Modelos bayesianos para representar el efecto de variables atmosféricas en series diarias de temperatura

Jorge Navarro Lapena

Trabajo de fin de grado de Matemáticas
Universidad de Zaragoza

Directores del trabajo: Jorge Castillo Mateo
y Jesús Asín Lafuente

4 de septiembre de 2023

Abstract

The growing evidence of climate change, driven by the increase in the concentration of greenhouse gases in the atmosphere, suggests a generalized increase in temperatures. However, this increase can vary significantly throughout the year and between different geographic regions. Precise identification of the existence and magnitude of these temporal and spatial trends is crucial for formulating effective management strategies that mitigate the adverse impacts of rising temperatures on human health, agriculture, and the economy. Therefore, the primary objective of this study is to develop data-driven models to analyze the effects of climate change on temperatures and their geographic variability in Spain, using daily temperature time series.

The methodology employed is based on papers published in *Journal of Agricultural, Biological and Environmental Statistics* and in *Annals of Applied Statistics* by Castillo-Mateo et al. [1], [2]. Those authors used Bayesian hierarchical models to make inferences about the distribution of daily maximum temperature in Aragón. Those works show that the daily temperature distribution has a seasonal pattern that affects to the mean value and the variance. According to Castillo-Mateo et al., we have decided to set our framework in the Bayesian frame due to the fact that it offers a flexible fitting and full inference. . Our joint model for the mean and the variance will consist of the following expressions:

$$\begin{aligned} E[Y_t(s)|\mathbf{X}_t(s)] &= \beta_0(s) + \sum_{i=1}^{2}(\beta_{i,s}(s)Si_t + \beta_{i,c}(s)Ci_t) + \Big(\alpha_0(s) + \sum_{i=1}^{2}(\alpha_{i,s}(s)Si_t + \alpha_{i,c}(s)Ci_t)\Big)Year_t \\ &+ \Big(\rho_0(s) + \sum_{i=1}^{2}(\rho_{i,s}(s)Si_t + \rho_{i,c}(s)Ci_t)\Big)Y_{t-1}(s) + \Big(\gamma_0(s) + \sum_{i=1}^{2}(\gamma_{i,s}(s)Si_t + \gamma_{i,c}(s)Ci_t)\Big)Y_{t-2}(s). \end{aligned} \tag{1}$$

$$\begin{aligned} Var[Y_t(s)|\mathbf{X}_t(s)] &= \exp\Big\{\beta_0^{\sigma}(s) + \sum_{i=1}^{2}(\beta_{i,s}^{\sigma}(s)Si_t + \beta_{i,c}^{\sigma}(s)Ci_t) + \Big(\alpha_0^{\sigma}(s) + \sum_{i=1}^{2}(\alpha_{i,s}^{\sigma}(s)Si_t + \alpha_{i,c}^{\sigma}(s)Ci_t)\Big)Year_t \\ &+ \Big(\rho_0^{\sigma}(s) + \sum_{i=1}^{2}(\rho_{i,s}^{\sigma}(s)Si_t + \rho_{i,c}^{\sigma}(s)Ci_t)\Big)Y_{t-1}(s) + \Big(\gamma_0^{\sigma}(s) + \sum_{i=1}^{2}(\gamma_{i,s}^{\sigma}(s)Si_t + \gamma_{i,c}^{\sigma}(s)Ci_t)\Big)Y_{t-2}(s)\Big\}. \end{aligned} \tag{2}$$

Where $Y_t(s)$ denotes the temperature on day t at location s in our study region, and the covariate vector $\mathbf{X}_t$(s) includes the information associated with the date: the harmonic components Si_t, Ci_t, with $Si_t = \sin\frac{2\pi ti}{365}$ and $Ci_t = \cos\frac{2\pi ti}{365}$; $Year_t$ the year; and the previous values $Y_{t-1}(s)$ and $Y_{t-2}(s)$. We use the superscript σ for the parameters of the variance model. In turn, we denote as $\alpha_i(s)$ the parameters associated with the temporal trend represented by the year covariate. $\rho_i(s)$ and $\gamma_i(s)$ are associated with the effects of the previous day and two days before, respectively. In the database, we have the observed sites $s_k \in \{s_1, \ldots, s_K\}$.

We focus on 40 Spanish stations, in the period from January 1, 1955, to December 31, 2022, with a wide climatic and geographical variability. The data series are provided by the European database ECA&D [8]. We use the R package *bamlss*, as referenced in articles [9] and [10], which provides tools for Bayesian inference for flexible additive models. This package is particularly suited for our analysis as we aim to develop a joint model for both the mean and the variance of daily maximum temperatures

To summarize, Chapter 1 is about the introduction, objectives and phases and procedures of this work.

