Comparing Bayesian statistical modelling with machine learning in spatio-temporal disease mapping

Diego Samper Ovejero

Escuela Técnica Superior de Ingeniería Agronómica y Biociencias

Universidad Pública de Navarra

Data Science & Statistics

Supervisor: Dr. María Dolores Ugarte

Contents

Chapter 1

Background & objectives

During my previous studies in Data Science, I had the opportunity to delve deeply into subjects related to statistics and modelling. These subjects allowed me to understand the importance of data analysis and its impact on real-life problems. However, throughout my career, I have had no contact with Bayesian inference, an approach that involves fitting a probability model to data and summarizing the outcome through a probability distribution on model parameters and unobserved quantities like predictions for new observations. This gap in my knowledge has led me to believe that there is much to learn about Bayesian inference and how it can improve my work in data analysis.

Another aspect that caught my attention is the lack of space-dependent models in my previous studies. These models are essential in the analysis of spatial data. Therefore, I believe that incorporating space-dependent models in my work could help me gain a more comprehensive understanding of spatial data and improve my modelling skills.

It is for these reasons that I have decided to embark on a new project to deepen my knowledge of Bayesian inference and space-time modelling. I am particularly interested in exploring the use of the Integrated Nested Laplace Approximation (INLA) methodology, which allows for fast and accurate approximations of posterior distributions, making it an ideal tool for analyzing large and complex datasets.

Additionally, I plan to compare classical machine learning models such as Extreme Gradient Boosting or Random Forest and deep learning models such as Long-Short Term Memory (LSTM) or Bayesian Neural Network (BNN) with Bayesian statistical models fitted with INLA to determine their strengths and weaknesses. By identifying which modelling approach is best suited for different types of datasets and analysis tasks, I aim to become a more versatile data analyst.

To these ends, we first introduce the theoretical framework explaining the concepts of Bayesian inference, classical machine learning and deep learning in Chapter 2. In Chapter 3, we perform an exploratory data analysis to gain a better understanding of the problem we are facing. Subsequently, rate modelling is presented in Chapter 4, where we outline the advantages and drawbacks of each method. We end this work in Chapter 5 with the conclusions and ideas on further work.

Chapter 2

Theoretical frameworks

2.1 Bayesian statistics

Bayesian statistics (see [1] and [2]) is a framework for understanding and quantifying uncertainty. The main difference between Bayesian and frequentist statistics is in how they approach the concept of probability. In frequentist statistics, probability is viewed as the long-run frequency of an event in repeated trials. On the other hand, Bayesian statistics sees probability as a measure of uncertainty about a particular event or phenomenon, a measure of the degree of belief that an individual has in that statement. This belief can be revised as new evidence becomes available.

To make probability statements about a certain parameter θ given observed data y, we first need a model for the joint probability distribution. The joint probability can be written as the product of the prior and the likelihood. That is, $p(\theta, y) = p(\theta)p(y|\theta)$.

Then, we use Bayes' theorem 2.1, where $p(\theta|y)$ is the posterior density, $p(y|\theta)$ is the likelihood, $p(\theta)$ is the prior and $p(y)$ the marginal distribution, to update our beliefs or knowledge about a particular event as we gather more information.

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta) \times p(\theta)}{p(y)} \tag{2.1}$$

As we can see, the posterior distribution will always be placed between the likelihood and the prior distribution.

One example of Bayesian statistics in the field of lung cancer is in determining the effectiveness of a new diagnostic test. Let's say we have a new test for lung cancer and we want to know how accurate it is at detecting the disease. We could use Bayesian statistics to update our prior belief about the accuracy of the test based on data from clinical trials. Suppose our prior belief is that the test is 80% accurate. We then collect data from a clinical trial that shows the test correctly identified 90 out of 100 lung cancer cases. Using Bayesian statistics, we can update our belief about the accuracy of the test to a posterior distribution that reflects our new knowledge. This can help us make more informed decisions about the use of the test in clinical practice.

In summary, both Bayesian and frequentist statistics can be used to evaluate the accuracy of the new diagnostic test for lung cancer, but they approach the problem differently. Bayesian statistics provides a probability distribution of the test's sensitivity based on prior knowledge and data, while frequentist statistics provides a point estimate with a confidence interval based only on the data.

2.2 Spatial data

Spatial data refers to any type of data that contains information about the physical location or spatial characteristics of objects, events, or phenomena. It can be found in various fields, including but not limited to geography, environmental science, urban planning, and epidemiology, and is used to

analyze, model, and visualize spatial patterns and relationships. Spatial data can be categorized into three primary types: point-reference data (also called geostatistical data), areal data (also known as lattice data), and point pattern data, each with its own unique characteristics and applications.

In this text, we will focus on areal data, where the region of study is partitioned into a finite number of areal units with well-defined boundaries.

2.3 Spatio-temporal models in disease mapping

The use of spatio-temporal models in disease mapping has gained significant attention in recent times due to its ability to offer a comprehensive understanding of how diseases are distributed and spread over both space and time. This modeling approach allows researchers to detect areas that have both high and low risks, and to identify any changes in the geographic patterns from one year to another. By taking into account both geographic and temporal dimensions, these models provide valuable insights into how the disease risk is evolving across the entire region, as well as how specific regions are affected. As a result, spatio-temporal disease mapping provides a holistic understanding of diseases, enabling the development of more effective interventions and policies that can inform public health decision-making.

Assuming the region of study is divided into n areas, and the time domain has T consecutive periods, we will establish some notation:

- Y_{it} is the number of observed deaths in the i-th area at time t
- E_{it} is the number of expected deaths in the i-th area at time t
- R_{it} is the relative risk of the i-th area at time t
- $i = 1, ..., n$ and $t = 1, ..., T$

Our assumption is that $Y_{it}|R_{it} \sim Poisson(E_{it}R_{it})$ independently and thus, the maximum likelihood estimator of R_{it} is SMR_{it}.

The Standardized Mortality Ratio (SMR, 2.2) is a statistical measure used to compare the observed number of deaths in a population to the number of deaths that would be expected, i f the risk in the i-th area and t-th year was the global risk over space and time. It assumes that the observed number of deaths follows a Poisson distribution, which implies that the events (deaths) occur independently of each other. When the data are independent and there is a reasonable amount of data, the SMR is an appropriate and reliable measure to estimate the relative risk of mortality in different populations or subgroups.

$$SMR_{it} = \frac{Y_{it}}{E_{it}} \tag{2.2}$$

In general, to compute the expected cases E_{it}, we usually stratify the population by variables such as age and gender. Therefore,

$$E_{it} = \sum_{j=1}^{J} n_{itj} R_j, \quad R_j = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} O_{itj}}{\sum_{i=1}^{n} \sum_{t=1}^{T} N_{itj}}$$

We can calculate the variance and the standard error as follows:

$$var[\hat{R}_{it}] = \frac{var[Y_{it}]}{E_{it}^2} = \frac{E_{it}R_{it}}{E_{it}^2} = \frac{R_{it}}{E_{it}}$$

$$s.e.[\hat{R}_{it}] = \sqrt{\frac{\hat{R}_{it}}{E_{it}}} = \sqrt{\frac{Y_{it}}{E_{it}^2}} = \frac{\sqrt{Y_{it}}}{E_{it}}$$

So, in fact the SMR compares the number of observed cases in the i-th area with the number of cases we would expect to observe if the small area had the same mortality rate as the whole region.

Notice that the SMR can be extremely variable when dealing with rare diseases or low-populated areas. This leads to methods capable of smoothing risks in space and time, such as hierarchical mixed Poisson models.

These models can be fitted from a frequentist point of view, or from a fully Bayes approach (most common in disease mapping).

where n_{itj} is the population at risk in the i-th area, t-th moment and j-th stratum and R_j is the mortality rate in the j-th stratum.

The model used in this field 2.3 is a generalized linear model GLM.

$$Y_{it}|R_{it} \sim Poisson(\mu_{it} = E_{it}R_{it}), \quad i = 1, .., n \quad and \quad t = 1, ..., T \tag{2.3}$$

$$\log(\mu_{it}) = \log(E_{it}) + \log(R_{it})$$

From now on, we will sometimes let $\log(R_{it}) = \beta_{it}$, and thus, $R_{it} = e^{\beta_{it}}$. The $\log(E_{it})$ is an offset.

The specification of $\log(R_{it})$ gives rise to different models. One of these is the [3] model where

$$\log(R_{it}) = \alpha + \xi_i + \gamma_t + \phi_t + \delta_{it} \tag{2.4}$$

Here, α is a global intercept, ξ_i is a spatial structured random effect with a Leroux prior assigned to it, γ_t is a temporary structured random effect, ϕ_t is a temporary unstructured random effect and δ_{it} is a random effect to take account of the spatio-temporal effect.

In practice, the temporal effect is usually structured, so the uncorrelated temporal component ϕ_t can be removed and the following reduced model is considered:

$$\log(R_{it}) = \alpha + \xi_i + \gamma_t + \delta_{it} \tag{2.5}$$

Let's define what each component exactly is:

- α quantifies the logarithm of the global risk and represents an overall risk.
- $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)' \sim N(\mathbf{0}, [\tau_\xi(\lambda_\xi \boldsymbol{R}_\xi + (1 - \lambda_\xi)\boldsymbol{I}_n)]^-)$ where $\lambda_\xi \in [0, 1]$ is a spatial smoothing parameter, $\boldsymbol{I}_n$ is the identity matrix and $\boldsymbol{R}_\xi$ is the neighborhood matrix ($R_{ij} = -1 \iff i \; and \; j \; are \; neighbors$, otherwise $R_{ij} = 0$).

  When $\lambda_\xi = 0$, the Leroux prior becomes $\boldsymbol{\xi} \sim N(\mathbf{0}, \tau_\xi^{-1}\boldsymbol{I}_n)$.

  On the other hand, when $\lambda_\xi = 1$, the Leroux prior reduces to $\boldsymbol{\xi} \sim N(\mathbf{0}, [\tau_\xi^{-1}\boldsymbol{R}_\xi]^-)$.
- γ_t is usually modeled with a random walk of order one or two (RW1 or RW2). For RW1, the conditional distribution takes the form (see [4] Theorem 2.3, pp 22- , pp 97 and pp 110):

$$\gamma_t|\gamma_{-t} \sim N\left(\frac{1}{2}(\gamma_{t-1} + \gamma_{t+1}), \frac{\sigma_\gamma^2}{2}\right)$$

The joint distribution of $\boldsymbol{\gamma}$ can be expressed as $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_\gamma^2 \boldsymbol{R}_\gamma^-)$, where $\boldsymbol{R}_\gamma$ for a RW1 takes the form [4]:

$$\boldsymbol{R}_\gamma = \begin{bmatrix} 1 & -1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ -1 & 2 & -1 & \ddots & & & & \vdots \\ 0 & -1 & 2 & -1 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & -1 & 2 & -1 & 0 \\ \vdots & & & & \ddots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & -1 & 1 \end{bmatrix} \tag{2.6}$$

- $\boldsymbol{\delta} = (\delta_{11}, ..., \delta_{nT})'$ is the vector of the space–time interaction effects, that is assumed to follow the following multivariate distribution: $\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma_\delta^2 \boldsymbol{R}_\delta^-)$. $\boldsymbol{R}_\delta$ is the structure matrix of the space-time interaction and is given by the Kronecker product of $\boldsymbol{R}_\gamma$ and $\boldsymbol{R}_\xi$: $\boldsymbol{R}_\delta = \boldsymbol{R}_\gamma \otimes \boldsymbol{R}_\xi$.

  Depending on the structure matrix of δ, different types of interactions arise:

Table 2.1: Possible types of space-time interactions

Space-time interaction	$\boldsymbol{R}_\delta$	Rank of $\boldsymbol{R}_\delta$
Type I	$\boldsymbol{I}_\xi \otimes \boldsymbol{I}_t$	$I \cdot T$
Type II	$\boldsymbol{I}_\xi \otimes \boldsymbol{R}_t$	$I \cdot (T-1)$
Type III	$\boldsymbol{R}_\xi \otimes \boldsymbol{I}_t$	$(I-1) \cdot T$
Type IV	$\boldsymbol{R}_\xi \otimes \boldsymbol{R}_t$	$(I-1) \cdot (T-1)$

  Type I interaction means δ_{it} are all independent. There is no structure in space and time. Type II is adequate when there is structure in time but not in space whereas Type III considers structure in space but not in time. Finally, Type IV is appropriate when δ_{it} are structured in space and time and is suitable if temporal trends from neighboring regions are likely to be similar.

Model 2.5 presents identifiability issues, as space and time effects have an implicit intercept which cannot be distinguished from the overall level, and the interaction terms are entangled with the main effects. To address this problem and achieve model identifiability, sum-to-zero constraints are typically imposed on the random effects of the model. In [5], they summarize the necessary identifiability constraints for different types of space-time interactions, displayed in Table 2.1 using RW1 priors for the temporally structured random effect.

Note that all the random effects in equation 2.5 are modeled as Gaussian Markov random fields (GMRF).

2.4 Integrated Nested Laplace Approximation

In recent years, the use of spatial and spatio-temporal data has become increasingly important in a wide range of scientific fields, including disease mapping, ecology, and climate modeling. These data often have complex structures, which require sophisticated statistical methods for analysis. Bayesian modeling is a popular approach for analyzing such data, but traditional methods for Bayesian inference, such as Markov Chain Monte Carlo (MCMC), can be computationally intensive and time-consuming, especially for complex models.

The Integrated Nested Laplace Approximation (INLA) approach, proposed by [6], uses a deterministic algorithm for Bayesian inference. It is especially designed for latent Gaussian models 2.5, a subclass of structured additive regression models which are flexible enough to be used in many different types of applications. INLA is a statistical method that provides a computationally efficient alternative to MCMC for fitting Bayesian models to spatial and spatio-temporal data. INLA, compared to MCMC, provides accurate results in shorter computing time. It also has the advantage of providing an approximation of the marginal likelihood, which can be used for model selection and comparison.

The key idea behind INLA is to approximate the posterior distribution of the model parameters using a sequence of nested Laplace approximations. The first Laplace approximation is used to estimate the posterior mode and curvature, while subsequent approximations are used to estimate the remaining posterior distribution. This approach allows for the efficient estimation of complex models, even for large datasets.

By fitting a Gaussian distribution with a mean equivalent to the Maximum A Posteriori (MAP) solution and a variance equivalent to the observed Fisher information, Laplace's approximation offers an analytical expression for a posterior probability distribution. This approximation is supported by the Bernstein-von Mises theorem, which asserts that in large samples, the posterior converges to a Gaussian distribution under regularity conditions.

Let's see how Laplace's approximation works. Suppose we want to compute the following integral:

$$\int f(x)dx = \int \exp(\log f(x))dx$$

where $f(x)$ is the density function of a random variable X. Using the Taylor series expansion for the $\log f(x)$ at $x = x_0$:

$$\log f(x) \approx \log f(x_0) + \left.\frac{\partial \log f(x)}{\partial x}\right|_{x=x_0} (x - x_0) + \left.\frac{\partial^2 \log f(x)}{\partial x^2}\right|_{x=x_0} \frac{(x - x_0)^2}{2}$$

If x_0 is equal to the mode $x^* = \text{argmax}_x \log f(x)$, then $\left.\frac{\partial \log f(x)}{\partial x}\right|_{x=x^*} = 0$ and the approximation becomes:

$$\log f(x) \approx \log f(x^*) + \left.\frac{\partial^2 \log f(x)}{\partial x^2}\right|_{x=x^*} \frac{(x - x^*)^2}{2}$$

Therefore, the integral of interest is approximated as follows:

$$\begin{aligned}\int f(x)dx &\approx \int \exp\left(\log f(x^*) + \left.\frac{\partial^2 \log f(x)}{\partial x^2}\right|_{x=x^*} \frac{(x - x^*)^2}{2}\right) dx \\ &= \exp(\log f(x^*)) \int \exp\left(\left.\frac{\partial^2 \log f(x)}{\partial x^2}\right|_{x=x^*} \frac{(x - x^*)^2}{2}\right) dx\end{aligned}$$

The integrand can be approximated by the density of a Normal distribution by setting

$$\sigma^{2*} = -1/\left(\left.\frac{\partial^2 \log f(x)}{\partial x^2}\right|_{x=x^*}\right)$$

$$\int f(x)dx \approx \exp(\log f(x^*)) \int \exp\left(\frac{-(x - x^*)^2}{2\sigma^{2*}}\right) dx$$

where the integrand is the kernel of a Normal distribution with mean equal to x^* and variance σ^{2*}. More precisely

$$\int_\alpha^\beta f(x)dx \approx f(x^*)\sqrt{2\pi\sigma^{2*}}(\Phi(\beta) - \Phi(\alpha))$$

where $\Phi(.)$ denotes the cumulative density function of the Normal $\left(x^*, \sigma^{2*}\right)$ distribution.

As an example, suppose we need to compute the following integral, where $f(x)$ is the Weibull density function. We will calculate the result by means of Laplace's approximation:

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, \quad \int_2^5 f(x)dx$$

For that, we will need the following quantities:

$$\log f(x) = \log\left(k\lambda^{-k}x^{k-1}\right) - \left(\frac{x}{\lambda}\right)^k$$

$$\frac{\partial}{\partial x}\left(\log f(x)\right) = \frac{\partial}{\partial x}\left(\log\left(k\lambda^{-k}x^{k-1}\right) - \left(\frac{x}{\lambda}\right)^k\right) = \frac{-k\left(\frac{x}{\lambda}\right)^k + k - 1}{x}$$

$$\frac{\partial^2}{\partial x^2}\left(\log f(x)\right) = \frac{\partial}{\partial x}\left(\frac{-1 + k - k\left(\frac{x}{\lambda}\right)^k}{x}\right) = -\frac{(k-1)\left(k\left(\frac{x}{\lambda}\right)^k + 1\right)}{x^2}$$

By solving $\frac{\partial \log f(x)}{\partial x} = 0$, we obtain the mode $x^* = \lambda\sqrt[k]{1 - \frac{1}{k}}$. The variance is obtained by evaluating $\sigma^{2*} = -1/\frac{\partial^2 \log f(x)}{\partial x^2}$ at the mode x^*:

$$\sigma^{2*}|_{x=x^*} = -1/\left(-\frac{(k-1)\left(k\left(\frac{x}{\lambda}\right)^k + 1\right)}{x^2}\right)\Bigg|_{x=x^*} = \frac{x^2}{(k-1)(k(\frac{x}{\lambda})^k + 1)}\Bigg|_{x=x^*}$$

$$\sigma^{2*}|_{x=x^*} = \frac{\left(\frac{k-1}{k}\right)^{2/k}\lambda^2}{(k-1)\left(k\sqrt[k]{\frac{k-1}{k}} + 1\right)}$$

Therefore,

$$\int_2^5 f(x)dx \approx f\left(x^*\right)\sqrt{2\pi\sigma^{2*}}(\Phi(\beta) - \Phi(\alpha))$$

Once we know how to perform Laplace's approximation, which is the core to INLA, let's remember a few basic properties of conditionals.

