08.011.

A. Tassan Mazzocco, P. Maranzano, and P. Borgoni. *Handle Air Quality Data from the European Environment Agency Data Portal*, 2023. URL `https://CRAN.R-project.org/package=EEAaq`. R package versione 0.0.3.

Core Writing Team, H. Lee, and J. Romero, editors. *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland, 2023. doi: 10.59327/IPCC/AR6-9789291691647.

A. C. Teucher, S. J. Albers, and S. L. Hazlitt. bcdata: An r package for searching & retrieving data from the bc data catalogue. *Journal of Open Source Software*, 6(61):2927, 2021. doi: 10.21105/joss.02927.

S. Thacker, D. Adshead, M. Fay, S. Hallegatte, M. Harvey, H. Meller, N. O'Regan, J. Rozenberg, G. Watkins, and J. W. Hall. Infrastructure for sustainable development. *Nature Sustainability*, 2(4):324–331, 2019. doi: 10.1038/s41893-019-0256-8.

C. Toso, F. Madricardo, E. Molinaroli, S. Fogarin, A. Kruss, A. Petrizzo, N.M. Pizzeghello, L. Sinapi, and F. Trincardi. Tidal inlet seafloor changes induced by recently built hard structures. *PloS one*, 14(10), 2019. doi: 10.1371/journal.pone.0223240.

Isabel F Trigo and Trevor D Davies. Meteorological conditions associated with sea surges in venice: a 40 year climatology. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 22(7):787–803, 2002.

M Pilar Tugores, Magdalena Iglesias, Núria Díaz, Dolores Onate, Joan Miquel, and Ana Giráldez. Latitudinal and interannual distribution of the European anchovy (Engraulis encrasicolus) and sardine (Sardina pilchardus) in the western Mediterranean, and sampling uncertainty in abundance estimates. *ICES Journal of Marine Science*, 2010. doi: 10.1093/icesjms/fsq057.

T Rolf Turner, Murray A Cameron, and Peter J Thomson. Hidden Markov chains in generalized linear models. *Canadian Journal of Statistics*, 26(1):107–125, 1998.

Georg Umgiesser, Marco Bajo, Christian Ferrarin, Andrea Cucco, Piero Lionello, Davide Zanchettin, Alvise Papa, Alessandro Tosoni, Maurizio Ferla, Elisa Coraci, et al. The prediction of floods in venice: methods, models and uncertainty. *Natural Hazards and Earth System Sciences*, 21(8):2679–2704, 2021.

United Nations Framework Convention on Climate Change (UNFCCC). Investment and financial flows to address climate change. UNFCCC Secretariat, 2007. URL `https://unfccc.int/resource/docs/publications/financial_flows.pdf`.

OAR US EPA. Climate change indicators: Oceans. Reports and assessments, US EPA, OAR, 2016. URL `https://www.epa.gov/climate-indicators/oceans`.

Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2): 260–269, 1967.

Andrew J Viterbi. A personal history of the viterbi algorithm. *IEEE Signal Processing Magazine*, 23(4):120–142, 2006.

Peiming Wang and Joseph D Alba. A zero-inflated negative binomial regression model with hidden Markov chain. *Economics Letters*, 92(2):209–213, 2006.

Sumio Watanabe. A widely applicable bayesian information criterion. *The Journal of Machine Learning Research*, 14(1):867–897, 2013. doi: 10.5555/2567709.2502609.

Michel Wedel. Concomitant variables in finite mixture models. *Statistica Neerlandica*, 56(3):362–375, 2002.

Danielle Welter, Nick Juty, Philippe Rocca-Serra, Fuqi Xu, David Henderson, Wei Gu, Jolanda Strubel, Robert T Giessmann, Ibrahim Emam, Yojana Gadiya, et al. Fair in action-a flexible framework to guide fairification. *Scientific data*, 10(1): 291, 2023. doi: 10.1038/s41597-023-02167-2.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Dawn Williams, James Haworth, Marta Blangiardo, and Tao Cheng. A spatiotemporal bayesian hierarchical approach to investigating patterns of confidence in the police at the neighborhood level. *Geographical Analysis*, 51(1):90–110, 2019. doi: 10.1111/gean.12160.

Shun-Zheng Yu. Hidden semi-Markov models. *Artificial intelligence*, 174(2):215–243, 2010.

Davide Zanchettin, Sara Bruni, Fabio Raicich, Piero Lionello, Fanny Adloff, Alexey Androsov, Fabrizio Antonioli, Vincenzo Artale, Eugenio Carminati, Christian Ferrarin, et al. Sea-level rise in venice: historic and future trends. *Natural Hazards and Earth System Sciences*, 21(8):2643–2678, 2021.

Chong Wei Zheng and Chong Yin Li. Variation of the wave energy and significant wave height in the china sea and adjacent waters. *Renewable and Sustainable Energy Reviews*, 43:381–387, 2015.

Igor S Zonn and Andrey G Kostianoy. The adriatic sea. *The Boka Kotorska Bay Environment*, pages 19–41, 2017.

Walter Zucchini, Iain L MacDonald, and Roland Langrock. *Hidden Markov models for time series: an introduction using R*. CRC press, 2017.

A. Zuiderwijk, M. Janssen, S. Choenni, R. Meijer, and R. S. Alibaks. Socio-technical impediments of open data. *Electronic Journal of e-Government*, 10(2): pp156–172, december 2012. URL `https://academic-publishing.org/index.php/ejeg/article/view/571/534`.

List of Figures

List of Tables