Chapter 2 contains a revision of basic ideas of linear models, Bayesian inference, and Bayesian hierarchical models. The main ideas and results of papers by Castillo-Mateo et al. [1] are summarized in order to consider the characteristics that must be represented by statistical models, such as autoregressive effects, seasonal behavior, and temporal trend. Results of Castillo-Mateo et al. [1] are revised because they show that the modeling of variance is required.

Chapter 3 will establish the procedure to conduct the exploratory analysis and latter the Bayesian inference. In particular, the structure of the model is proposed. A brief description of the developed R function to make inference is included.

Chapter 4 summarizes the main results obtained by applying the methodology to daily temperature series in 40 Spanish weather stations. We will start with an exploratory data analysis, through various procedures such as the use of 30-day moving quantiles. We will continue with the construction of linear models for the mean and variance, and their subsequent analysis. In turn, we will study the spatial distribution of linear predictors. Next, we develop a local and a global Bayesian model and compare them, in addition to analyzing the posterior distribution of parameters associated to the temporal trend, i.e., to the climate change signals. It is found that the elements of the model are required, both memory of order 2 and trend in the mean value and variance. It is also found that there is a geographical variability that is not perfectly reproduced using only geographical covariates.

Finally in Chapter 5, we highlight the main results and implications of this work. And, we conclude with the proposal of future works and research lines.

Índice general

Capítulo 1

Introducción

El objetivo del trabajo fin de grado (TFG) es el estudio de los modelos bayesianos de tipo regresión aplicados sobre series temporales en un punto, como herramienta estadística para representar sus características. Las series temporales son secuencias de datos de una variable registrada en instantes de tiempo sucesivos y, en general, equiespaciados. Estas técnicas serán utilizadas para modelar la serie diaria de temperatura máxima (Y) en un punto s de la Península Ibérica, en el día t, $Y_t(s)$. Para ello, hemos utilizado los datos de 40 lugares, en el periodo comprendido entre 1955 y 2022. En particular, los modelos serán capaces de representar la variabilidad espacial del clima y la evolución temporal de la temperatura, que incluye estacionalidad, dependencia de la situación en días previos y un incremento de nivel en el largo plazo (cambio climático observado). Esta variabilidad temporal debe reflejarse en la estimación de la media y de la varianza. Para ello, usaremos la perspectiva dada por los artículos de Castillo-Mateo y colaboradores [1] y [2].

El modelo estadístico tendrá dos requerimientos adicionales. Debe representar la estructura de la autocorrelación, es decir, la dependencia respecto a la situación descrita por la propia respuesta en días previos, que constituye parte del vector de covariables, $X_t(s)$, y presentar varianza no constante. Estos requerimientos implican que se necesitan modelos de tipo regresión para $Y_t(s), t = 1, \ldots, T$, donde la distribución es gaussiana con esperanza $\mu_t(s) = E[Y_t(s)|X_t(s)]$ y varianza $\sigma_t^2(s) = Var[Y_t(s)|X_t(s)]$, es decir, $Y_t(s) \sim \mathcal{N}(\mu_t(s), \sigma_t(s))$. Este modelo excede a los estudiados en las asignatura de grado, puesto que el modelo lineal usual impone que $\sigma_t(s) = \sigma(s)$ es constante. Veremos también la necesidad de incluir términos armónicos para reflejar la estacionalidad. Además, se pretende una modelización conjunta incluyendo covariables espaciales, es decir de ubicación geográfica o de caracterización climática. Por ello, el marco de estimación que se ha seleccionado es el de los modelos jerárquicos bayesianos.

El trabajo práctico se realizará sobre la base de datos Tmax-ECA, `https://www.ecad.eu` [8], en la que se hace un trabajo preliminar de depuración, como la eliminación de observatorios con demasiados datos faltantes, o la inclusión de más variables del ámbito temporal. Se ha diseñado un procedimiento que constará de dos fases:

- **Primera fase:** Análisis exploratorio de cada lugar, utilizando modelos lineales (basados en máxima verosimilitud), que empleen la propia variable respuesta para explicar la estructura del predictor lineal en la media (variables autorregresivas), y usando como respuesta el cuadrado de los residuos del modelo anterior para explorar la estructura de la varianza de la respuesta diaria. A partir de las estimaciones en cada punto, se pretende estudiar su dependencia/relación con variables geográficas o climáticas que caracterizan los tipos de clima.

- **Segunda fase:**

  1. Ajuste de un modelo local bayesiano, que utilizaremos para comparar con su respectivo modelo local de regresión lineal, lo que nos dará información útil sobre algunas covariables usadas en el modelo.
  2. Ajuste de un modelo global bayesiano, que incluya tanto variables geográficas como temporales, con el fin de poder modelar la temperatura máxima diaria en cualquier punto de la

Península.