In what follows, we provide a concise recap of several properties associated with conditional distributions. This will serve as a foundation for understanding how INLA operates, which we will discuss subsequently.

Given any pair of variables (x, y) and provided that $p(y) > 0$,

$$p(x \mid y) = \frac{p(x,y)}{p(y)} \Longrightarrow p(x,y) = p(x \mid y)p(y)$$

then

$$p(y) = \frac{p(x,y)}{p(x \mid y)}$$

If we consider a third variable z we can write

$$p(y \mid z) = \frac{p(x, y \mid z)}{p(x \mid y, z)}$$

which is particularly relevant to the Bayesian case.

The general idea is that we can approximate a generic conditional (posterior) distribution as

$$p(y \mid z) \approx \frac{p(x, y \mid z)}{\tilde{p}(x \mid y, z)}$$

where $\tilde{p}(x \mid y, z)$ is the Laplace approximation to the conditional distribution of x given y and z.

This idea can be used to approximate any generic required posterior distribution.

The main aim of Bayesian inference for Latent Gaussian Models (LGM, explained in Section 2.5) are the posterior marginals. We denote the latent gaussian field as θ and the hyperparameters as ψ:

$$p(\theta_i \mid \boldsymbol{y}) = \int p(\theta_i, \psi \mid \boldsymbol{y})\, d\psi = \int p(\theta_i \mid \psi, \boldsymbol{y})\, p(\psi \mid \boldsymbol{y}) d\psi$$
$$p(\psi_k \mid \boldsymbol{y}) = \int p(\psi \mid \boldsymbol{y}) d\psi_{-k}$$

On the other hand, the approximated posterior marginals of interest returned by INLA have the following form:

$$\tilde{p}(\theta_i \mid \boldsymbol{y}) = \ldots = \int \tilde{p}(\theta_i \mid \psi, \boldsymbol{y})\, \tilde{p}(\psi \mid \boldsymbol{y}) d\psi$$
$$\tilde{p}(\psi_k \mid \boldsymbol{y}) = \int \tilde{p}(\psi \mid \boldsymbol{y}) d\psi_{-k}$$

Thus, we need to find expressions for $p(\theta_i \mid \psi, \boldsymbol{y})$ and $p(\psi \mid \boldsymbol{y})$.

The approximation to the second term is straightforward and presented right below, whereas the one to the first term is slightly more complex, because most of the times there will be more elements in $\boldsymbol{\theta}$ than there are in ψ, making the computation more expensive.

$$\begin{aligned}
p(\psi \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{\theta}, \psi \mid \boldsymbol{y})}{p(\boldsymbol{\theta} \mid \psi, \boldsymbol{y})} \\
&= \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}, \psi) p(\boldsymbol{\theta}, \psi)}{p(\boldsymbol{y})} \frac{1}{p(\boldsymbol{\theta} \mid \psi, \boldsymbol{y})} \\
&= \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}, \psi) p(\boldsymbol{\theta} \mid \psi) p(\psi)}{p(\boldsymbol{y})} \frac{1}{p(\boldsymbol{\theta} \mid \psi, \boldsymbol{y})} \\
&\propto \frac{p(\psi) p(\boldsymbol{\theta} \mid \psi) p(\boldsymbol{y} \mid \boldsymbol{\theta}, \psi)}{p(\boldsymbol{\theta} \mid \psi, \boldsymbol{y})} \\
&\approx \left. \frac{p(\psi) p(\boldsymbol{\theta} \mid \psi) p(\boldsymbol{y} \mid \boldsymbol{\theta}, \psi)}{\tilde{p}(\boldsymbol{\theta} \mid \psi, \boldsymbol{y})} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*(\psi)} =: \tilde{p}(\psi \mid \boldsymbol{y})
\end{aligned}$$

where $\tilde{p}(\theta \mid \psi, \boldsymbol{y})$ is the Laplace approximation of $p(\theta \mid \psi, \boldsymbol{y})$ and $\theta = \theta^*(\psi)$ is its mode for a given ψ. Since *psi* has few dimensions, we can get the marginals for $\psi_k|\boldsymbol{y}$ directly from the approximation to $\psi|\boldsymbol{y}$.

Moving on to the first term we wanted to approximate, there are two possibilities:

1. Gaussian Approximation using $\tilde{p}(\boldsymbol{\theta} \mid \psi, \boldsymbol{y})$.
2. Full Laplace Approximation: we can write $\theta = \{\theta_i, \theta_{-i}\}$, use the definition of conditional probability and again Laplace approximation to obtain

$$
\begin{aligned}
p\left(\theta_i \mid \psi, \boldsymbol{y}\right) &= \frac{p\left(\left(\theta_i, \boldsymbol{\theta}_{-i}\right) \mid \psi, \boldsymbol{y}\right)}{p\left(\boldsymbol{\theta}_{-i} \mid \theta_i, \boldsymbol{\psi}, \boldsymbol{y}\right)} = \frac{p\left(\left(\theta_i, \boldsymbol{\theta}_{-i}\right), \boldsymbol{\psi} \mid \boldsymbol{y}\right)}{p(\psi \mid \boldsymbol{y})} \frac{1}{p\left(\theta_{-i} \mid \theta_i, \psi, \boldsymbol{y}\right)} \\
&\propto \frac{p(\boldsymbol{\theta}, \psi \mid \boldsymbol{y})}{p\left(\boldsymbol{\theta}_{-i} \mid \theta_i, \psi, \boldsymbol{y}\right)} \quad \propto \frac{p(\psi) p(\boldsymbol{\theta} \mid \psi) p(\boldsymbol{y} \mid \boldsymbol{\theta}, \psi)}{p\left(\boldsymbol{\theta}_{-i} \mid \theta_i, \psi, \boldsymbol{y}\right)} \\
&\approx \left.\frac{p(\psi) p(\boldsymbol{\theta} \mid \psi) p(\boldsymbol{y} \mid \boldsymbol{\theta}, \psi)}{\tilde{p}\left(\boldsymbol{\theta}_{-i} \mid \theta_i, \psi, \boldsymbol{y}\right)}\right|_{\boldsymbol{\theta}_{-i}=\boldsymbol{\theta}_{-i}^*(\theta_i, \psi)} =: \tilde{p}\left(\theta_i \mid \psi, \boldsymbol{y}\right)
\end{aligned}
$$

This approximation is generally correct but is computationally expensive. Once we get $\tilde{p}\left(\theta_i \mid \psi, \mathbf{y}\right)$ and $\tilde{p}(\psi \mid \mathbf{y})$, the marginal posterior distribution $p\left(\theta_i \mid \mathbf{y}\right)$ is then approximated by

$$
\tilde{p}\left(\theta_i \mid \mathbf{y}\right) \approx \int \tilde{p}\left(\theta_i \mid \psi, \mathbf{y}\right) \tilde{p}(\psi \mid \mathbf{y}) d\psi
$$

and the integral can be solved numerically through a finite weighted sum

$$
\tilde{p}\left(\theta_i \mid \mathbf{y}\right) \approx \sum_j \tilde{p}\left(\theta_i \mid \psi^{(*)}, \mathbf{y}\right) \tilde{p}\left(\psi^{(*)} \mid \mathbf{y}\right) \Delta^*
$$

for some relevant integration points $\psi^{(*)}$ with a corresponding set of weights Δ^*.

Taking all the preceding information into account, we are now able to outline the specific actions that INLA performs:

1. Examine the combined posterior distribution of the hyperparameters, $\tilde{p}(\psi|\boldsymbol{y})$, and generate a collection of favorable integration points ψ^* that are linked with the majority of the distribution's mass. Additionally, create a corresponding set of associated area weights Δ^*.
2. Following the grid search, compute the marginal posterior distribution $\tilde{p}(\psi_k|\boldsymbol{y})$ by applying an interpolation method that relies on the density values of $\tilde{p}(\psi|\boldsymbol{y})$ evaluated at the integration points ψ^*.
3. Calculate the estimated marginal distribution $\tilde{p}(\theta_i|\psi^*, \boldsymbol{y})$ for certain chosen values of parameter θ_i at each integration point in ψ^*.
4. For each i, use numerical integration to obtain

$$
\tilde{p}(\theta_i|\boldsymbol{y}) \approx \sum_{\psi^*} \tilde{p}(\theta_i|\psi^*, \boldsymbol{y}) \tilde{p}(\psi^*|\boldsymbol{y}) \Delta^*
$$

The so called *Simplified Laplace Approximation* is a computationally efficient alternative to the Full Laplace Approximation, which is much faster and provides similar results.

2.5 Latent Gaussian models

For a statistician, the crux of the issue in (parametric) inference lies in establishing a probability model that can account for the observed data, by leveraging pertinent parameters that are central to the statistical problem at hand. Conditional Independence is assumed in the first hierarchy level.

$$
\mathbf{y} \mid \boldsymbol{\theta}, \psi \sim p(\mathbf{y} \mid \boldsymbol{\theta}, \psi) = \prod_{i=1}^{n} p\left(y_i \mid \theta_i, \psi\right)
$$

In the second one, we assume that the parameters are described by a Gaussian Markov Random Field (GMRF) $\theta \mid \psi \sim \text{Normal}\left(\mathbf{0}, \mathbf{Q}^{-1}(\psi)\right)$, where $\mathbf{Q}$ is the precision matrix and verifies that, for a general pair i and j, with $i \neq j$ it holds that the corresponding element of the precision matrix is null

$$\theta_i \perp\!\!\!\perp \theta_j \mid \theta_{-i,j} \Leftrightarrow Q_{ij}(\psi) = 0$$

Therefore, it is clear that in GMRF we have sparse precision matrices, which can be exploited for very quick computations for the Gaussian part of the model. Latent Gaussian models are flexible prior models which explicitly model dependence among samples and which allow for efficient learning of predictor functions and for making probabilistic predictions.

The joint posterior distribution of θ and ψ is given by the product of the likelihood, the GMRF density and the hyperparameter prior distribution,

$$p(\theta, \psi|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta, \psi) \times p(\theta|\psi) \times p(\psi)$$

We can restate a LGM by partitioning $\psi = (\psi_1, \psi_2)$. Hence, the LGM will take the following form:

$$\begin{aligned}
\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\psi} &\sim \prod_i p\left(y_i \mid \theta_i, \psi_2\right) && \text{(data model)} \\
\boldsymbol{\theta} \mid \psi &\sim \text{Normal}\left(\mathbf{0}, \mathbf{Q}^{-1}\left(\psi_1\right)\right) && \text{(GMRF prior)} \\
\psi &\sim p(\psi) && \text{(hyperprior)}
\end{aligned}$$

where (a) ψ_1 are the hyper-parameters and ψ_2 are the nuisance parameters, (b) the dimension of θ can be very large and (3) the dimension of ψ must be relatively small (not to exceed 20) to avoid an exponential increase in computational time computational time.

2.6 Variational Bayes

The low-rank Variational Bayes correction to the posterior means of a Gaussian latent field is a technique used in Bayesian inference to approximate the posterior distribution of a high-dimensional Gaussian random field, proposed by [7]. In many applications, such as spatial modeling or image analysis, the dimensionality of the data is very large, making the computation of the posterior distribution infeasible.

This technique involves approximating the high-dimensional Gaussian random field using a lower-dimensional approximation, such as a low-rank matrix. This approximation reduces the computational burden of the inference algorithm and allows for faster and more efficient computation of the posterior distribution.

The posterior means of the Gaussian latent field are corrected using the low-rank approximation to improve the accuracy of the inference results. The correction is based on a second-order Taylor series expansion of the posterior distribution, which approximates the curvature of the posterior distribution around its mode.

It can significantly improve the accuracy and efficiency of Bayesian inference algorithms and is the default option in the `R-INLA` package since 2022 (see [8] on how to use the package).

2.7 Model selection criteria: INLA

Model selection is a critical task in statistical analysis, where the goal is to identify the model that best represents the underlying data generating process. Several model selection criteria are used to compare and evaluate different models and choose the one that is most appropriate for the data. In

this context, two widely used model selection criteria are Deviance Information Criterion (DIC, [9]) and Watanabe-Akaike Information Criterion (WAIC, [10]).

DIC is a measure of the goodness-of-fit of a model that takes into account both the complexity of the model and the degree of overfitting. It is computed as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameter estimates. A lower DIC value indicates a better fitting model.

On the other hand, WAIC is a model selection criterion that measures the out-of-sample predictive accuracy of a model. It is based on the Kullback-Leibler (KL) divergence between the predictive distribution of the model and the true distribution of the data. Unlike DIC, WAIC takes into account the entire posterior distribution of the model parameters and adjusts for overfitting by penalizing complex models. A lower WAIC value implies better predictive accuracy of the model.

In summary, while WAIC focuses on the predictive accuracy of a model, DIC considers both the goodness-of-fit and the model complexity. Therefore, both WAIC and DIC are useful tools for selecting the best model among different alternatives.

A new measure has been proposed for model selection in terms of the predictive capacity of the models, which is a *leave-group-out* technique (see [11] for more information). In INLA, this option is achieved by using the `inla.group.cv` function. The higher the value, the better the predictive capacity.

2.8 Classical machine learning

In recent years, there has been a growing interest in modelling and predicting risk in various domains. In the field of machine learning, regression models have been widely used to make predictions on input features. These models aim to find the relationship between the input features and the target variable, using various statistical and mathematical techniques. In this project, we will explore classical machine learning algorithms such as linear regression, Decision Tree, Random Forest or Extreme Gradient Boosting and their application in predicting death rates based on demographic information. Specifically, we will be using the dataset introduced in chapter 3.

Not only will we fit several models, but we will also develop a nested approach to maximize the model's performance. It consists of using different sets of variables, such that in each step, we consider extra information (hopefully useful) to the model. We will do this using *python*.

2.9 Cross validation

Cross-validation is a technique used in machine learning and other areas to assess the performance of a model on an independent dataset. The basic idea is to split the dataset into two parts: a training set, which is used to train the model, and a validation set, which is used to evaluate the model's performance. Cross-validation involves repeatedly splitting the data into different training and validation sets, and averaging the performance over these splits to obtain a more robust estimate of the model's performance.

It works well when the data is independently and identically distributed (i.i.d), meaning that the observations are independent of each other and are drawn from the same distribution. However, it can lead to biased estimates in certain cases, such as when the data has a temporal or spatial dependency structure. In time-dependent datasets, the performance may be better estimated using time-series cross-validation, which takes into account the temporal dependencies between observations. In spatially dependent datasets, leave-one-out cross-validation can be used to ensure that the validation set is not too similar to the training set.

This is our case, so we had to use a type of hand-coded cross-validation since there are no automatic implementations for this type of data. The validation used involves splitting a training set containing

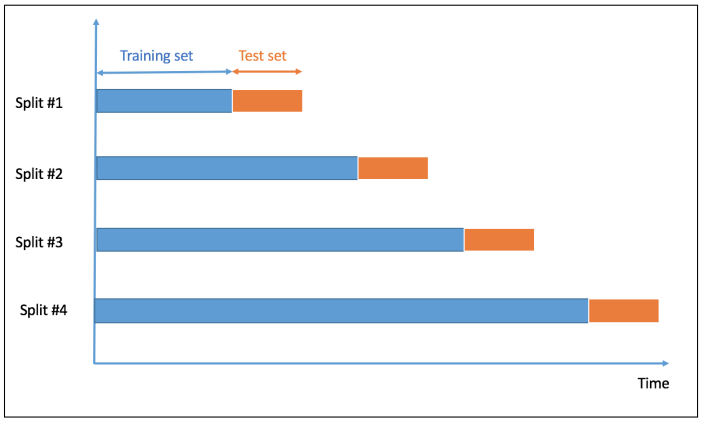

Figure 2.1: Time series split representation.

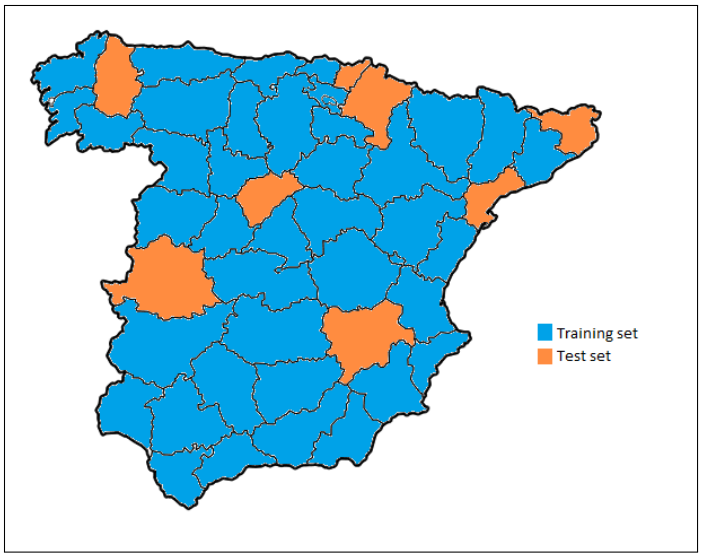

Figure 2.2: Space dependent split representation.

information from certain years and a test set containing information from subsequent years (to avoid training with variables to be predicted in each case). Additionally, we will include certain provinces in the training set and use different ones in the test set (to avoid training the model with data from the same province)."

The resulting cross validation can be explained as the combination of time series split (2.1) with space dependent split (2.2).

2.10 Model selection criteria: ML

Model selection criteria play a crucial role in machine learning as they provide a means of comparing and selecting among different models. One commonly used metric is the R-squared (R^2) value, which is a measure of how well the model fits the data. It is defined as the proportion of variance in the dependent variable that is explained by the independent variables in the model. A value of 1 indicates a perfect fit, while a value of 0 indicates that the model does not explain any of the variance in the dependent variable.

Another widely used metric is the root mean square error (RMSE), which is a measure of the difference between the predicted values and the actual values. It is calculated by taking the square root of the

average of the squared differences between the predicted and actual values. The RMSE provides a measure of the accuracy of the model, and a lower value indicates a better fit.

It is important to note that while R^2 and RMSE are commonly used metrics, they have their limitations. R^2 does not provide any information on the absolute error of the model, and it can be biased towards models with more parameters. On the other hand, RMSE is sensitive to outliers and can be affected by the scale of the data.

2.11 Deep learning

Deep learning models are capable of extracting and identifying intricate patterns and relationships within the data, without the need for explicit feature engineering. In our case, a deep learning model could learn the underlying patterns that link the predictors to the number/rate of deaths. By using deep learning, we could potentially improve the accuracy of our predictions and uncover hidden patterns that would not be captured by traditional machine learning models. However, deep learning models require more computational resources, more training data, and more time to train compared to classical machine learning models.

Deep learning approaches for space-time data [12], including disease mapping, typically involve the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

CNNs are well-suited for analyzing spatial data, such as satellite imagery or maps, and have been used for various applications in disease mapping. For example, researchers have used CNNs to identify areas with high rates of malaria transmission by analyzing satellite imagery of vegetation in [13], which can serve as a proxy for the presence of breeding sites for the mosquitoes that carry the disease.

RNNs, on the other hand, are particularly useful for analyzing temporal data, such as time series data. In disease mapping, RNNs have been used to model the spread of infectious diseases over time and space. For example, researchers have used RNNs to forecast the spread of dengue fever by analyzing spatio-temporal data on climate, mosquito abundance, and reported cases of the disease.

However, despite the spatio-temporal nature of the data, neither CNN nor RNN architectures have been utilized in this study. The data characteristics do not naturally lend themselves to a suitable representation for CNNs, which are primarily designed for image data. Additionally, RNNs have been surpassed by more advanced network architectures such as GRU (Gated Recurrent Unit) or LSTM (Long-Short-Term-Memory).

Moreover, upon discovering the existence of Bayesian Neural Networks (BNNs), there was a strong motivation to explore their implementation in this study, given their potential advantages and relevance to the research topic.

In a nutshell, we will implement three kinds of neural networs: MLP, LSTM and BNN and compare them.

Chapter 3

Exploratory data analysis

3.1 Introduction & Preprocessing

The present exploratory data analysis (EDA) focuses on a dataset of lung cancer mortality in Spain, spanning from 1999 to 2020, for each province year by year. The study is motivated by the importance of gaining insight into cancer mortality patterns and trends, and the relevance of understanding such trends for public health and policy-making purposes, as well as to serve as an example on which to apply the techniques under study.

Cancer continues to be a major public health concern worldwide, and lung cancer, in particular, is a leading cause of cancer mortality in Spain. With approximately 20% of all cancer-related deaths attributed to lung cancer [14], it is critical to investigate the factors contributing to the spatio-temporal distribution of lung cancer mortality rates across the country. While our study does not directly investigate risk factors, our contribution lies in enhancing the comprehension of the temporal evolution of geographical patterns. By doing so, we aim to provide insights that may help identify potential risk factor. The EDA will involve applying data analysis techniques to examine patterns and trends in cancer mortality data in Spain.

Through the investigation, we aim to identify potential hotspots associated with lung cancer mortality in Spain, which can inform the development of targeted public health interventions and policies.

The top 5 rows of the provided data are show in Table 3.1.

Table 3.1: Top 5 rows of the original lung cancer data.

	PROV	ANO	SEX	EDADGR	O	Pop
0	1	1999	1	1	0	5454.0
1	1	1999	1	10	7	10292.0
2	1	1999	1	11	3	10216.0
3	1	1999	1	12	15	7930.0
4	1	1999	1	13	13	7375.0

The dataset contains the following columns:

- PROV: province's ID (from 1 to 50) corresponding to their zip codes,
- ANO: year (from 1999 until 2020),
- SEX: wheter mortality refers to males (SEX=1) or females (SEX=6),
- EDADGR: age group to which individuals belong (from 1 to 18, where the group 1 contains people of 0-4 years, the group number 2 from 5-9 etc.)
- O: number of observed cases (deaths),
- Pop: population.

Table 3.2: Top 5 rows of the preprocessed lung cancer dataset.

	ProvinceID	Year	Sex	AgeGr	Obs	Population	Risk	Exp	Rate
0	1	1999	Man	1	0	5454	2.2385e-07	0.0012	0.00
1	1	1999	Man	10	7	10292	2.7430e-04	2.8231	0.68
2	1	1999	Man	11	3	10216	6.2067e-04	6.3408	0.29
3	1	1999	Man	12	15	7930	1.1574e-03	9.1782	1.89
4	1	1999	Man	13	13	7375	1.8102e-03	13.3502	1.76

We have done some preprocessing work, as well as calculations to obtain the expected number of cases for each row and the SMR, according to the expressions in section 2.3. Results are presented in Table 3.2. We wanted to express the gender with its *string* representation, change some of the column names and make sure the data types were appropriate.

Apart from the SMR, we have calculated the rate per thousand individuals, defined as $Rate_{ijt} = \frac{O_{ijt}}{Population_{ijt}} \times 1000$ because of the following reasons:

- Interpretation: The rate per thousand is more straightforward to interpret compared to the SMR, which requires understanding of the reference population.
- Comparability: The rate per thousand allows for easier comparisons between groups or over time, as the denominator (population at risk) is constant. In contrast, the SMR requires adjusting for differences in the reference population, which can complicate comparisons.
- Precision: The rate per thousand may be more precise in some instances as it is based on actual population denominators, whereas the SMR relies on estimated denominators.

Therefore, from now on the target variable will be the $Rate$. After these steps, we merged the Table 3.2 with the *spanish cartography dataset*, performing a left join by *ProvinceID* to get the region names and the geometry multi-polygons (provinces' boundaries). The table is not presented in this work because of its dimensions.

Once the data is in the appropriate format, exploratory analysis begins.

3.2 Data analysis

Data analysis is a crucial component in making informed decisions and understanding complex phenomena in a variety of fields. In this study, we analyze the previously introduced dataset to gain insights into the relationships and patterns among different variables. Through the use of descriptive statistics and data visualization techniques, we aim to uncover hidden trends and correlations that may be useful for our objectives.

To begin our analysis, we first investigate how the individuals in the dataset are distributed according to their gender and age. To achieve this, we will determine the proportion of deaths of male and female subjects, as well as the distribution of deaths across different age groups. The respective plots are presented in Figure 3.1.

One can observe huge differences regarding gender with 75% of deaths corresponding to males. If we focus on age groups, lung cancer mortality is nearly nonexistent in people younger than 40. For the older age groups, the proportion of deaths is similar.

To get an idea of the distribution of each variable, we describe data through numerical summaries, shown in Table 3.3.

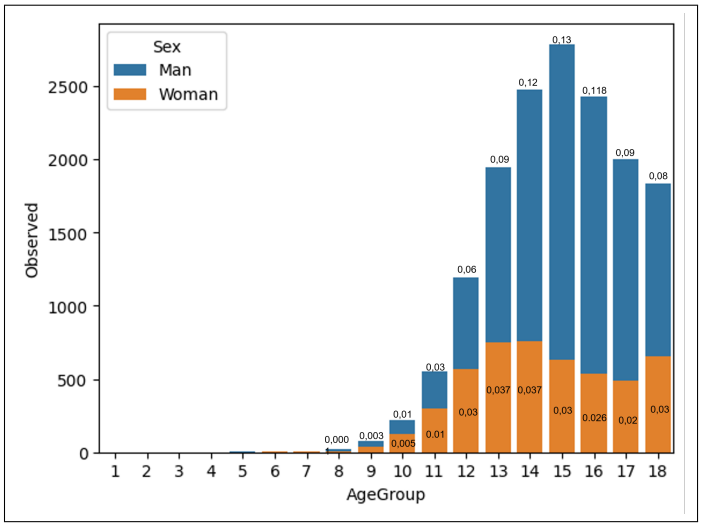

Figure 3.1: Proportion of deaths of each sex and age group.

Table 3.3: Numerical summaries of the variables in our dataset.

	min	max	mean	sd	cv
Observed	0.0	401.0	11.24	30.12	2.68
Population	1.23e+03	321933.00	24772.27	34946.26	1.41
Expected	6.74e-05	425.76	11.24	29.31	2.61
Rate	0.0	8.65	0.67	1.28	1.91

When the coefficient of variation (cv) is above 1, it indicates a relatively high degree of variability or dispersion in the dataset compared to the mean and suggests that the standard deviation is larger than the mean, indicating that the data points are spread out over a wider range.

The mean and standard deviation of the Observed variable suggest that there is quite a bit of variability in the number of observed cases. The standard deviation is larger than the mean, indicating that the distribution is likely skewed to the right.

The mean and standard deviation of the Population variable suggest that the population size also varies quite a bit, and there are likely some very large values that are driving up the standard deviation.

Expected cases have a similar mean to the Observed variable, but a smaller standard deviation. This indicates that the expected number of cases is more consistent across the population than the actual number of observed cases.

Finally, the rate has a mean of 0.68 cases per 1000 inhabitants and a standard deviation of 1.28. In this case, the highest value is 8.66 cases per 1000 inhabitants.

In Figure 3.2 we show how rates evolved for men and women between 1999 and 2020. Males exhibit a declining tendency, whereas females display a contrasting pattern.

Despite the fact that lung cancer deaths are primarily among men, there has been a decreasing trend in this group in recent years. On the other hand, women, who historically had a lower number of deaths, have experienced a consolidated upward trend. This trend may be attributed to the fact that in the past, the majority of women did not smoke. However, with significant advancements in gender equality and personal rights, women have started smoking, leading to an increase in lung cancer cases. Additionally, a notable discrepancy in the magnitude of the variable is evident between the two sexes.

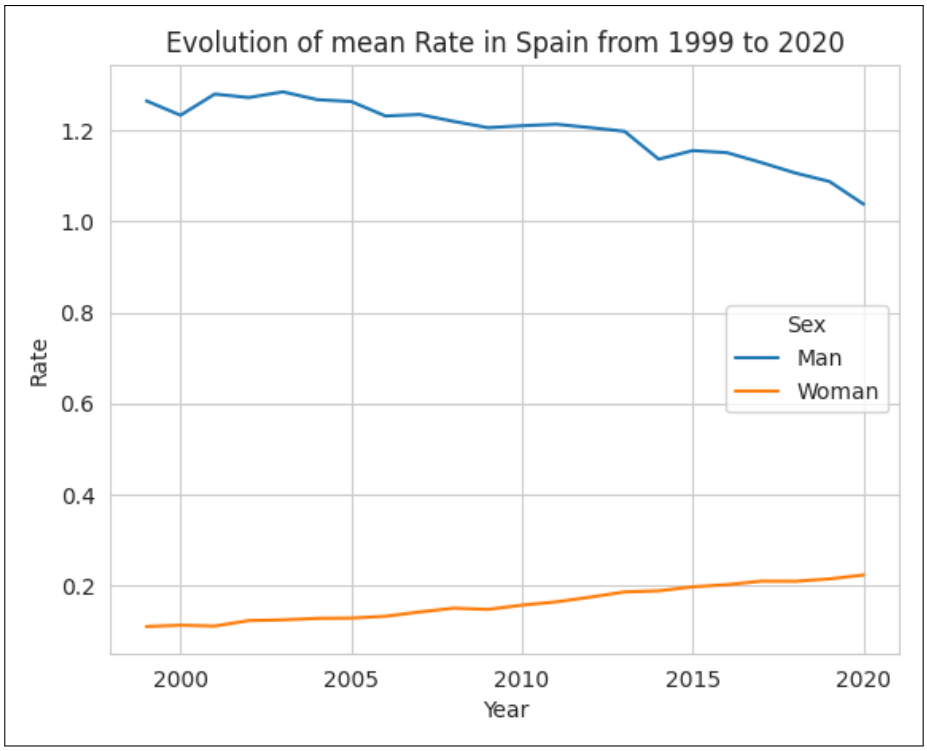

Figure 3.2: Evolution of mean Rate by sex.

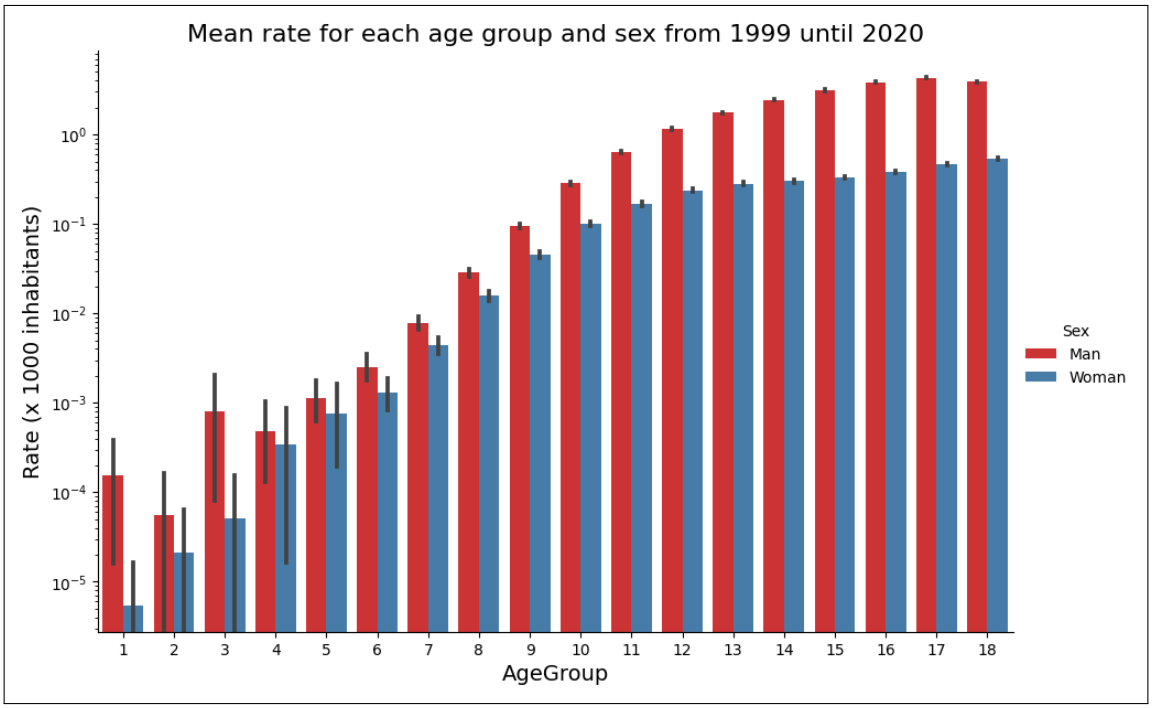

Figure 3.3: Mean Rate for each age group and sex from 1999 until 2020.

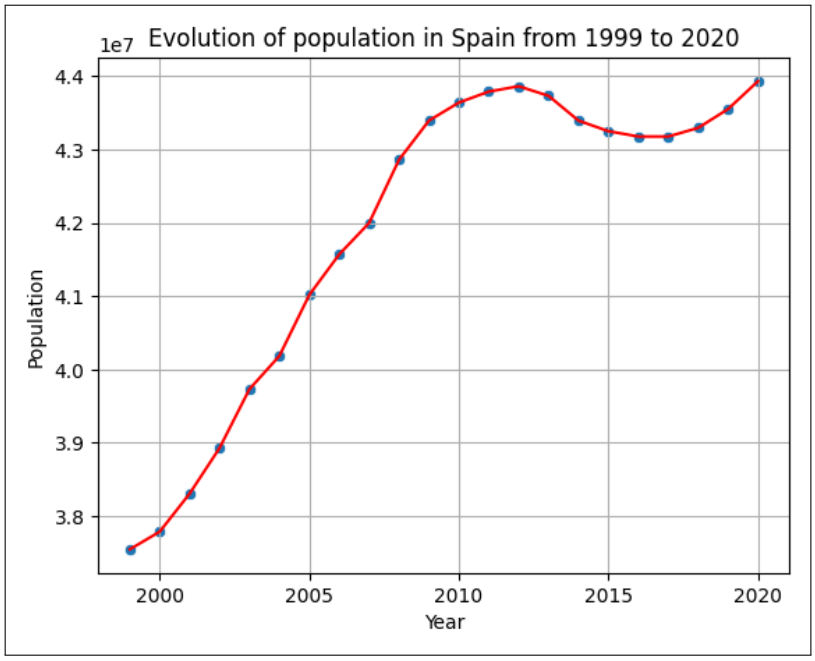

Figure 3.4: Evolution of Spanish population between 1999 and 2020.

Regarding Figure 3.3, the lower age groups present higher deviations from the mean (see error bars). This is due to the fact that a single death in early ages has a bigger impact in the mean that one death among the older ones, because close to 0 rates are expected for the youngest. We must also take into account that people in higher age groups are more likely to die, not just because of cancer but from other diseases too. That is why, even if there is less population in older ages, rates remain high.

We present a plot in Figure 3.4 to see the population's temporal evolution. If the rate remained constant, we would expect a higher absolute number of deaths.

We are now interested in identifying the regions and years with a higher rate. The heatmap displayed in Figure 3.5 makes it easy to find this information. The results in the rate heatmap are consistent (time neighbors exhibit similar rates) and is hard to find extreme values. Overall, the later the year, the smaller the rate is. The darkest region is Extremadura, which means the rate is higher there.

To compare multiple models and mitigate the computational cost, dimensionality reduction techniques are necessary. We employed a straightforward clustering method called DBSCAN on the entire population, considering only age groups and the proportion of deaths within each group.

By clustering the data into 5 groups, we selected 2 clusters with a sufficient number of non-zero observations. Age groups 14 and 15 were merged into one cluster (65-74 years), while age groups 16 and 17 were grouped together in another cluster (75-84). This allowed us to simplify the analysis by grouping up these individuals into single groups. Therefore, from now on, we will work with these four groups:

- Group 1: Men, from 65 to 74 years,
- Group 2: Women, from 65 to 74 year,
- Group 3: Men, from 75 to 84 years,
- Group 4: Women, from 75 to 84 years

For the selected groups, the evolution of the rate is shown in Figure 3.6. As expected, the ones corresponding to men are decreasing, while the trends for women are increasing.

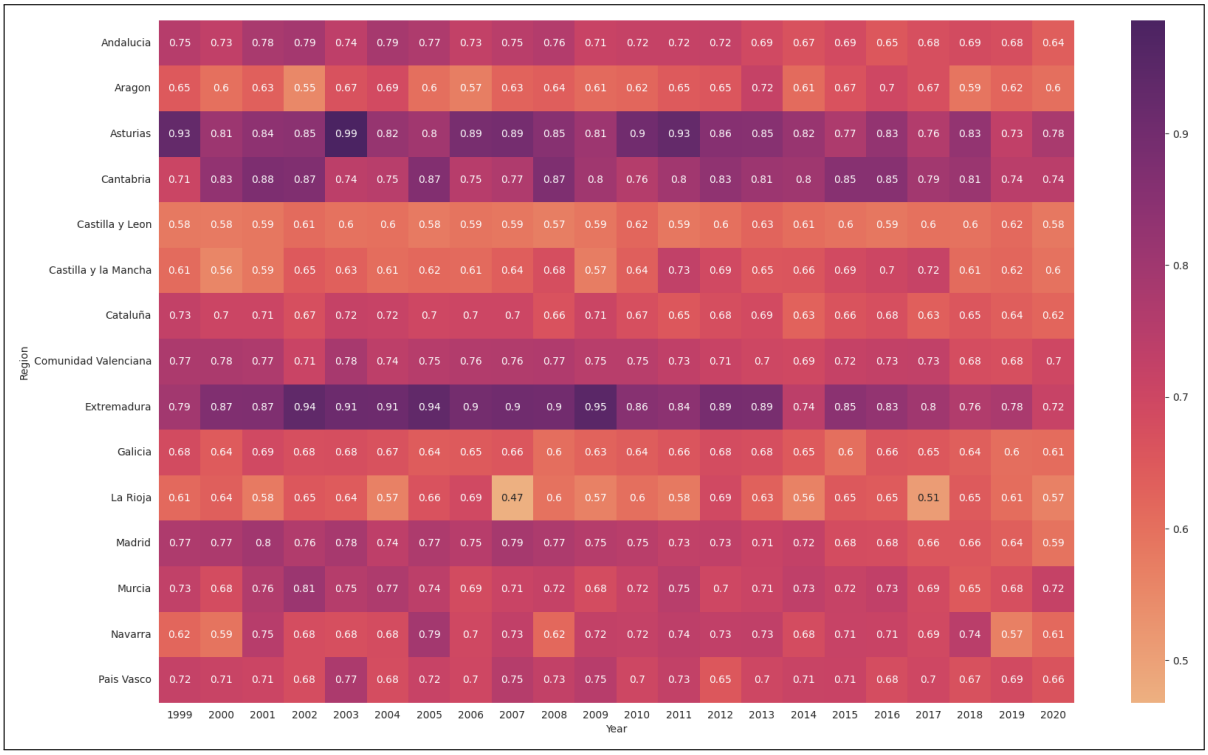

Figure 3.5: Heatmap of Rate for each region from 1999 to 2020.