Se trabajará con librerías de R adecuadas para este tipo de modelización, *bamlss*, librería descrita en los trabajos de Umlauf et al [9] y [10], que permite la construcción de modelos bayesianos para la media y la desviación típica, conjuntamente.

Se han obtenido otros resultados adicionales. En particular, se estudiará la variabilidad espacial de los parámetros de los predictores de los modelos de media y varianza. Para ello se plantearán modelos que usarán como covariables las variables geográficas. Su capacidad para explicar la variabilidad espacial de los modelos locales se comparará con la mejora en su representación que implica utilizar variables de tipo climático, como las características estacionales en un periodo de referencia. Además, se ha generado una librería de funciones R que se ha puesto a disposición de la comunidad a través de Github.

El contenido de la memoria incluye un capítulo donde se revisarán los conceptos básicos sobre modelos de regresión y en el ámbito de la estadística bayesiana. El siguiente capítulo plantea la modelización y el procedimiento diseñado para analizar la temperatura diaria máxima. A continuación, se resumen los resultados sobre las series de la España peninsular. Por último, se incluyen las conclusiones y opciones de trabajo futuro. Se han incluido Anexos relativos a mostrar algunos resultados de interés que completan la información de la memoria.

Capítulo 2

Fundamentos teóricos

Este capítulo plantea una breve introducción a los modelos bayesianos de regresión que permiten abordar la modelización para la media y la varianza y además incorporar componentes espaciales. Se incluyen subsecciones para revisar los conceptos básicos de modelos de regresión, de inferencia bayesiana y de modelos jerárquicos. Por último, se revisa la línea de trabajo sobre la modelización de series de temperatura máxima diaria que se va a seguir en el TFG.

2.1. Modelos bayesianos de regresión

2.1.1. Modelos de regresión

Comenzamos dentro del marco clásico de inferencia estadística. Revisamos los conceptos básicos sobre modelos de regresión lineal múltiple (MRLM). La notación empleada a continuación difiere con la habitual en el modelo MRLM. Esta decisión es con el fin de tener una notación compacta con respecto a los modelos que propondremos posteriormente. Merece la pena aclarar que s es fija y única para cada modelo ajustado. Sea $Y_t(s)$ una variable aleatoria normal conocida como la variable dependiente o respuesta. Dada una muestra, de tamaño n, $(X_{t,1}(s), X_{t,2}(s), ..., X_{t,n}(s))$ (variables independientes, predictoras o covariables del modelo), definimos la ecuación:

$$Y_t(s) = \beta_0(s) + \beta_1(s)X_{t,1}(s) + \beta_2(s)X_{t,2}(s) + ... + \beta_n(s)X_{t,n}(s) + \varepsilon(s). \tag{2.1}$$

Donde $\varepsilon(s) \sim \mathcal{N}(0, \sigma(s))$ y aparecen los parámetros fijos $\beta_i(s) \in \mathbb{R} \quad \forall i \in \{0, 1, 2, ..., n\}$. De esto deducimos que

$$E[Y_t(s)|X_t(s)] = \beta_0(s) + \beta_1(s)X_{t,1}(s) + \beta_2(s)X_{t,2}(s) + ... + \beta_n(s)X_{t,n}(s) \tag{2.2}$$

y

$$Var(Y_t(s)|X_t(s)) = \sigma^2(s), \tag{2.3}$$

es decir,

$$Y_t(s)|X_t(s) \sim \mathcal{N}(\beta_0(s) + \beta_1(s)X_{t,1}(s) + \beta_2(s)X_{t,2}(s) + ... + \beta_n(s)X_{t,n}(s), \sigma(s)). \tag{2.4}$$

Notar que la varianza es constante $\forall t$. Esto se conoce como homocedasticidad del modelo y es importante, ya que indica que $\sigma(s)$ es independiente de las covariables observadas. El principal objetivo de los MRLM es, por tanto, estimar los parámetros $\beta_0(s), \beta_1(s), \beta_2(s), ..., \beta_n(s), \sigma(s)$. Para realizar esta estimación se utiliza el estimador máximo verosímil (EMV), cuyo objetivo es estimar los coeficientes de regresión que maximizan la función de verosimilitud. En este caso la función de verosimilitud es

$$L(\boldsymbol{\beta(s)}, \sigma(s)|\mathbf{Y(s)}) = \prod_t \frac{1}{\sqrt{2\pi\sigma^2(s)}} \exp\left\{ - \frac{(y_t(s) - (\beta_0(s) + \beta_1(s)X_{t,1}(s) + \ldots + \beta_n(s)X_{t,n}(s)))^2}{2\sigma^2(s)} \right\} \tag{2.5}$$