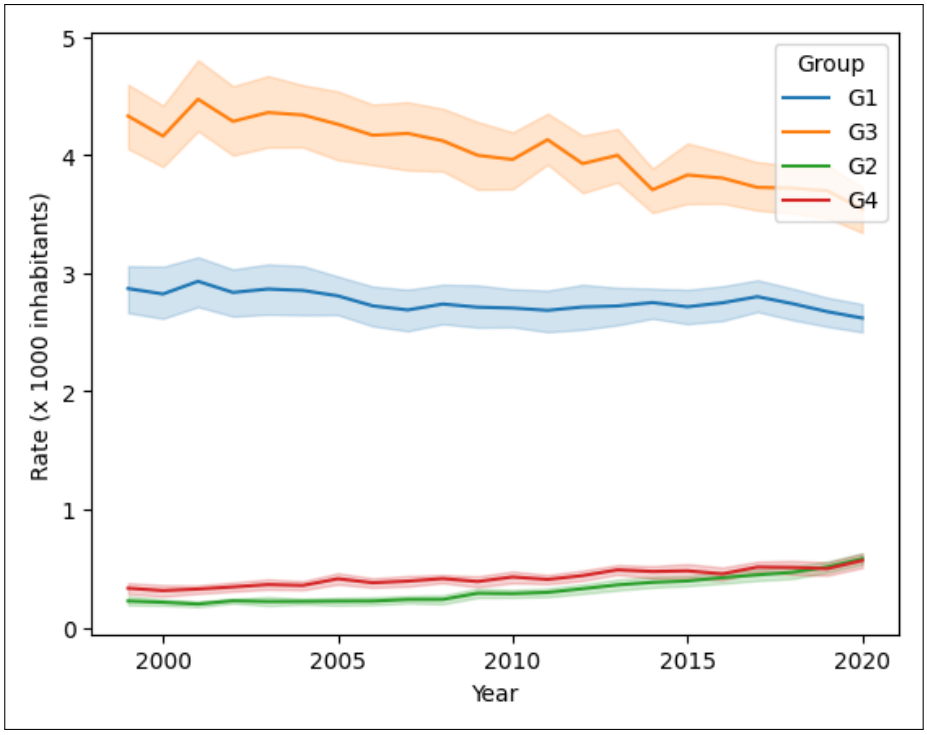

Figure 3.6: Evolution of mean rate by groups.

Another way of visualizing this is through map plots. Figures 3.7, 3.8, 3.9 and 3.10 help us identify areas with higher and lower rates at a glance. Finally, to account for the variability of each province, we found interesting to show 3.11, 3.12, 3.13 and 3.14, where the rates for each province are shown in different lines and the overall mean is presented in blue. Despite the visible peaks and valleys observed for each province and the variability among them, the overall trend within each group is clear. If we pay attention to Figure 3.12, we can see that there are some provinces with rates higher than the mean and others that are below the mean rate. Some lines exhibit higher variability than others, and probably belong to lower populated areas.

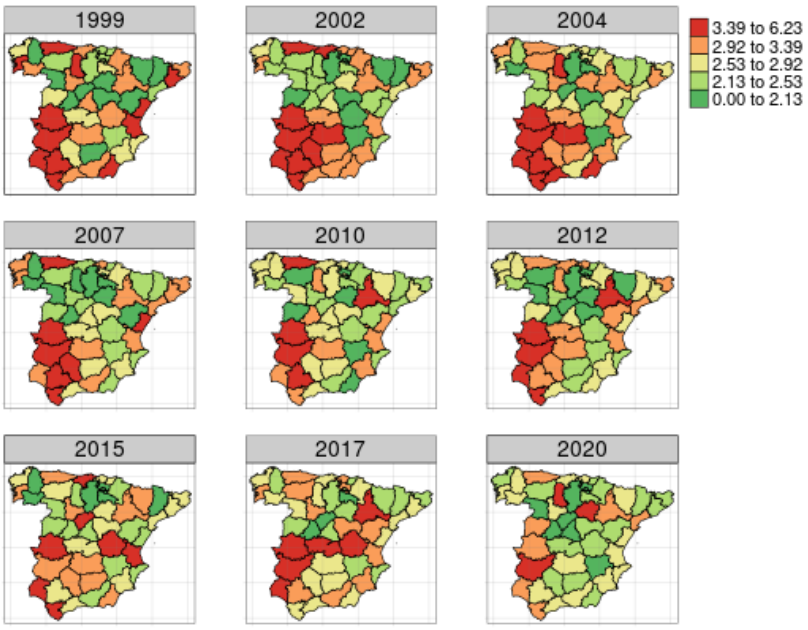

Figure 3.7: Evolution of mortality rates per 1000 inhabitants for men between 65 and 74 years, group 1.

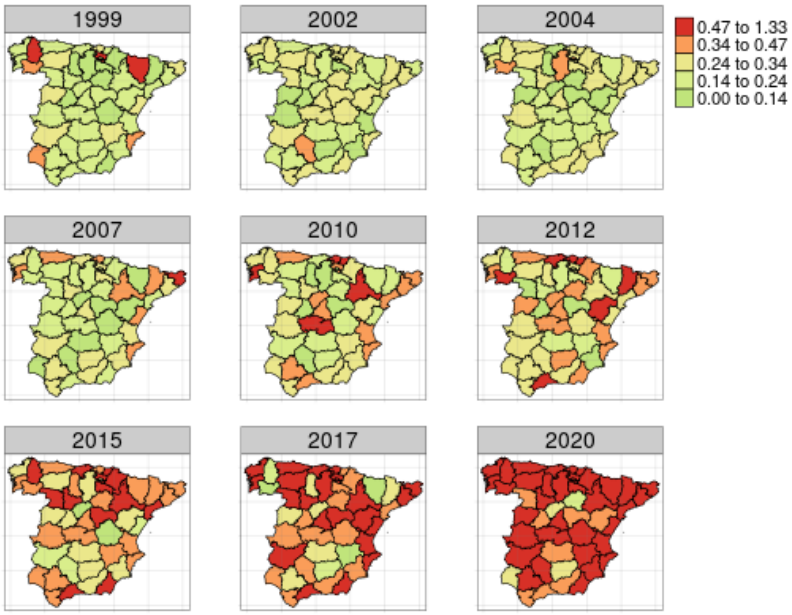

Figure 3.8: Evolution of mortality rates per 1000 inhabitants for women between 65 and 74 years, group 2.

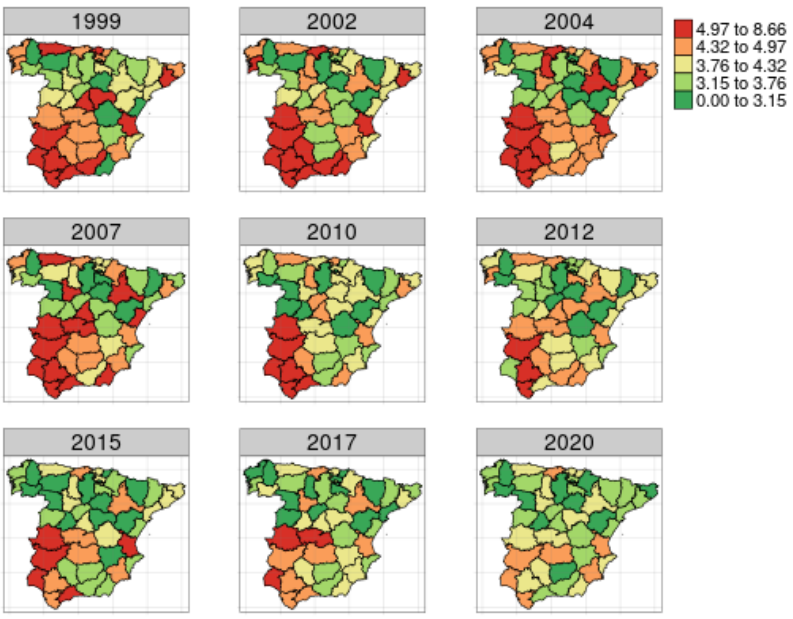

Figure 3.9: Evolution of mortality rates per 1000 inhabitants for men between 75 and 84 years, group 3.

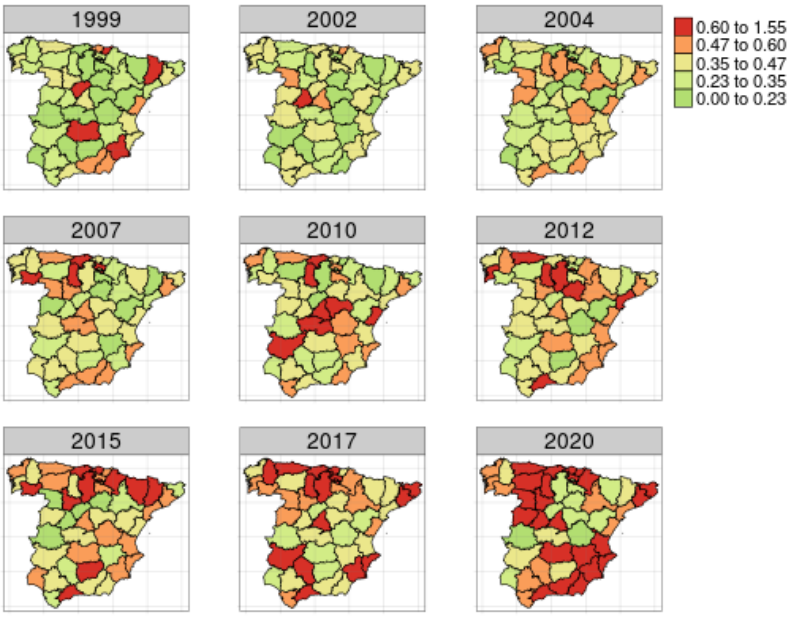

Figure 3.10: Evolution of mortality rates per 1000 inhabitants for women between 75 and 84 years, group 4.

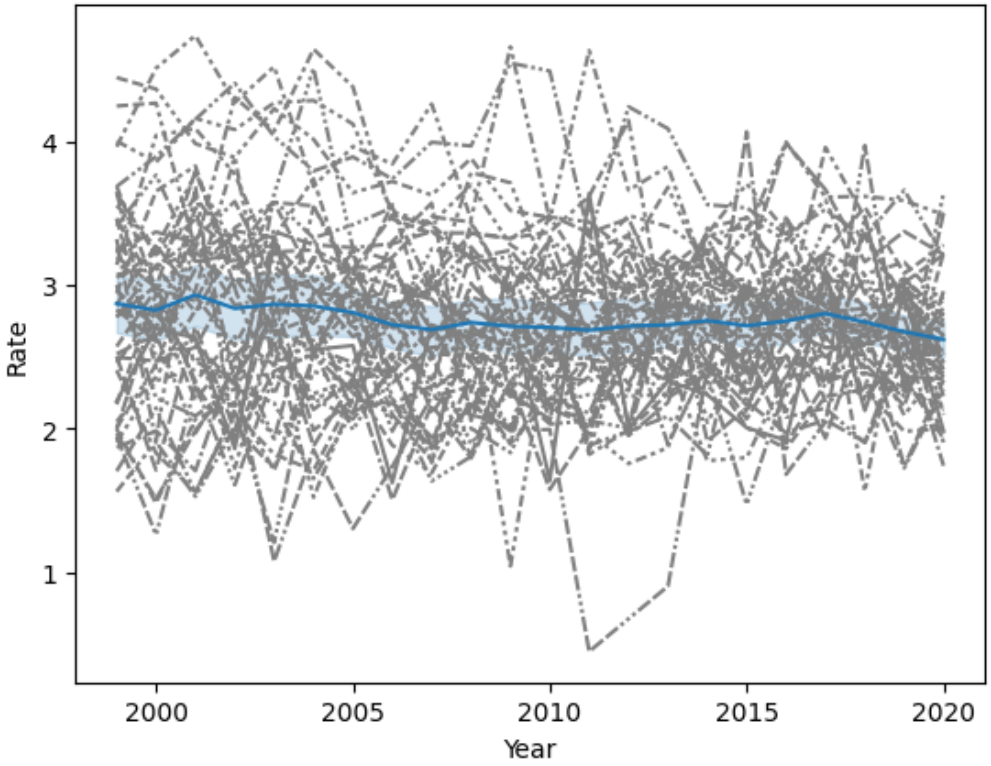

Figure 3.11: Evolution of mortality rates per 1000 inhabitants for men between 65 and 74 years for each province. The blue line indicates the mean rates per year.

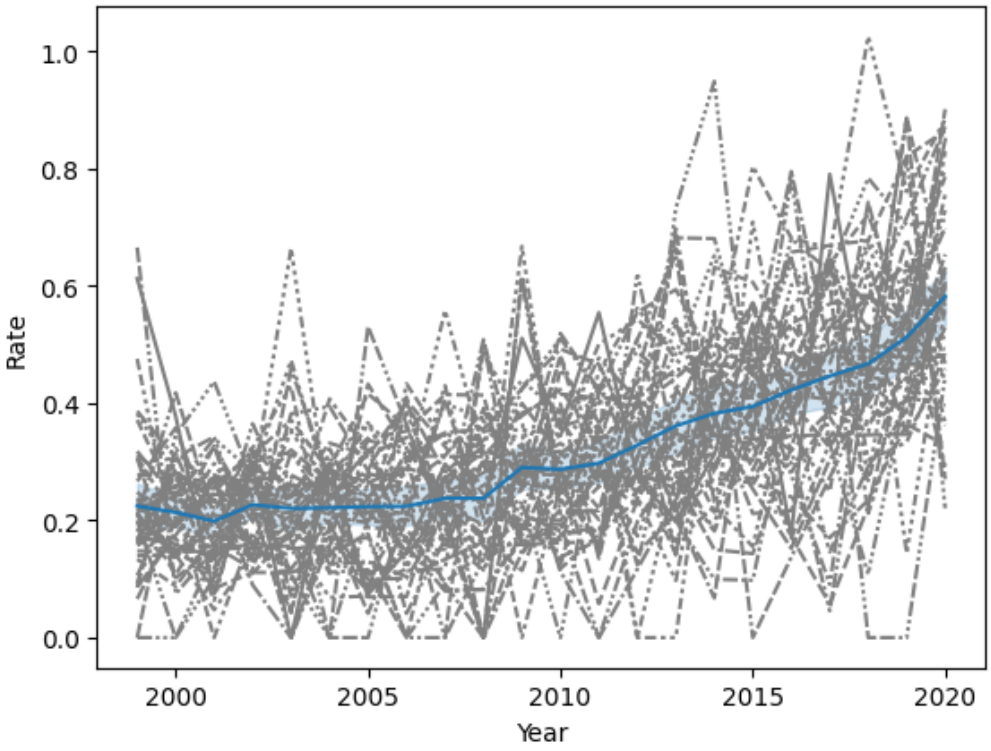

Figure 3.12: Evolution of mortality rates per 1000 inhabitants for women between 65 and 74 years for each province. The blue line indicates the mean rates per year.

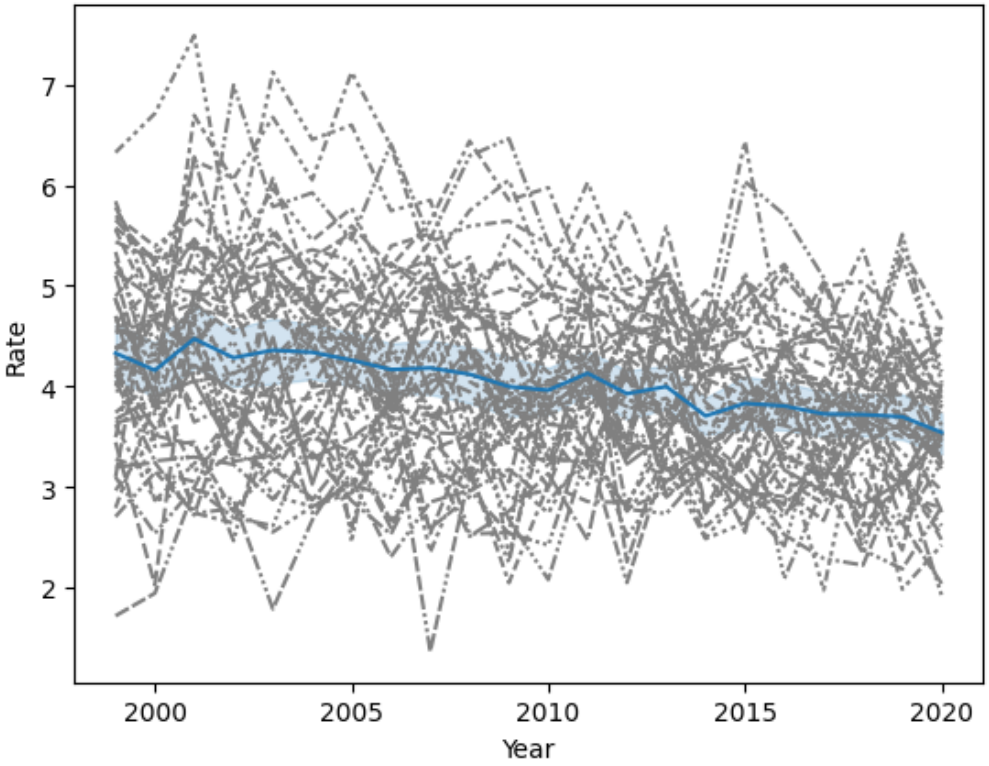

Figure 3.13: Evolution of mortality rates per 1000 inhabitants for men between 75 and 84 years for each province. The blue line indicates the mean rates per year.

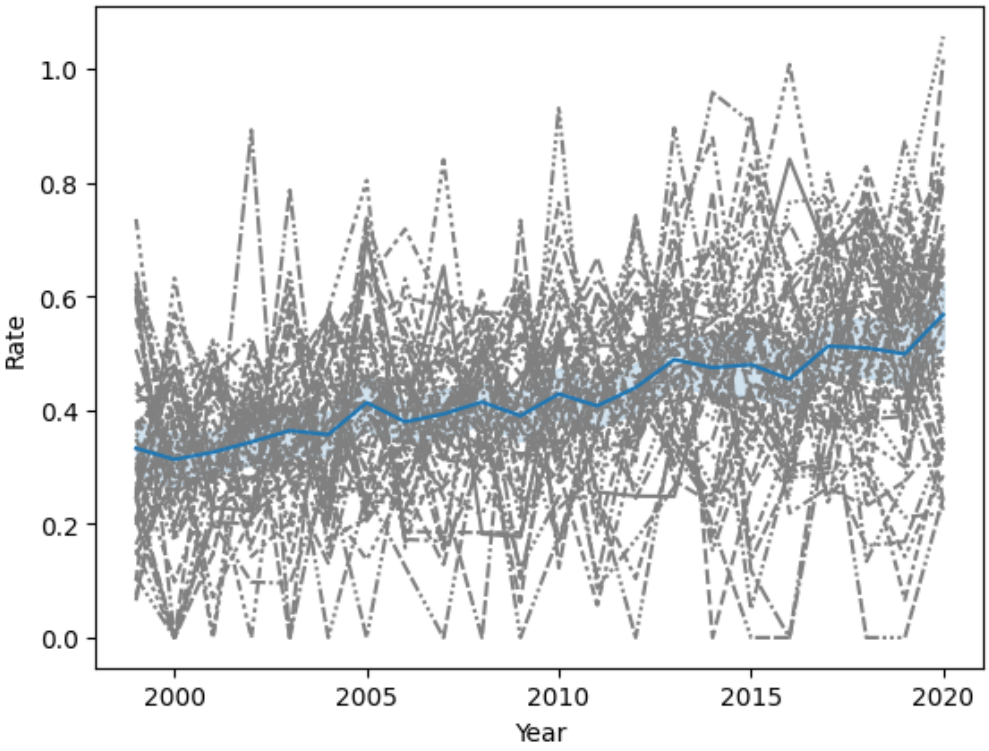

Figure 3.14: Evolution of mortality rates per 1000 inhabitants for women between 75 and 84 years for each province. The blue line indicates the mean rates per year.

Chapter 4

Rate modelling

4.1 INLA models

This section details the INLA-fitted models for each of the four strata of interest, namely: gender (male/female) and age group (65-74/75-84).

The present study uses the BigDM library developed by Spatial Statistics UPNA, which features a function with adjustable parameters for fitting the desired model. This library is built upon the widely used `R-INLA` package and has been optimized to effectively analyze large datasets by utilizing disjoint models or k-order neighborhoods instead of the global one.

4.1.1 Fitted models & results

We fitted 4 models for each stratum / group, for different types of interaction [15]. Results are shown in Table 4.1.

As explained before, we make decisions based on DIC and WAIC, primarily. Both criteria yielded the same results. The selected interactions are presented in bold in Table 4.1. For the male groups the selected interaction type was II, which means that each area has a temporal correlation structure, but neighboring areas have independent temporal correlations, i.e., trends of adjacent provinces don't have to be similar. On the other hand, female groups obtained better DIC values with interaction type IV. That is, at each time point and for contiguous periods, there are spatial correlations, and vice-versa. Note that, for group number 2, the DIC and WAIC are the same for interaction types II and IV, but we selected the second option because of the leave-group-out value, which is slightly higher for type IV.

Plotting the rate versus the posterior median estimate, we obtain Figure 4.1 (for the subsequent models, we will show the predictions for the 4 groups in the same plot, but the behavior is the same). We did not expect to recover the observed rates (we do not believe they are true) and we expect some shrinkage due to the smoothing of the model. In group number 4, rates are more unstable and INLA smoothes more, which allows us to see the temporal trend.

The posterior median estimates are shown, in stratum order, in Figures 4.2, 4.3, 4.4 and 4.5. We must keep into account, that this approach gives us the full posterior distribution, not only point estimates. This is a huge advantage of this method and will keep an important role in interpretation and comparison with the two other proposed model families.

In Figures 4.6, 4.7 and 4.8, crude rates are shown against posterior median estimates together with the credible intervals for each group, in Madrid, Navarra and Valencia, respectively. INLA captures the tendencies and credible intervals are good at reflecting the variations in the data. As expected, men were easier to adjust than women, because of the amount of available data.

As the EDA showed, trends for men are decreasing, while those for women have been increasing in the last years. Note that the legends vary from one sex to the other, because scales for these two groups are still very different. These plots are also useful in identifying whether a specific time period

Table 4.1: Results for INLA models fitted to each stratum and interaction type.

Group	Interaction	DIC	WAIC	N. parameters	Leave-group-out
1	Type I	7639	7636	294	0.0322
	Type II	**7526**	**7535**	157	0.0339
	Type III	7639	7655	231	0.0331
	Type IV	7531	7544	151	0.0342
2	Type I	5264	5265	112	0.1051
	Type II	5249	5253	94	0.1056
	Type III	5268	5271	86	0.1055
	Type IV	**5249**	**5253**	81	0.1058
3	Type I	7418	7417	179	0.0354
	Type II	**7361**	**7361**	130	0.0362
	Type III	7429	7429	126	0.0358
	Type IV	7364	7364	131	0.03623
4	Type I	5352	5352	79	0.0973
	Type II	5348	5349	69	0.0974
	Type III	5352	5353	65	0.0975
	Type IV	**5345**	**5347**	68	0.0975

exhibits a higher (above the median) or lower (below the median) than expected rate. Provinces with higher populations (Madrid) tend to have smoother crude rate values, whereas low populated areas (Navarra and Cáceres) exhibit strong changes in consecutive years. The credibility intervals provided by INLA are, therefore, more accurate as the areas become more populated.

Comparing these maps with the ones presented in Chapter 3, we can see that the posterior median estimates are "smoother" that the original rates, not only in space, but also in time. It seem clear that Extremadura is one of the regions where the mortality risk is higher for male groups, whereas for females, the zones with higher risks are at the coast.

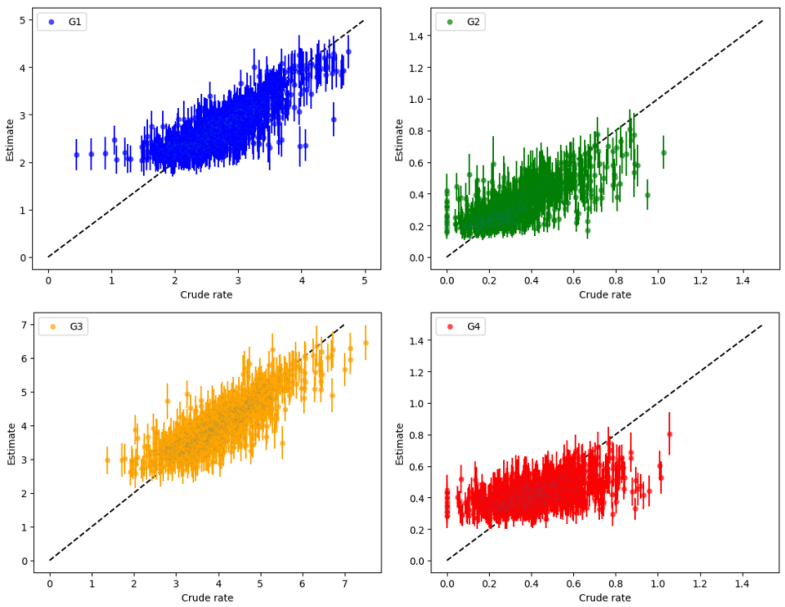

Figure 4.1: Posterior median estimates for the rate with the corresponding 95% credible intervals for each group.

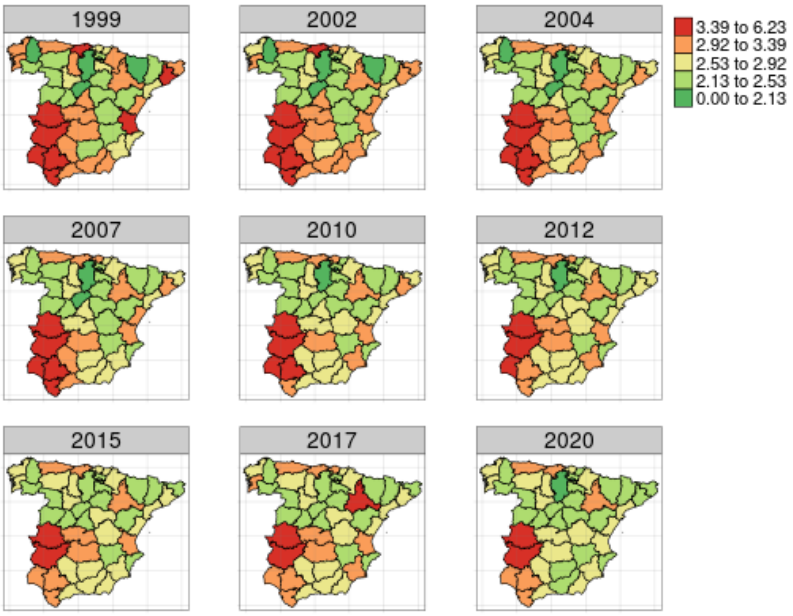

Figure 4.2: Posterior median estimates for the rates for men between 65 and 74 years.

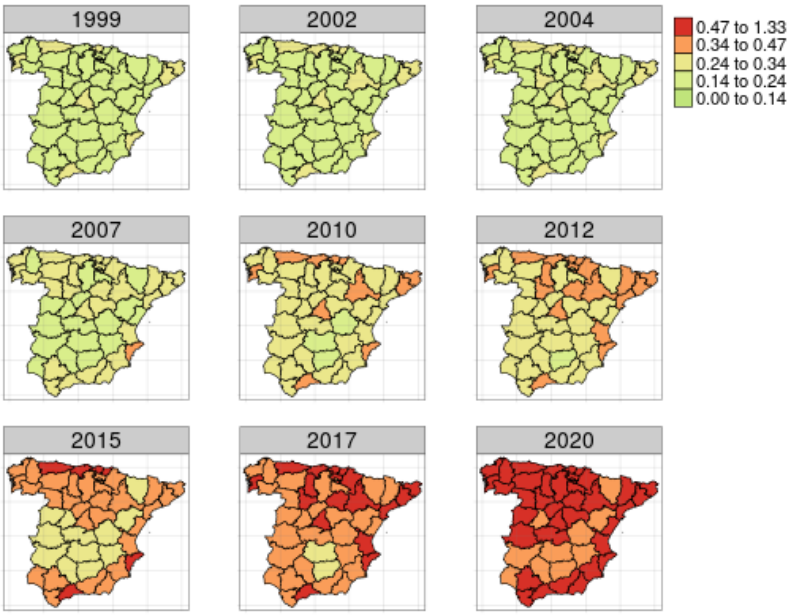

Figure 4.3: Posterior median estimates for the rates for women between 65 and 74 years.

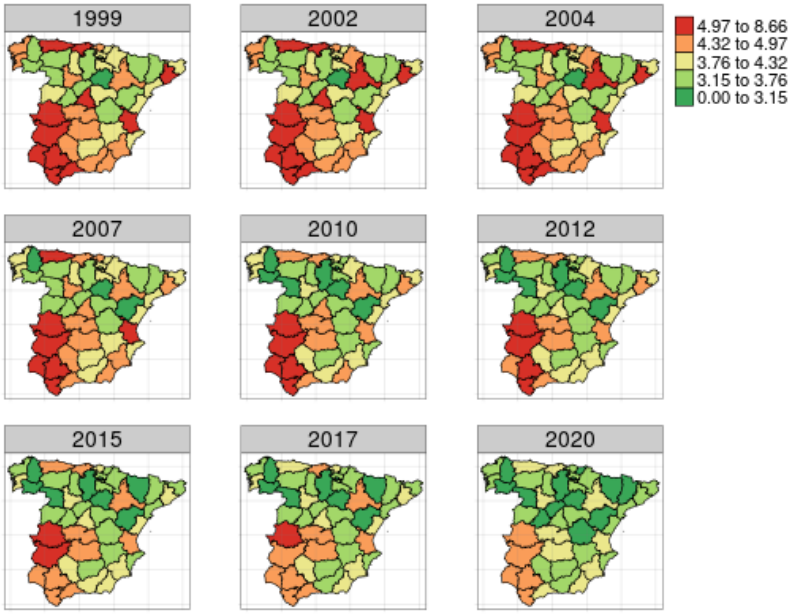

Figure 4.4: Posterior median estimates for the rates for men between 75 and 84 years.

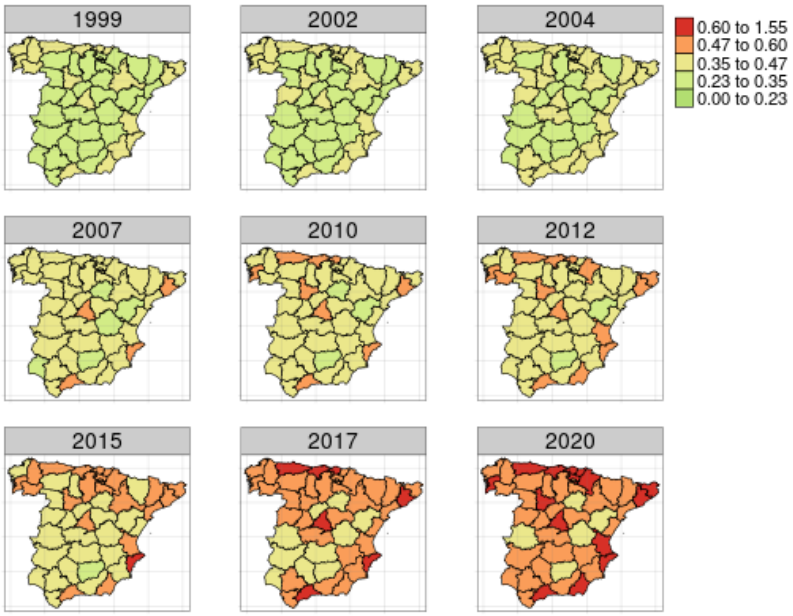

Figure 4.5: Posterior median estimates for the rates for women between 75 and 84 years.

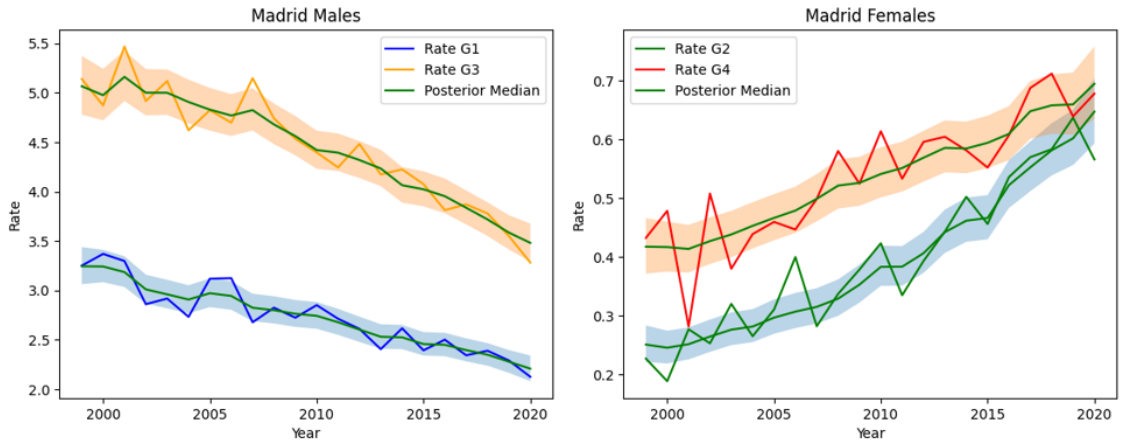

Figure 4.6: Evolution of rate and posterior median in Madrid, from first to last year, with the corresponding credible intervals.

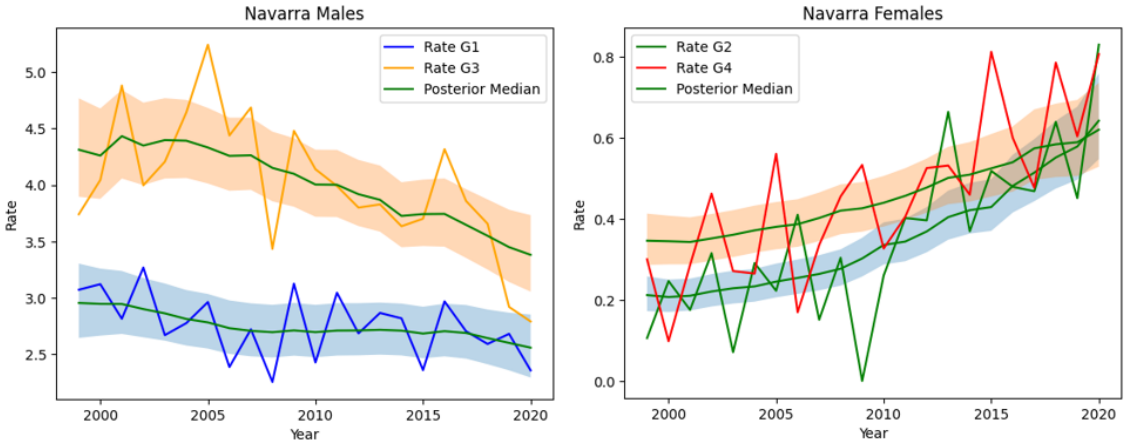

Figure 4.7: Evolution of rate and posterior median in Navarra, from first to last year, with the corresponding credible intervals.

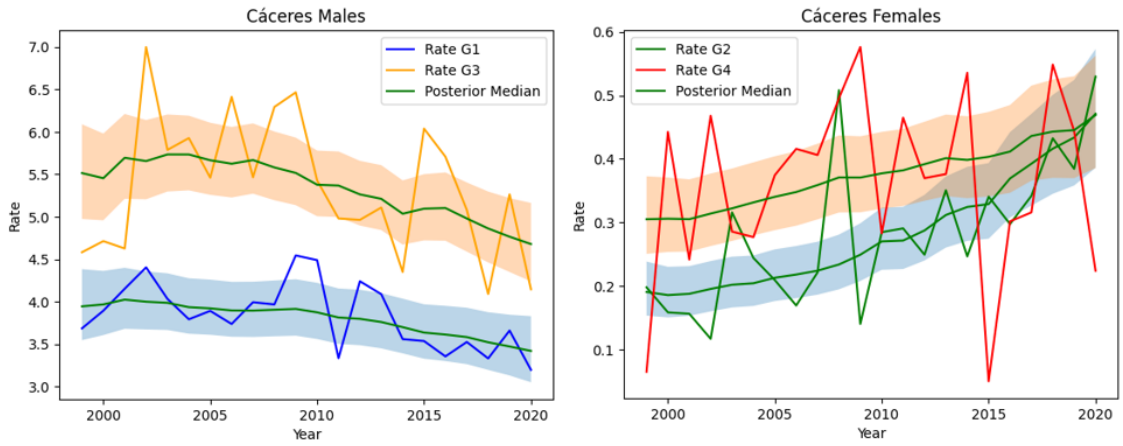

Figure 4.8: Evolution of rate and posterior median in Cáceres, from first to last year, with the corresponding credible intervals.

4.2 Classical machine learning

4.2.1 ML models

Machine learning models in this study have been implemented in *python*, using the `scikit-learn` package. There are plenty of available models in this library and therefore, we decided to use the most popular ones. We studied the behavior of each model with different predictor variables, by applying feature engineering[1].

4.2.2 Fitted models & results

Here we present the list of predictor variables used to fit each model and the obtained results.