donde $y_t(s)$ es el valor observado de la variable dependiente para el punto de observación t, $\boldsymbol{\beta(s)}$ es el vector de coeficientes e $\mathbf{Y(s)}$ el vector que contiene a $y_t(s)$ en el instante temporal t. En este caso, el EMV de los parámetros del predictor lineal coincide con el estimador de mínimos cuadrados ordinario. El estimador EMV de $\sigma^2(s)$ es ligeramente distinto, ya que en mínimos cuadrado el cociente es entre los grados de libertad. Para comprobar si una covariable es importante o no para explicar la variable respuesta en el modelo, utilizamos el test t. Una vez obtenida la estimación de los parámetros, $(\tilde{\beta}_0(s), \tilde{\beta}_1(s), ..., \tilde{\beta}_n(s))$, los usamos para obtener lo que conocemos como valores ajustados, $\tilde{Y}_t(s)$, mediante la expresión:

$$\tilde{Y}_t(s) = \tilde{\beta}_0(s) + \tilde{\beta}_1(s)X_{t,1}(s) + ... + \tilde{\beta}_n(s)X_{t,n}(s). \tag{2.6}$$

Definimos los residuos del modelo como $e = \mathbf{Y}(s) - \tilde{\mathbf{Y}}(s)$. Cuando el modelo es correcto se deberían cumplir las siguientes propiedades:

- $\sum_t e_t(s) = 0$
- Los residuos deben ser independientes entre sí.
- Los residuos deben seguir una distribución normal.
- La homocedasticidad de las variables dependientes.

En los modelos anteriores se ha supuesto que la varianza es constante (homocedasticidad). Sin embargo, en muchos contextos, esta suposición no es realista. Por ejemplo, en el problema que vamos a abordar a lo largo de este trabajo. En este caso, se trabajará con un modelo GAMLSS (Generalized Additive Models for Location, Scale and Shape), Stasinopoulos & Rigby (2008) [7], que tiene como fórmula condicional:

$$\mathbf{Y}(s)|\mathbf{X}(s) \sim \mathcal{N}(\mu(s) = \mathbf{X}(s)\boldsymbol{\beta}^{\mu}(s), \sigma(s) = g(\mathbf{X}(s)\boldsymbol{\beta}^{\sigma}(s))). \tag{2.7}$$

Denotamos por $\boldsymbol{\beta}^{\mu}$ a los parámetros del modelo de la media, mientras que por $\boldsymbol{\beta}^{\sigma}$ a los del de la desviación típica. Vamos a considerar el caso $g(x) = e^x$. Las estimaciones de estos parámetros se obtienen, de nuevo, mediante la maximización de la función de verosimilitud, pero, en este caso, σ ya no es constante. La función a maximizar es

$$L(\boldsymbol{\beta}^{\mu}(s), \boldsymbol{\beta}^{\sigma}(s)|\mathbf{Y}(s)) = \prod_t f(y_t(s)|\mu_t(s), \sigma_t(s)), \tag{2.8}$$

donde $f(.|\mu, \sigma)$ es la función de densidad de una $\mathcal{N}(\mu, \sigma)$. Una vez obtenidas las estimaciones de los parámetros, se dispone de un modelo con una distribución normal definida como sigue:
Si $(\tilde{\beta}_0^{\mu}(s), \tilde{\beta}_1^{\mu}(s), ..., \tilde{\beta}_n^{\mu}(s), \tilde{\beta}_0^{\sigma}(s), ..., \tilde{\beta}_n^{\sigma}(s))$ son las estimaciones de los parámetros del modelo y de ellas se obtienen

$$\tilde{\mu}_t(s) = \tilde{\beta}_0^{\mu}(s) + \tilde{\beta}_1^{\mu}(s)X_{t,1}(s) + ... + \tilde{\beta}_n^{\mu}(s)X_{t,n}(s) \tag{2.9}$$

y

$$\tilde{\sigma}_t(s) = exp(\tilde{\beta}_0^{\sigma}(s) + \tilde{\beta}_1^{\sigma}(s)X_{t,1}(s) + ... + \tilde{\beta}_n^{\sigma}(s)X_{t,n}(s)), \tag{2.10}$$

entonces $\tilde{Y}_t(s) \sim \mathcal{N}(\tilde{\mu}_t(s), \tilde{\sigma}_t(s))$.