1. Encoded province and population.
2. Encoded province, population and sex.
3. Encoded province, population, sex and age group.
4. Encoded province, population, sex, age group and year.
5. Population, sex, age group, year, longitude and latitude of the centroids of each province.

   We obtained the information from the cartography dataframe.
6. Population, sex, age group, year, x and y coordinates of the centroids of each province.

   We transformed the longitude and latitude values to x and y coordinates of a plane. The representation of these centroids is shown in Figure 4.9.

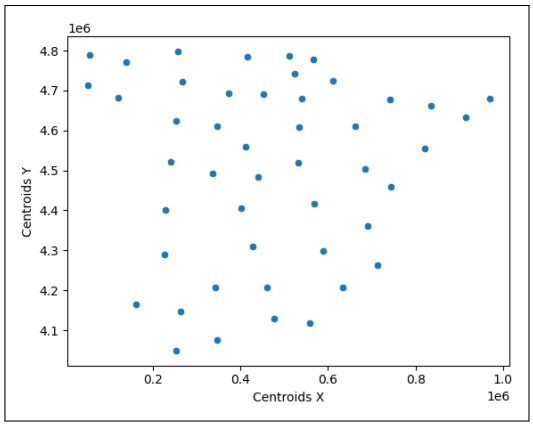

Figure 4.9: Centroids of each province in Cartesian coordinates.

7. Population, sex, age group, x and y coordinates and temporal lags.

   We created a function which adds columns to our dataset. These columns are the previous years' rates (number of years decided by the data scientist), for the same province and stratum. We used 2 columns to take into consideration the two previous years for each observation. This aims to capture the time dependencies in data.

   Notice that for the first 2 years, there is no information and thus, we have missing values for them. We could apply some imputation techniques, but we didn't want to include a bias in our data, so we ended up removing the first 2 years of observations.

   A glance at these temporal lags is shown in Figure 4.10.

[1] Feature engineering refers to the process of using domain knowledge to select and transform the most relevant variables from raw data when creating a predictive model using machine learning or statistical modeling.

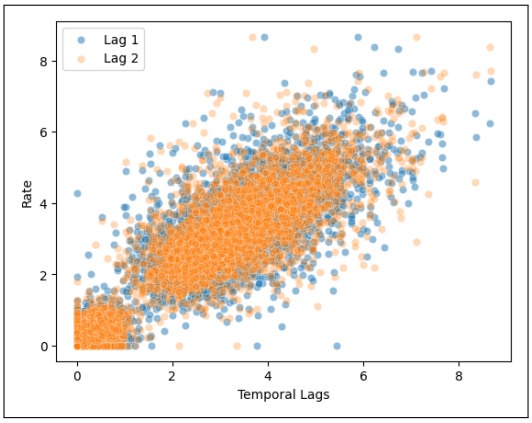

Figure 4.10: Temporal lags 1 and 2 versus Rate.

8. Population, sex, age group, x and y coordinates, temporal lags and spatial lags.

   We have defined a new variable named *spatial_lag* (see [16]), to account for the spatial auto-correlation. It is calculated based on the adjacency matrix and calculates the average of the rates of the adjacent provinces. It also helps reduce noise.

   Spatial lag of order 1 (considering only adjacent provinces) presents a positive correlation with the original Rate. See figure 4.11.

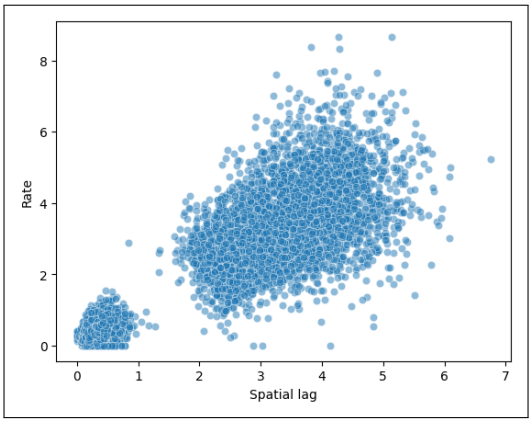

Figure 4.11: Spatial lag versus Rate.

9. Population, sex, age group, x and y coordinates, temporal lags, spatial lags and spatio-temporal lags.

   Incorporating a spatio-temporal lag is a natural approach to account for the potential interplay between temporal and spatial auto-correlations. It is the same as the spatial lag but for previous years.

   The spatio-temporal lag is shown in Figure 4.12 and looks similar to the *spatial_lag*.

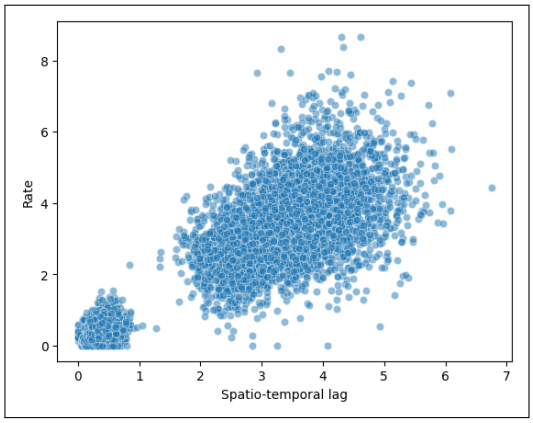

Figure 4.12: Spatio-temporal lag versus Rate.

10. Population, sex, age group, x and y coordinates, temporal lags, spatial lags and spatio-temporal lags and basis functions for each variable.

    Basis functions are a set of functions used in machine learning to transform input data into a new feature space. This technique can improve model performance by capturing complex patterns in the data, reducing dimensionality, and increasing flexibility.

    In this work we included polynomial features, gaussian basis and sigmoid-like basis.

Each predictor subset has been trained with the above mentioned models using the *grid search* strategy with the cross validation explained before. This is the way we fine-tuned the hyperparameters too. The full training of all models had an estimated time of 15h.

These models were selected based on their specific features: linear regression and decision tree models offer interpretability, random forest models enable the estimation of confidence intervals (as we will see later), and XGBoost is well-known for its exceptional performance.

The final model selected, together with the preprocessing pipeline, can be represented as in Figure 4.13.

It is important to highlight that due to the striking similarity in results among the four regression models used, one might be inclined to favor the decision tree model, given its high interpretability. However, it is noteworthy that the predictions from this model appear to be rather unusual (for the problem, not for a decision tree), as demonstrated in Figure 4.14.

The results are presented in Table 4.2 and a nice way to visualize what we obtained is presented in Figures 4.15 and 4.16.

We achieved a better fit by incorporating the information we believed would be useful. Our models didn't have to face to overfitting because we selected the hyperparameters with the cross validation explained before. It is worth noting that for models 7 and beyond, the differences were relatively small. This could indicate that lung cancer does not exhibit significant spatial autocorrelation or that its effect is really small, at least for province lattice data.

The use of geographic coordinates in the modeling process has yielded interesting results. Surprisingly, the longitude and latitude variables do not seem to have a strong relationship with the response variable. However, the inclusion of X and Y coordinates has been found to improve the model, as evidenced by the lower RMSE and higher R^2 values, though the difference is small.

We must keep in mind that, even if the *lag* variables do not seem to improve the model by much, they are variables substituting the previous encoded $ProvinceID$ and $Year$ variables and though, seem to capture the essence of the autocorrelations. Ultimately, the basis functions had almost no effect on the performance of the model.

Upon testing the aforementioned techniques, it appears that we have encountered the limits of model fitting given the available data. It is possible that the inclusion of additional covariates, such as the proportion of smokers in each group, could enhance the model's performance and mitigate the variability of the fitted values, but we managed to improve the quality from the first to the last model including knowledge in the form of covariates

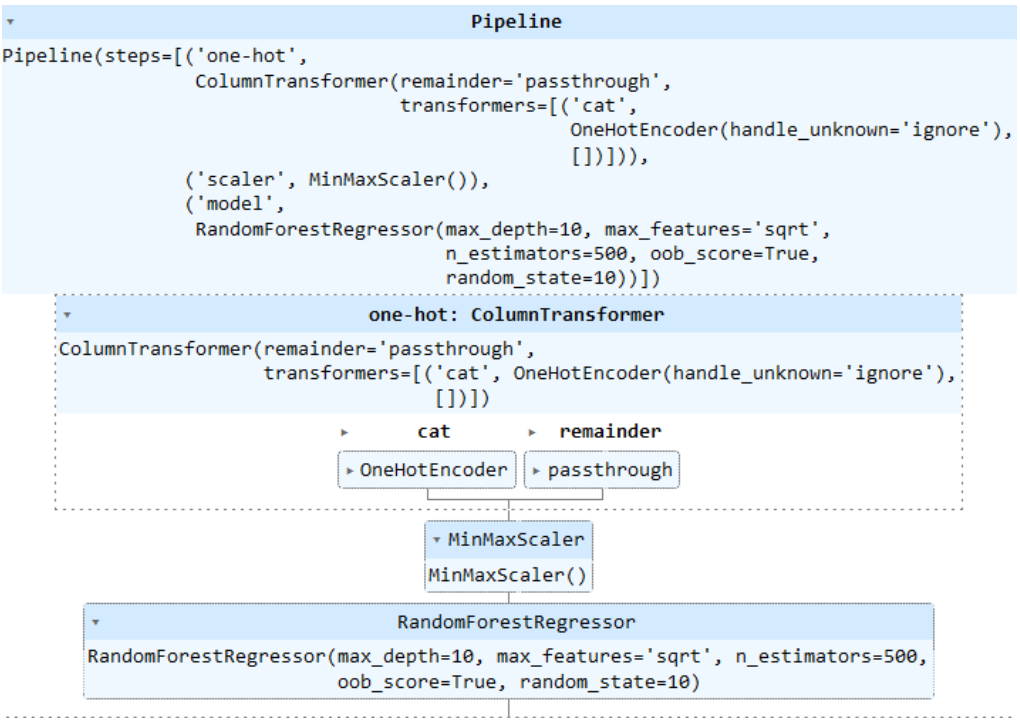

Figure 4.13: Preprocessing and modeling pipeline scheme for the selected model.

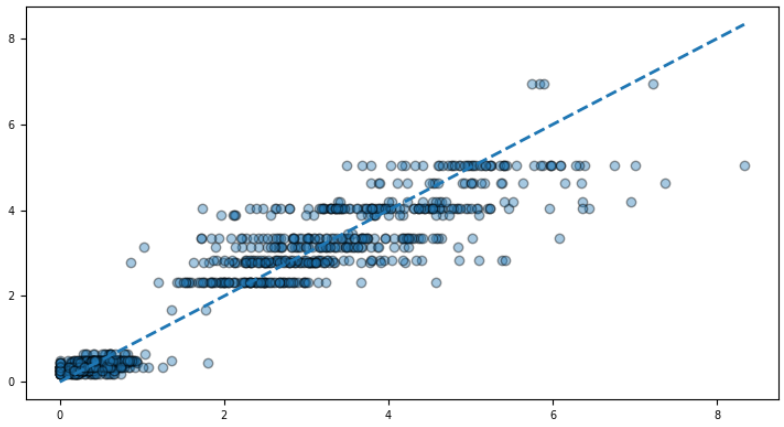

Figure 4.14: Random forest test predictions (x) versus real values (y).

Table 4.2: Results for the classical machine learning models.

Predictors	Model	RMSE train	R^2 train	RMSE test	R^2 test
1	LR	1.37	0.39	1.37	0.36
	RF	1.55	0.21	1.63	0.15
	XGB	1.14	0.58	1.45	0.28
	DT	1.46	0.31	1.58	0.14
2	LR	0.58	0.89	0.57	0.89
	RF	0.73	0.82	0.82	0.78
	XGB	0.49	0.92	0.61	0.87
	DT	0.71	0.84	0.73	0.82
3	LR	0.50	0.92	0.51	0.91
	RF	0.67	0.85	0.71	0.84
	XGB	0.42	0.94	0.54	0.89
	DT	0.54	0.91	0.57	0.89
4	LR	0.50	0.92	0.49	0.91
	RF	0.67	0.85	0.71	0.84
	XGB	0.42	0.94	0.53	0.90
	DT	0.54	0.91	0.57	0.89
5	LR	0.55	0.90	0.56	0.89
	RF	0.39	0.94	0.51	0.89
	XGB	0.46	0.93	0.56	0.89
	DT	0.57	0.89	0.60	0.87
6	LR	0.56	0.90	0.57	0.91
	RF	0.45	0.92	0.54	0.89
	XGB	0.49	0.92	0.55	0.89
	DT	0.51	0.91	0.58	0.89
7	LR	0.55	0.90	0.54	0.91
	RF	0.42	0.93	0.58	0.90
	XGB	0.47	0.93	0.56	0.90
	DT	0.57	0.89	0.58	0.90
8	LR	0.55	0.90	0.53	0.91
	RF	0.41	0.95	0.53	0.92
	XGB	0.46	0.93	0.55	0.90
	DT	0.57	0.89	0.57	0.90
9	LR	0.52	0.91	0.57	0.91
	RF	**0.41**	**0.95**	**0.51**	**0.91**
	XGB	0.43	0.94	0.55	0.90
	DT	0.55	0.90	0.60	0.90
10	LR	0.55	0.90	0.53	0.91
	RF	**0.40**	**0.95**	**0.51**	**0.91**
	XGB	0.39	0.95	0.53	0.91
	DT	0.56	0.89	0.56	0.90

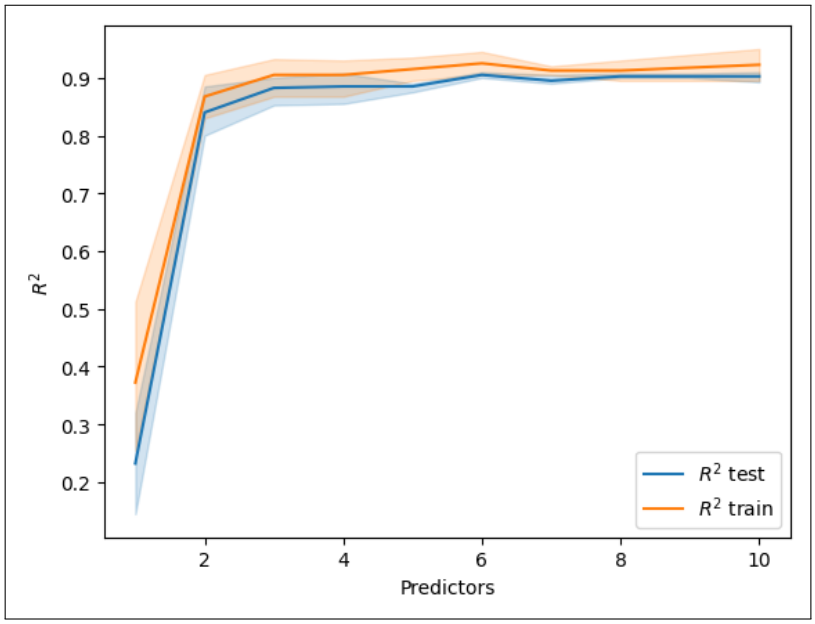

Figure 4.15: R^2 in train and test for the different variable configurations.

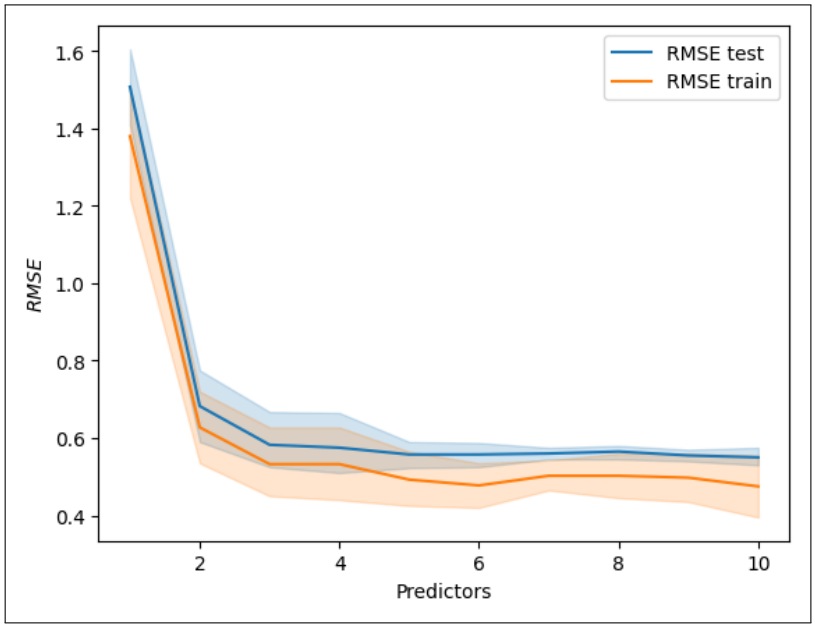

Figure 4.16: RMSE in train and test for the different variable configurations.

Based on the principle of parsimony, we will choose the simplest model among the two considered (in bold), which is model 9. The residual diagnosis plot is presented in Figure 4.17.

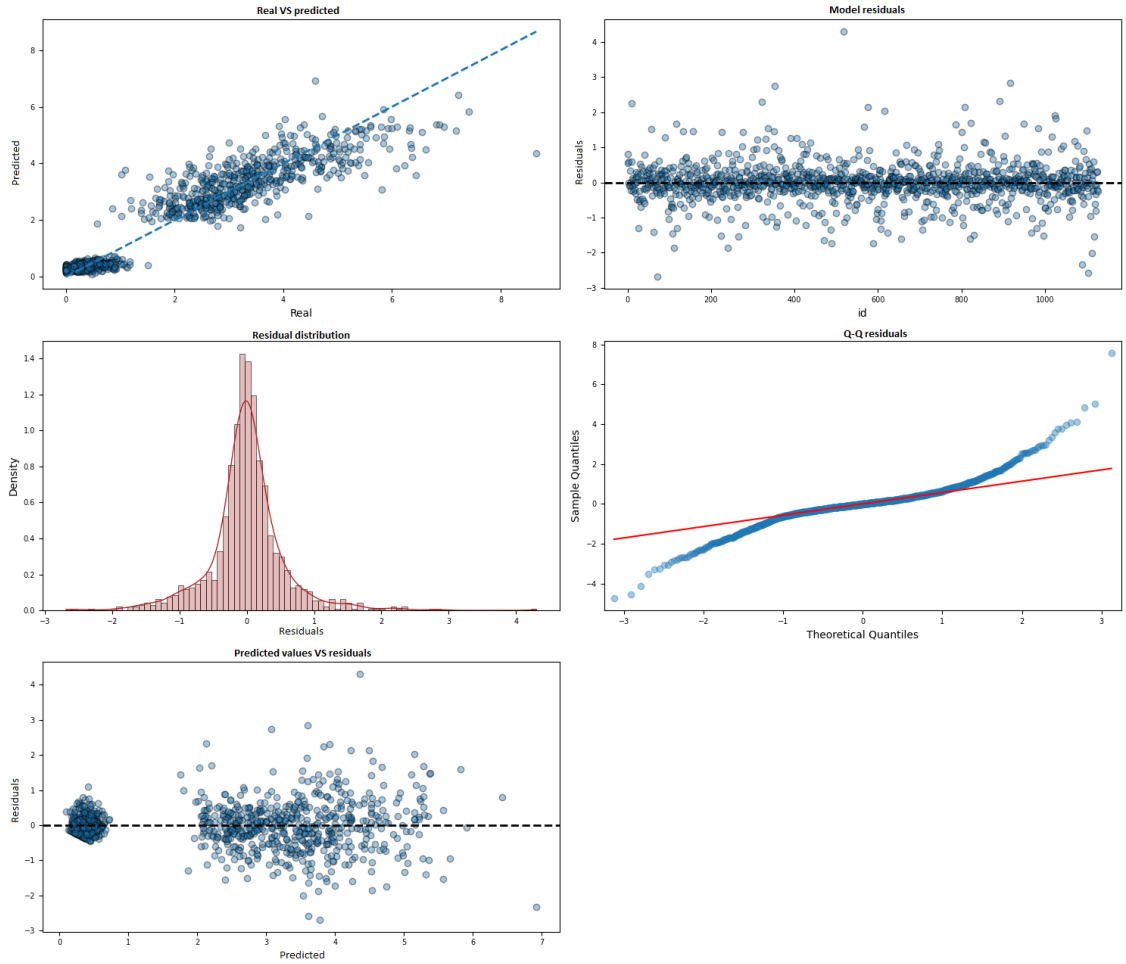

Figure 4.17: Residual diagnosis for random forest selected model.

In the context of lung cancer mortality rate per 1000 inhabitants, an RMSE of 0.51 indicates that, on average, the model's predicted values deviate from the actual values by 0.51 deaths per 1000 inhabitants. This could be considered a relatively high error, but it seems that the model is having trouble with group number 4 more than with the others, which increases the overall RMSE. The behavior of the model is similar to the one reflected in Figure 4.6, which looks adequate, indicating that the model is performing reasonably well in its predictions.

An R^2 value of 0.91 suggests that the model explains 91% of the variance in the data. This means that the model is able to capture most of the variation in lung cancer mortality rates and provides a good fit to the data. It also suggests that the model is likely to be reliable in making predictions about lung cancer mortality rates for new data, as it has explained a large proportion of the variance in the existing data. The behavior of the model's predictions on the training and test sets can be observed in Figure 4.18.

So far, we have obtained classical machine learning results, specifically point estimates. Now, let's consider the scenario where we want to incorporate confidence intervals, similar to what we do in Bayesian statistical models. To achieve this, we utilized the *python* library `forestci` to compute the unbiased sample variance using the infinitesimal jackknife method (explained in [17]). In Figure 4.19,

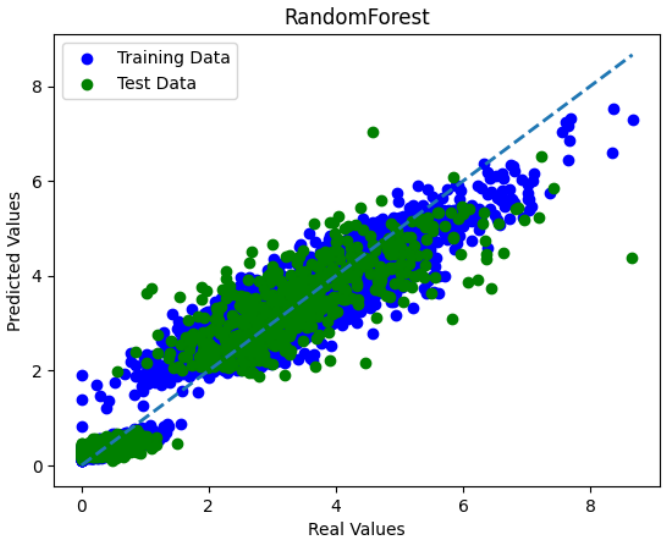

Figure 4.18: Real versus predicted Rates, colored by pertenence to the training or to the test dataset.

we depict this sample variance. Subsequently, we apply that

$$CI_{1-\alpha}(pred) = \overline{pred} \pm t_{n-1,\alpha/2} * \sqrt{\frac{S^2}{n}}$$

to derive the confidence intervals from the point estimates $\overline{pred}$ and the obtained sample variance, S^2.

Now, we are able to plot the point estimates together with the prediction intervals for each row of the dataset, 4.20.

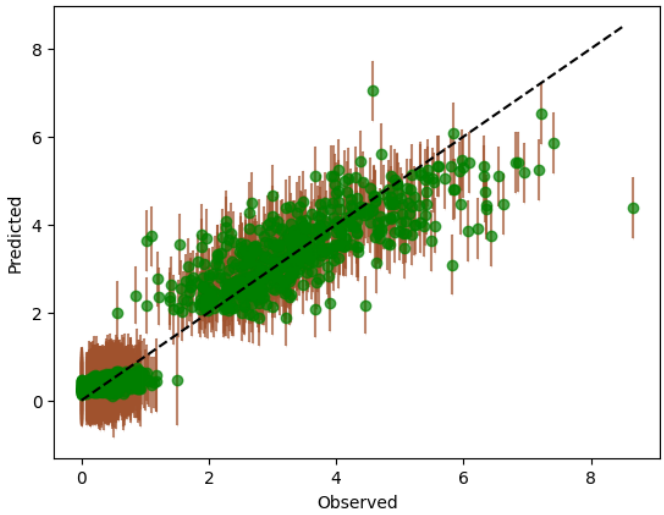

Figure 4.19: Point estimates for the test dataset together with each point's standard deviation acording to *forestci*.

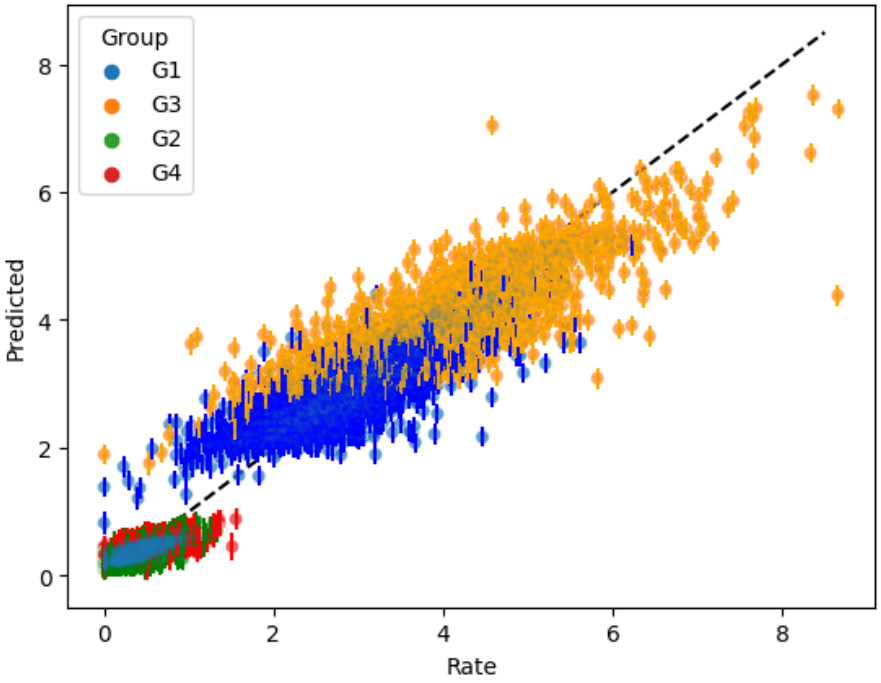

Figure 4.20: Real versus predicted values for each group with their 95% prediction intervals.

4.3 Deep learning models

The objective is to train a deep learning model that can effectively capture and model spatio-temporal data patterns. To achieve this, a preprocessing step has been conducted to transform the input data: the construction of temporal windows. Specifically, temporal windows of 3 years have been defined for each province, enabling the consideration of spatial and temporal dependencies during both the training and validation processes. By incorporating these temporal windows, the model can account for the complex interplay between spatial and temporal factors. Furthermore, it is because of these windows that the model accounts for the temporal autocorrelations.

In this work, three types of neural networks have been implemented: Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), and Bayesian Neural Networks (BNN). Each of these neural network architectures possesses distinct features and is worth comparing them.

The MLP is a fundamental type of neural network commonly used for supervised learning tasks. It consists of multiple layers of interconnected nodes, where each node applies a (non)linear activation function to the weighted sum of its inputs. MLPs are known for their ability to approximate complex nonlinear mappings and are widely used for tasks such as classification and regression.

The LSTM network is a specialized type of recurrent neural network (RNN) that addresses the vanishing gradient problem associated with traditional RNNs. LSTMs are designed to effectively capture and model dependencies over time, making them suitable for tasks involving sequential or time-series data. They utilize memory cells and gating mechanisms to selectively remember or forget information, enabling them to capture long-term dependencies and handle sequential patterns more effectively.

BNNs (introduced by [18]) are neural networks that incorporate Bayesian inference techniques. Unlike traditional neural networks, BNNs provide not only point estimates but also uncertainty estimates for their predictions. This is achieved by representing the weights of the network as probability distributions and sampling for them, allowing for the propagation of uncertainty throughout the

Table 4.3: Results for deep learning models.

Network type	RMSE train	R^2 train	RMSE test	R^2 test
MLP	0.46	0.93	0.51	0.91
LSTM	0.49	0.92	0.51	0.91
BNN	0.55	0.84	0.63	0.84

network. BNNs offer benefits such as robustness to overfitting, better calibration of uncertainty, and the ability to incorporate prior knowledge.

The best results obtained for each network are shown in Table 4.3. For each network type, we performed several trials with different configurations of hyperparameters and architectures. Apart from that, we used a learning rate scheduler and *tensorboard* to visualize the loss evolution and weights distribution among other magnitudes. Then we selected the best configuration of each kind. The loss function we used was the RMSE for the MLP and LSTM and the negative log likelihood for the BNN. Since the output of the BNN model is a distribution rather than a point estimate, the loss function should be used to compute how likely it is to see the true data from the estimated distribution produced by the model.

The selected model was the BNN because it provides credible intervals and good point estimates. The MLP and LSTM obtain similar results to the ones by Random Forest, but are not able to give these intervals. The BNN model architecture is shown in Figure 4.21. The results are quite nice, as we can see in Figure 4.22.

An example of what we visualize when using *tensorboard* on the fully connected 30 nodes layer is presented in Figure 4.23. At first, weights are randomly initialized and is easy to see how they are learnt (calculated to reduce the loss function through *backpropagation*), getting fixed around the 100-th epoch.

Finally, we conducted the same analysis we did before with INLA, for Madrid, Navarra and Cáceres using our BNN. Results are displayed in Figures 4.24, 4.25, 4.26. As we can see, the posterior median is much closer to the crude rate than INLA models, but the intervals are much wider. For women, predictions didn't work very well. Some observations fall under the credible intervals because they are too wide and the median estimates are not smooth. Differences are due to the fact that BNN credible intervals are created to provide a range of values on which each observation will fall with a 95% probability.

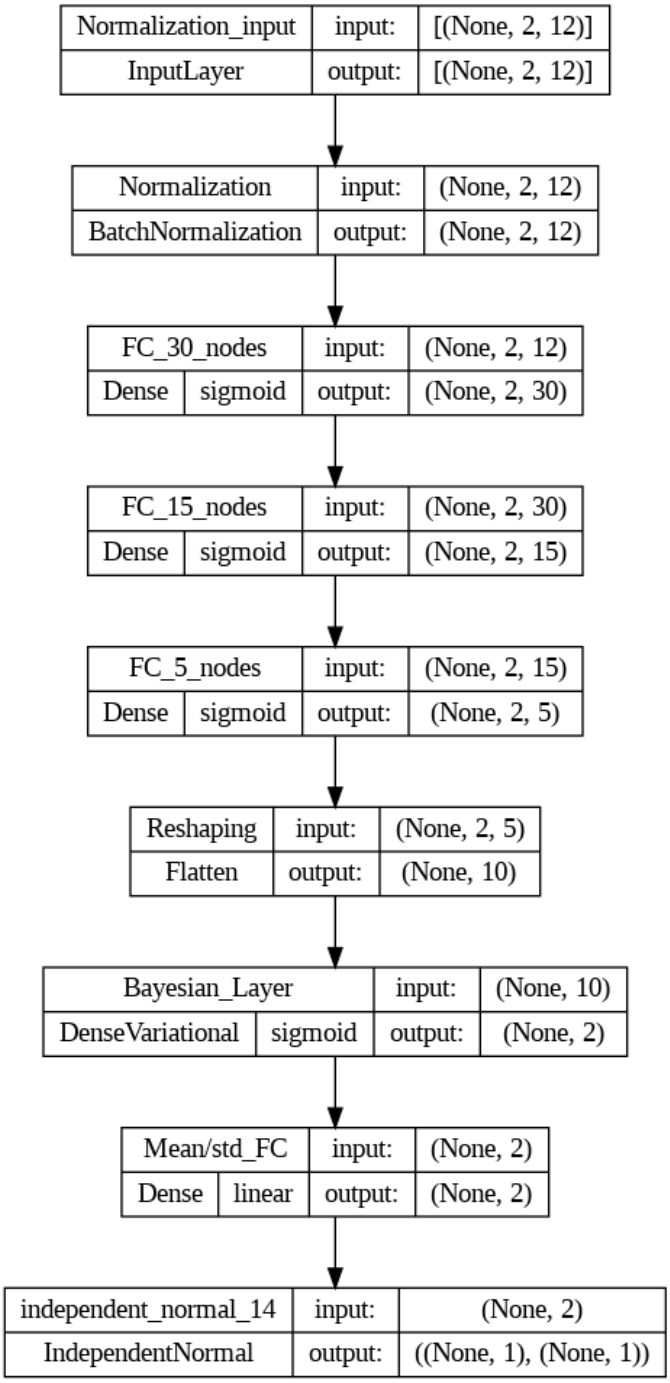

Figure 4.21: Selected Bayesian Neural Network architecture.

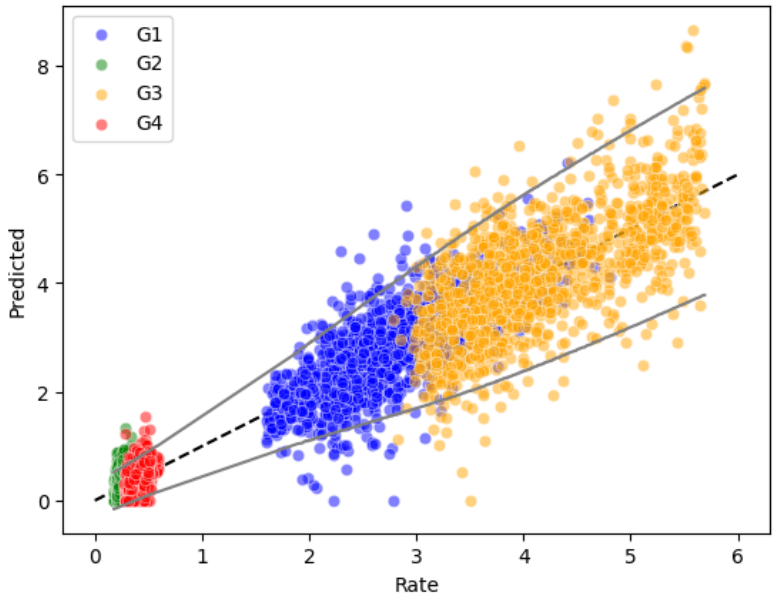

Figure 4.22: Posterior median estimates for the rate with the corresponding 95% credible intervals.

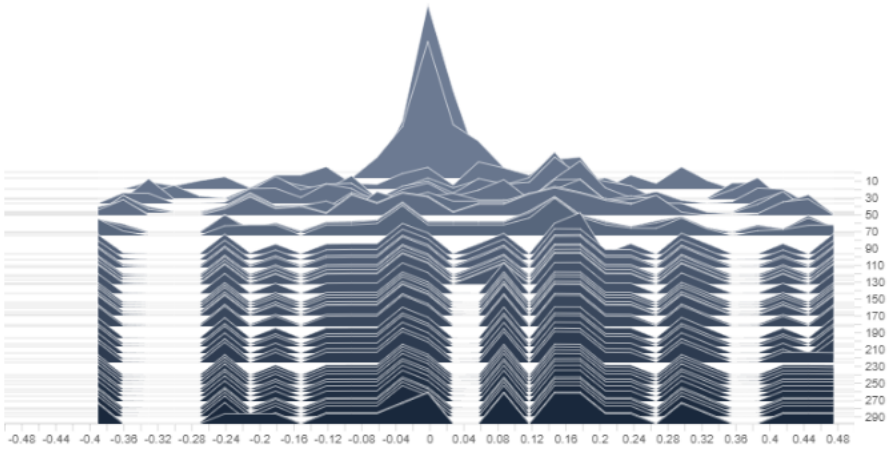

Figure 4.23: Weight distributions for the first dense layer of the BNN.

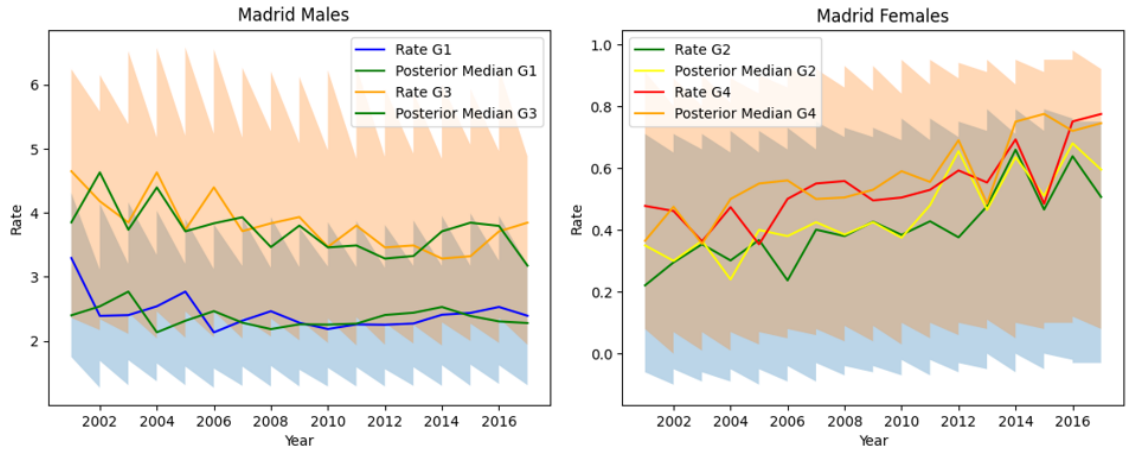

Figure 4.24: Evolution of rate and posterior median in Madrid, from first to last year, with the corresponding credible intervals.