2.1.2. Inferencia bayesiana

Hasta ahora hemos presentado modelos en los cuales todos los parámetros a estimar se consideran valores fijos. La inferencia bayesiana se basa en la idea principal de que estos parámetros no son constantes, sino que siguen una distribución, es decir, son variables aleatorias. Se basa en la actualización de nuestras creencias a priori sobre un parámetro, θ, desconocido, después de observar nuevos datos, Y. La inferencia bayesiana se describe con detalle en los primeros capítulos del libro "Bayesian Data Analysis" del profesor Andrew Gelman et al. [3]. Usaremos la notación de este libro, θ para los parámetros desconocidos, Y para los datos observados y $p(.)$ para la función de densidad (o masa de probabilidad en el caso de distribuciones discretas). Este método de estimación se basa principalmente en la regla de Bayes,

para calcular la distribución de probabilidad a posteriori de los parámetros θ, dados los datos observados Y. El problema principal es obtener el modelo de probabilidad conjunto, $p(\theta, Y)$, y, a partir de él, extraer la densidad a posteriori $p(\theta|Y)$ (notar que Y es fijo). Se dispone de la expresión dada por

$$p(\theta|Y) = \frac{p(\theta)p(Y|\theta)}{p(Y)}, \tag{2.11}$$

donde $p(\theta)$ es la función de densidad de la distribución a priori de los parámetros y $p(Y|\theta)$ es la función de verosimilitud. De esta fórmula extraemos que $p(\theta|Y) \propto p(\theta)p(Y|\theta)$, lo cual es suficiente para obtener la distribución a posteriori, sin la necesidad de conocer la distribución marginal de Y. De la fórmula 2.11 se deduce la siguiente expresión:

$$p(\theta, Y) = p(\theta|Y)p(Y) = p(Y|\theta)p(\theta). \tag{2.12}$$

Se puede realizar inferencia predictiva sobre datos observables desconocidos, $\tilde{Y}$. Se define la densidad a priori de $\tilde{Y}$, usando la fórmula

$$p(\tilde{Y}) = \int_\theta p(\tilde{Y}, \theta) d\theta = \int_\theta p(\theta)p(\tilde{Y}|\theta) d\theta. \tag{2.13}$$

Una vez observados los datos Y, la distribución a posteriori de $\tilde{Y}$ se define mediante la distribución predictiva a posteriori, $p(\tilde{Y}|Y)$, que se obtiene mediante la fórmula

$$p(\tilde{Y}|Y) = \int_\theta p(\tilde{Y}, \theta|Y) d\theta = \int_\theta p(\tilde{Y}|\theta, Y)p(\theta|Y) d\theta = \int_\theta p(\tilde{Y}|\theta)p(\theta|Y) d\theta. \tag{2.14}$$

Además, la esperanza condicional, en el caso de normalidad de los datos y parámetros, $Y \sim \mathcal{N}(\mu, \sigma)$ sigue la fórmula

$$E(\tilde{Y}|Y) = E(E(\tilde{Y}|\mu, Y)|Y) = E(\mu|Y). \tag{2.15}$$

Y para la varianza condicional tenemos: $Var(\tilde{Y}|Y) = E(\sigma^2|Y) + Var(\mu|Y)$.

La inferencia bayesiana implica calcular la distribución a posteriori de los parámetros, lo que para módelos muy complejos puede ser muy costoso. A este problema le viene acompañado una solución, el método de Markov Chain Monte Carlo (MCMC), técnica estadística utilizada para conseguir muestras de la distribución a posteriori de los parámetros, sin la necesidad de obtenerla explícitamente. Este método se basa en la construcción de una cadena de Markov, cuya distribución estacionaria corresponde con la distribución a posteriori buscada. Después de simular la cadena una cantidad de iteraciones adecuada, se obtienen muestras representativas de la distribución a posteriori.

Un concepto de interés, es el de los intervalos de credibilidad al 95% para un parámetro θ. Este intervalo tiene como límites inferior y superior los cuantiles 0.025 y 0.975 de la distribución a posteriori del parámetro de interés, respectivamente. Se trata de un intervalo en el cual la probabilidad de que un valor no observado del parámetro caiga es de 0.95.

2.1.3. Modelos bayesianos jerárquicos

Se presenta el marco de los modelos bayesiano jerárquico, al que se ajustarán los modelos en este TFG donde la variable respuesta $Y_t(s)$ tiene distribución normal. Su esperanza es $\mu_t(s) = E[Y_t(s)|X_t(s)]$ y su varianza $\sigma_t^2(s) = Var[Y_t(s)|X_t(s)]$, es decir $Y_t(s)|X_t(s) \sim N(\mu_t(s), \sigma_t(s))$. En un modelo bayesiano los parámetros de los predictores lineales son variables aleatorias y, siendo jerárquico, a su vez las distribuciones a priori de dichos parámetros dependen de hiperparámetros con distribuciones que, en general, son no informativas.

Los modelos BAMLSS (Bayesian additive models for location, scale, and shape), Umlauf et al (2018) [9], extienden los GAMLSS al marco bayesiano. La estimación mediante la librería *bamlss* Umlauf et al (2021) [10] establece que los parámetros del predictor lineal de la media y de la desviación típica son variables aleatorias con distribución $N(\mu_{\beta_i(s)}, \sigma_{\beta_i(s)})$. Esos hiperparámetros $\mu_{\beta_i(s)}$ tienen una distribución

a priori normal no informativa. Para los hiperparámetros de variabilidad $\sigma_{\beta_i(s)}$ se utiliza como distribución a priori una distribución inversa Gamma, IG(0.001, 0.001).

El algoritmo de estimación obtiene la distribución a posteriori mediante un MCMC específico denominado algoritmo Metropolis-Hastings. Esta librería permite la extracción de la distribución a posteriori con funciones adecuadas, tanto para inferencia en la esperanza como en la varianza.

2.2. La modelización de las series diarias de temperatura

El problema de la modelización de la temperatura diaria máxima se ha abordado por Castillo-Mateo et al (2022) [1]. El artículo propone un modelo bayesiano jerárquico espacio-temporal que modela la temperatura máxima diaria durante los veranos, en la región de la Cuenca del Ebro. Pese a su tamaño reducido, la región muestra un relieve diverso, con un clima cálido y poco variable en el centro y una mayor variabilidad en las montañas. Los resultados instan la necesidad de incluir armónicos para explicar la estacionalidad y considerar una tendencia temporal a largo plazo y estructura autorregresiva. La autorregresión se refleja mediante la inclusión de covariables basadas en la temperatura máxima registrada el día anterior. Además, la dependencia espacial del modelo se aborda modelando los interceptos, los coeficientes de las covariables y las varianzas, utilizando variables geográficas. La dependencia espacial del modelo se captura por una parte mediante covariables geográficas y por otra parte haciendo los coeficientes espacialmente variables modelados mediante procesos gaussianos. Como resultado de la inferencia se encuentra una variabilidad espacial importante en la tendencia. Asímismo, los autores muestran que el modelo propuesto es adaptable a cualquier ubicación, incluso donde no hay observaciones, incluyendo covariables geográficas y usando un kriging bayesiano para obtener el valor de los parámetros espacialmente variables en localizaciones no observadas.

En el artículo de Castillo-Mateo et al (2023) [2], se presenta un modelo bayesiano para los cuantiles de la temperatura máxima diaria, que es autorregresivo y espacial, con el objetivo de estudiar los cambios en la distribución de las temperaturas máximas diarias en los veranos, a lo largo de 60 años, en la región de la Cuenca del Ebro. La altitud es la única covariable espacial que se considera en el modelo. Una vez ajustados los modelos para los cuantiles $\tau = 0.05, 0.5, 0.95$, vemos distintos patrones espaciales y temporales. Esto nos muestra la diferencia entre el comportamiento en los cuantiles extremos, frente al cuantil central de la temperatura máxima diaria. Por ejemplo, el cuantil 0.95 ha aumentado más en áreas de la depresión del Ebro; mientras que en las zonas del noroeste, el cuantil 0.05 no ha aumentado. Estas diferencias son un indicativo de la necesidad de modelar toda la distribución de la temperatura máxima diaria, y no solo la media. En particular, indican que existe heterocedasticidad a lo largo del tiempo.

Las conclusiones de ambos estudios indican la obligación de representar la variabilidad espacial y temporal en el modelado de la temperatura máxima diaria. Con estos antecedentes, un modelo para la temperatura diaria debe incluir una modelización también para la varianza, a diferencia de los modelos descritos en la subsección 2.1.1. Si bien usaremos los resultados de estos artículos como referencias de interés, en nuestro trabajo utilizaremos los datos de todo el año, no solo los del verano, además de trabajar con toda España, a diferencia de la región más reducida que usan estos artículos.

Capítulo 3

Metodología propuesta para la estimación del modelo

En este capítulo se presenta la metodología diseñada en el TFG para la estimación de un modelo de regresión para la serie de la temperatura diaria. En primer lugar, se propone un modelo que recoge las características de la distribución de la temperatura diaria que se han descrito previamente, estacionalidad, memoria y cambio en la variabilidad. Se ha diseñado un procedimiento para plantear la modelización. En segundo lugar, se presentan las herramientas gráficas y numéricas utilizadas para explorar las fuentes de variación que afectan a la respuesta, así como la forma de su efecto sobre la respuesta. Para este objetivo, se usan como instrumentos exploratorios algunos modelos lineales. En tercer lugar, se presenta la estrategia de modelización en el ámbito bayesiano. La última sección presenta resumidamente la librería de funciones desarrollada en el software libre R.