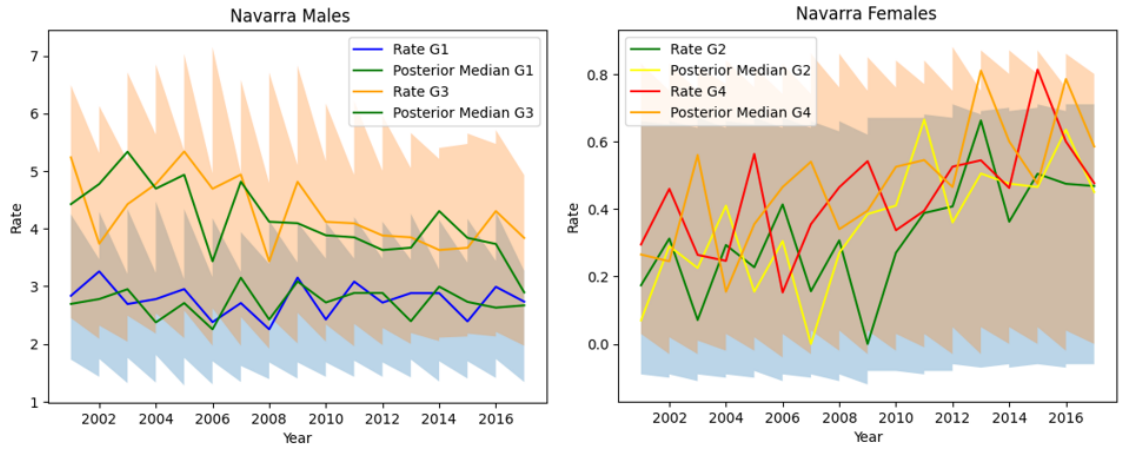

Figure 4.25: Evolution of rate and posterior median in Navarra, from first to last year, with the corresponding credible intervals.

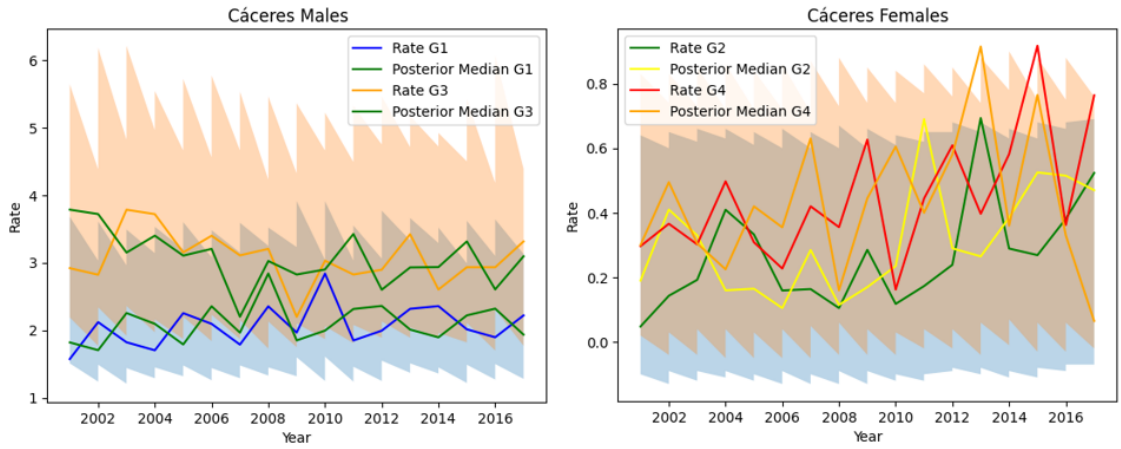

Figure 4.26: Evolution of rate and posterior median in Cáceres, from first to last year, with the corresponding credible intervals.

4.4 Comparison and results

In this section we compare the results obtained with each method and discuss their main strengths and weaknesses.

4.4.1 Simulation

To rigorously evaluate the performance of the different models, we conduct a simulation study inspired by [19]. This simulation involves considering the observations we have for each province, year, and stratum as realizations of a stochastic process. While we have a specific sample, it is important to acknowledge that alternative samples could have been generated from an unknown distribution for each location, time, and stratum.

Therefore, we are interested in evaluating how well our models would perform on different realizations of the stochastic process. However, the challenge lies in the fact that we do not have access to the means of these Poisson distributions. To address this, we propose treating each observation as the mean of a Poisson distribution, allowing us to generate additional samples. It is worth noting that this approach may introduce some bias, as the observed values may not necessarily align with the true means of the Poisson distributions, although we assume they are close.

To mitigate this bias, we plan to estimate the means of the Poisson distributions by taking a weighted average of the observations in each province, year, and stratum, considering the spatial proximity of neighboring provinces. This will help to smooth out the values by incorporating information from neighboring provinces.

Therefore, we calculate

$$\lambda_{ijt} = 0.8 * Rate_{ijt} + 0.2 * spatial\_lag_{ijt}$$

and assign, for the k-ith simulation,

$$SimulatedDeaths_{ijt}^{(k)} = Poisson(\lambda_{ijt} n_{ijt})$$

Once we have the simulated observations, which can be interpreted as other possible outcomes of the stochastic process, we calculate the simulated rates dividing by the population. Consequently, we have successfully generated simulations for the rates. We have conducted a total of 10 simulations. The results of the simulations are shown in Figure 4.27.

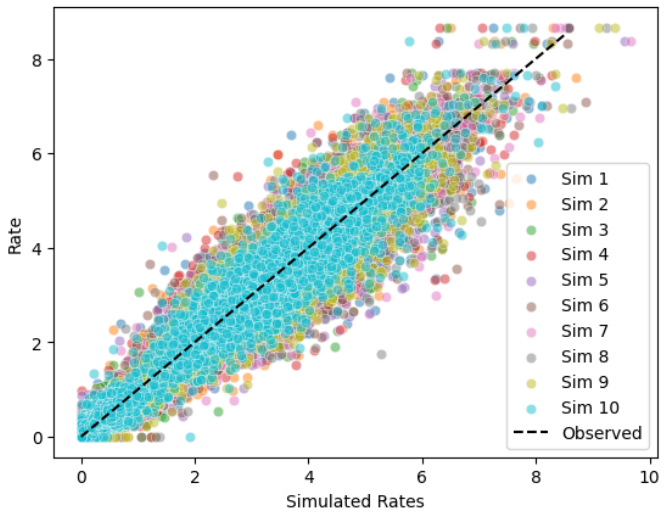

Figure 4.27: Simulated rates versus original rates for 10 replications.

4.4.2 Results

The simulation allows us to assess whether our models have been overfitted to the specific instance of the spatio-temporal stochastic process in the original data, or if they can effectively generalize to alternative outcomes of the same process.

To evaluate the three approaches, we will utilize the following metrics:

- Mean absolute difference: $\sum_{k=1}^{K=10} \frac{|SimulatedRate_{ijt}^{(k)} - prediction_{ijt}|}{10}$, $\forall i, j, t$. It represents the mean difference in absolute magnitude between the simulated rates and the predicted rates for the 10 simulations.
- Mean R^2: $\sum_{k=1}^{K=10} \frac{R^{2(k)}}{10}$. Represents the mean R^2 for one realization of the stochastic process.
- Mean RMSE: $\sum_{k=1}^{K=10} \frac{RMSE^{(k)}}{10}$. It is the expected RMSE for one realization of the stochastic process.
- Number of simulated rates that fall inside the 95% credible/prediction intervals. It helps us get an idea of how well the confidence/credible/prediction intervals capture the rates and simulated rates.
- Mean interval amplitude.
- Interpretability of the model.
- Fitting time.

We present the results for the numerical metrics in Figure 4.28. INLA was able to get the best results in R^2, RMSE, and Absolute difference, but the three model types were excellent. It is worth noting that the BNN exchanges R^2 and RMSE to provide credible intervals, compared to the other neural networks. The random forest approach has a very small interval amplitude, due to the fact that those are prediction intervals and there is small variability in the predictions of the model. On the other hand, BNN intervals are wide to capture 95% of the rates. In contrast, INLA has small credible intervals with accurate median estimates. INLA captured 58% of the simulated rates in its credible intervals, random forest captured 44% and BNN captured 96%.

Neural networks have demonstrated the ability to achieve comparable accuracy to classical machine learning models, eliminating the need for explicit feature engineering. They autonomously uncover the intrinsic patterns within the data, even with a small dataset. INLA operates similarly by utilizing informative prior distributions that contribute valuable insights to the model. Random Forest models offer feature importance measures, which indicate the relative contribution of predictors to the model's predictions. However, the interpretability of individual trees within the Random Forest ensemble is limited. Apart from that, machine learning approaches do not require the design of neural network architectures; instead, established models are employed and applied directly to the data at hand.

In terms of computational time required to fit the data, the classical machine learning models took the longest, with approximately 15 hours for the entire process (ranging from 45 minutes to 1 hour and 30 minutes for each model). The Bayesian Neural Network fit with 300 epochs took around 5 minutes, but exploring different networks and configurations extended the total fitting time to about 2 hours and 30 minutes. Fitting a BNN is not an easy task, as only small to medium size datasets are appropriate. For large amounts of data it would be non-viable. In contrast, fitting the INLA models involved 4 models per group (16 models in total), and each model took approximately 10 seconds to fit when using the `compact` mode, which makes INLA the fastest in this case. INLA leverages optimization techniques and approximate methods to provide fast results, even for large datasets. BNNs exhibit black box characteristics in prediction, yet they provide valuable insights into the posterior distribution of their parameters.

INLA allows for easy interpretation of the effects of predictor variables and can incorporate prior information effectively. BNNs, however, can be more challenging to interpret due to the complexity

of neural network architectures. They often focus on prediction rather than explicit parameter estimation, making it harder to derive interpretable insights from the network weights.

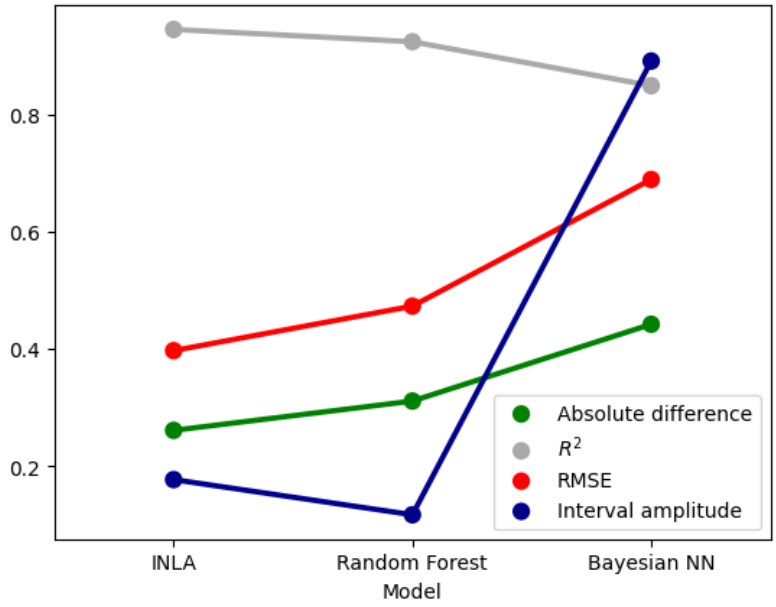

Figure 4.28: Absolute difference, mean R^2, mean RMSE and mean interval amplitude for each approach.

INLA is capable of making predictions based on the estimated spatio-temporal patterns and incorporates uncertainty quantification through posterior distributions. ML models, on the other hand, are designed to optimize predictive accuracy. They leverage large amounts of data to learn patterns and relationships, allowing for more accurate predictions in many cases. ML models can handle complex and non-linear relationships, and with appropriate training, they can generalize well to new spatio-temporal scenarios. However, ML models often lack explicit uncertainty quantification, which can be a disadvantage when making decisions in uncertain or risk sensitive contexts. Overall, while INLA provides detailed description and uncertainty quantification, ML models prioritize predictive accuracy by learning patterns directly from the data. The choice between the two approaches depends on the specific goals of the analysis, the available data, the need for interpretability, and the importance of uncertainty estimation in the context of the spatio-temporal dataset.

Classical ML models have well-established theoretical foundations and are often easier to understand conceptually. They rely on mathematical algorithms and statistical principles that are relatively straightforward to grasp, making them accessible to users with a solid understanding of statistical concepts. INLA is a Bayesian approach that requires a good understanding of Bayesian statistics and modeling assumptions. The theoretical foundations of INLA involve complex statistical concepts like Bayesian inference and approximations using the Laplace method. Implementing INLA effectively requires a strong grasp of these advanced statistical principles, making it more challenging from a theoretical perspective. Deep learning models, especially deep neural networks, involve highly complex architectures and mathematical concepts like backpropagation, gradient optimization, and activation functions. Understanding the theoretical foundations of deep learning can be challenging. Deep learning models often require a solid understanding of linear algebra, calculus, and optimization algorithms.

Chapter 5

Conclusions and further work

Throughout my journey as a Data Science student, I have gained valuable knowledge and practical experience in various aspects of the field. My end-of-studies work has provided me with a comprehensive understanding of Bayesian inference, its fundamental principles, and its application to real-life problems. I have also explored innovative techniques to incorporate spatio-temporal dependencies into classical machine learning models.

One of the areas I focused on was enhancing the performance of machine learning models by incorporating spatial and temporal information. By leveraging coordinates, space, time, and space-time lags, as well as utilizing basis functions, I was able to capture the underlying dependencies present in the data. This approach proved effective in improving the performance of classical machine learning models when dealing with spatio-temporal data.

In addition, I extended my exploration to Random Forest models and successfully obtained prediction intervals. These prediction intervals provided valuable insights into the uncertainty associated with the model's predictions, enabling a more comprehensive assessment of the model's reliability and the potential variability in the outcomes.

Moreover, I delved into the realm of neural networks and developed the architectures of three distinct types: Multi-Layer Perceptron, Long Short-Term Memory, and Bayesian Neural Network. The Bayesian Neural Network, in particular, allowed me to provide credible intervals for the model's predictions. By incorporating Bayesian principles, I was able to quantify the uncertainty in the model's output and provide more informative predictions with associated confidence.

Due to the academic nature of this work and its main objectives, not every proposed model has been fully exploited. However, a hypothetical continuation of the study could enhance the results and interpretability of the models. One potential avenue for improvement is the incorporation of explainability techniques, such as SHAP (SHapley Additive exPlanations) values, to classical machine learning models. SHAP values provide insights into the contribution of each feature to the model's predictions, thereby improving interpretability.

To further capture spatio-temporal correlations in data within machine learning models, additional exploration is needed. More sophisticated approaches, such as graph neural networks, could be investigated. Graph neural networks are specifically designed to handle spatial and temporal dependencies and are considered a promising solution for analyzing space-time dependent datasets. Their utilization could enhance the accuracy and predictive capabilities of the models.

The outcomes derived from INLA can be leveraged extensively to extract maximum benefits. The results obtained from INLA can be incorporated into the training process of machine learning models or utilized as priors in Bayesian frameworks. This incorporation enhances the interpretability and robustness of the models, enabling better decision-making and improved performance.

In conclusion, while this study has provided valuable insights into Bayesian inference, spatio-temporal dependencies, and neural network architectures, there are several areas for further exploration and improvement, which I am sure I will pursue in future research.

Bibliography

[1] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, Nov. 2013. doi: 10.1201/b16018.

[2] R. van de Schoot, S. Depaoli, R. King, *et al.*, "Bayesian statistics and modelling," *Nature Reviews Methods Primers*, vol. 1, no. 1, Jan. 2021. doi: 10.1038/s43586-020-00001-2.

[3] M. D. Ugarte, A. Adin, T. Goicoa, and A. F. Militino, "On fitting spatio-temporal disease mapping models using approximate bayesian inference," *Statistical Methods in Medical Research*, vol. 23, no. 6, pp. 507–530, Apr. 2014. doi: 10.1177/0962280214527528.

[4] H. Rue and L. Held, *Gaussian Markov Random Fields*. Chapman and Hall/CRC, Feb. 2005. doi: 10.1201/9780203492024.

[5] T. Goicoa, A. Adin, M. D. Ugarte, and J. S. Hodges, "In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results," *Stochastic Environmental Research and Risk Assessment*, vol. 32, no. 3, pp. 749–770, Mar. 2017. doi: 10.1007/s00477-017-1405-0.

[6] H. Rue, S. Martino, and N. Chopin, "Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 71, no. 2, pp. 319–392, Apr. 2009. doi: 10.1111/j.1467-9868.2008.00700.x.

[7] J. V. Niekerk, E. Krainski, D. Rustand, and H. Rue, "A new avenue for bayesian inference with INLA," *Computational Statistics &amp Data Analysis*, vol. 181, p. 107 692, May 2023. doi: 10.1016/j.csda.2023.107692.

[8] M. Blangiardo and M. Cameletti, *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Wiley, Apr. 2015. doi: 10.1002/9781118950203.

[9] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. V. D. Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 64, no. 4, pp. 583–639, Oct. 2002. doi: 10.1111/1467-9868.00353. [Online]. Available: https://doi.org/10.1111/1467-9868.00353.

[10] S. Watanabe, "Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory," *J. Mach. Learn. Res.*, vol. 11, pp. 3571–3594, Dec. 2010, issn: 1532-4435.

[11] R. Liu Zhedong, "Leave-group-out cross-validation for latent gaussian models," 2022. doi: 10.48550/ARXIV.2210.04482.

[12] C. K. Wikle and A. Zammit-Mangion, "Statistical deep learning for spatial and spatiotemporal data," *Annual Review of Statistics and Its Application*, vol. 10, no. 1, pp. 247–270, Mar. 2023. doi: 10.1146/annurev-statistics-033021-112628.

[13] O. S. Zhao, N. Kolluri, A. Anand, *et al.*, "Convolutional neural networks to automate the screening of malaria in low-resource countries," *PeerJ*, vol. 8, e9674, Aug. 2020. doi: 10.7717/peerj.9674. [Online]. Available: https://doi.org/10.7717/peerj.9674.

[14] A. Galindo-Utrero, J. M. San-Román-Montero, R. Gil-Prieto, and Á. Gil-de-Miguel, "Trends in hospitalization and in-hospital mortality rates among patients with lung cancer in spain between 2010 and 2020," *BMC Cancer*, vol. 22, no. 1, Nov. 2022. doi: 10.1186/s12885-022-10205-2.

[15] E. A. Fattah and H. Rue, "Approximate bayesian inference for the interaction types 1, 2, 3 and 4 with application in disease mapping," 2022. doi: 10.48550/ARXIV.2206.09287.

[16] X. Liu, O. Kounadi, and R. Zurita-Milla, "Incorporating spatial autocorrelation in machine learning models using spatial lag and eigenvector spatial filtering features," *ISPRS International Journal of Geo-Information*, vol. 11, no. 4, p. 242, Apr. 2022. doi: 10.3390/ijgi11040242.

[17] S. Wager, T. Hastie, and B. Efron, *Confidence intervals for random forests: The jackknife and the infinitesimal jackknife*, 2013. doi: 10.48550/ARXIV.1311.4555.

[18] I. Kononenko, "Bayesian neural networks," *Biological Cybernetics*, vol. 61, no. 5, pp. 361–370, Sep. 1989. doi: 10.1007/bf00200801.

[19] E. Orozco-Acosta, A. Adin, and M. D. Ugarte, "Big problems in spatio-temporal disease mapping: Methods and software," *Computer Methods and Programs in Biomedicine*, vol. 231, p. 107 403, Apr. 2023. doi: 10.1016/j.cmpb.2023.107403.