3.1. Modelo propuesto para la temperatura diaria

El modelo propuesto para la temperatura diaria corresponde a una distribución normal cuyos parámetros esperanza y varianza se expresan en las ecuaciones 3.1 y 3.2. La estructura recoge las ideas de Castillo-Mateo et al (2023) [2], aunque el número de armónicos debe atender a que la base de datos contiene información de todo el año y no solo del verano. Otra diferencia con los trabajos de Castillo-Mateo et al es que se considera una estructura autorregresiva que incorpora información de los dos días previos, atendendiendo a que la puede haber una mayor inercia térmica fuera del verano.

$$\begin{aligned} E[Y_t(s)|X_t(s)] = \beta_0(s) + \sum_{i=1}^{2}(\beta_{i,s}(s)Si_t + \beta_{i,c}(s)Ci_t) + \Big(\alpha_0(s) + \sum_{i=1}^{2}(\alpha_{i,s}(s)Si_t + \alpha_{i,c}(s)Ci_t)\Big)Year_t \\ + \Big(\rho_0(s) + \sum_{i=1}^{2}(\rho_{i,s}(s)Si_t + \rho_{i,c}(s)Ci_t)\Big)Y_{t-1}(s) + \Big(\gamma_0(s) + \sum_{i=1}^{2}(\gamma_{i,s}(s)Si_t + \gamma_{i,c}(s)Ci_t)\Big)Y_{t-2}(s). \end{aligned} \tag{3.1}$$

$$\begin{aligned} Var[Y_t(s)|X_t(s) = \exp\Big\{\beta_0^{\sigma}(s) + \sum_{i=1}^{2}(\beta_{i,s}^{\sigma}(s)Si_t + \beta_{i,c}^{\sigma}(s)Ci_t) + \Big(\alpha_0^{\sigma}(s) + \sum_{i=1}^{2}(\alpha_{i,s}^{\sigma}(s)Si_t + \alpha_{i,c}^{\sigma}(s)Ci_t)\Big)Year_t \\ + \Big(\rho_0^{\sigma}(s) + \sum_{i=1}^{2}(\rho_{i,s}^{\sigma}(s)Si_t + \rho_{i,c}^{\sigma}(s)Ci_t)\Big)Y_{t-1}(s) + \Big(\gamma_0^{\sigma}(s) + \sum_{i=1}^{2}(\gamma_{i,s}^{\sigma}(s)Si_t + \gamma_{i,c}^{\sigma}(s)Ci_t)\Big)Y_{t-2}(s)\Big\}. \end{aligned} \tag{3.2}$$

Donde la serie $Y_t(s)$ corresponde al valor de la temperatura en el día t y en la posición s, el vector de covariables incluye la información asociada a la fecha: los componentes armómicos Si_t, Ci_t, con $Si_t = \sin\frac{2\pi ti}{365}$ y $Ci_t = \cos\frac{2\pi ti}{365}$, $Year_t$ el año y los valores previos $Y_{t-1}(s)$ e $Y_{t-2}(s)$. Utilizamos el superíndice σ para los parámetros del modelo de la varianza. A su vez, denotamos como $\alpha_i(s)$ a los parámetros asociados la tendencia temporal que representa la covariable year. $\rho_i(s)$ y $\gamma_i(s)$ están asociados a los

efectos del día previo y dos días antes, respectivamente. En la base de datos se dispone de los lugares $s_k \in \{s_1, \ldots, s_K\}$.

3.2. Herramientas exploratorias

El análisis exploratorio representa mediante la media móvil y los cuantiles móviles, tanto la evolución a lo largo del año de estas características, para mostrar su comportamiento estacional, como el cambio observado entre el inicio y el final de las series, es decir, para describir el efecto del cambio en el largo plazo.

Posteriormente, se ajustan modelos con la misma estructura para la media en cada lugar, y se analiza si la variabilidad espacial de los parámetros estimados puede explicarse mediante variables geográficas o climáticas.

Se ajusta un modelo de regresión lineal para estudiar la media de la serie de temperatura diaria máxima en cada lugar s_k, con la estructura de la expresión 3.1. Para fijar el número de armónicos en este modelo, se ha analizado su efecto en todos los lugares, de modo que se decide la eliminación de un armónico de orden superior atendiendo a la mejora del R^2_{adj}. Para ello, se representa mediante un boxplot el cambio en R^2_{adj} para cada lugar cuando se incluye un nuevo armónico. Cuando la caja incluye al valor 0 o se sitúa en valores negativos, entonces se considera que está sobreparametrizado y que el armónico no se requiere en el modelo.

El modelo se ajusta mediante la función lm de R.

El valor estimado para el coeficiente de cada término (incluido el intercepto) en el lugar s_k se incluye en la posición k del vector correspondiente. Así, obtenemos 20 vectores que contienen cada uno el coeficiente estimado en cada lugar asociado a una covariable. Por ejemplo, el primer vector tiene almacenados los interceptos de todos los lugares.

Se estima un modelo para la varianza a partir de un ajuste lm sobre la respuesta definida por el cuadrado de los residuos del modelo anterior. Una vez obtenido en R el modelo 3.1 para cada lugar s_k, se compara $\mathbf{Y}(s)$ con el valor medio ajustado $\tilde{\mathbf{Y}}(s)$, que se ha obtenido sustituyendo los valores observados de las variables independientes (como la temperatura máxima del día anterior, el año, etc.) en la regresión ajustada, es decir, $E[\mathbf{Y}(s)|\mathbf{X}(s)]$. Esto genera una "predicción"de $\mathbf{Y}(s)$ basada en el modelo. Calculamos $Y_t^*(s) = (Y_t(x) - \tilde{Y}_t(s))^2$. Como antes, buscamos un modelo para obtener $E[Y_t^*(s)|X_t(s)] = E[(Y_t(s) - \tilde{Y}_t(s))^2|X_t(s)]$. Puesto que $\tilde{Y}_t(s)$ estima $E[Y(s)|\mathbf{X}]$, $E[Y_t^*(s)|X_t(s)]$ no es más que $Var[Y_t(s)|X_t(s)]$. Le aplicamos el logaritmo neperiano a $Y_t^*(s)$. Operativamente, se ajusta un modelo de regresión lineal para $\log(Y_t^*(s))$ en cada lugar s_k, utilizando el predictor lineal que aparece en la expresión 3.2. Estaríamos creando un modelo para estimar $Var[Y_t(s)|\mathbf{X}_t]$. Con estos modelos se trabajará de modo similar a como se ha propuesto con los modelos exploratorios para el valor medio.

3.2.1. Variabilidad espacial de los parámetros de los predictores lineales

A partir de las estimaciones en cada punto, se pretende estudiar si la variabilidad que muestran los modelos ajustados puede explicarse con variables geográficas (altitud y posición), o con variables climáticas que caracterizan los tipos de clima.

En primer lugar, se ajustan modelos de regresión para representar las relaciones con las variables geográficas. Para cada parámetro, se consideran como respuesta las estimaciones en cada punto de la red de observatorios y se ajusta un modelo de regresión frente a longitud, latitud y altitud, que intervienen mediante splines y considerando interacciones. Se compara el grado de ajuste de estos modelos con el obtenido por modelos que incorporan 4 variables que caracterizan climáticamente cada observatorio. Estas variables representan el valor medio y la variabilidad de la temperatura diaria máxima en verano y en invierno. Se considera, para cada lugar, la media de las temperaturas máximas diarias de junio, julio y agosto de los últimos 30 años, la media para diciembre, enero y febrero, el rango intercuartílico de los mismos datos de junio, julio y agosto (diferencia entre sus cuantiles 0.75 y 0.25) y el rango intercuartílico de invierno. Con esto obtendríamos una valoración sobre la capacidad de tener un modelo para

representar la temperatura máxima diaria a partir de la expresión 3.1 para un lugar del cual conocemos únicamente su posición geográfica o resúmenes climáticos.

3.3. Modelizacion bayesiana

Procedemos a usar la librería de R *bamlss*, exclusivamente para ajustar un modelo a la serie de un lugar. Con esta función lo que pretendemos es estimar $\tilde{\mu}_t$ (la media) y $\tilde{\sigma}_t$ (la desviación típica) conjuntamente, de manera que, en el día t, $\tilde{Y}_t \sim \mathcal{N}(\tilde{\mu}_t, \tilde{\sigma}_t)$. Para modelar la media y la desviación típica utilizaremos los predictores lineales de las expresiones 3.1 y 3.2, usando las a priori e hiperparámetros que se han citado en la sección 2.1.3. Una vez obtenidas las distribuciones a posteriori de los parámetros de las covariables existen opciones para estimar $\tilde{\mu}_t$ y $\tilde{\sigma}_t$, con la media a posteriori de cada parámetro, con la moda o con el percentil 50. Nosotros hemos decidido utilizar la media.

Tras estimar el modelo, se obtienen los intervalos de credibilidad de cada parámetro, tanto para la media como para la desviación típica. Por una parte, estos intervalos permiten considerar la eliminación de alguna covariable, mediante el siguiente procedimiento: si el valor 0 pertenece al intervalo de credibilidad del parámetro asociado a la covariable i, entonces creamos un nuevo modelo eliminando la covariable i del respectivo predictor lineal. Para ver cual de los dos modelos es mejor, mediante la función DIC en R, el que presenta menor DIC es considerado "mejor". Una vez aplicado este proceso las ocasiones necesarias, obtendremos un modelo final. Este procedimiento no se aplica individualmente a un término seno o coseno de un armónico para evitar que la estacionalidad no esté bien representada.

Una vez obtenido el modelo final, se han diseñado herramientas de crítica para comprobar su calidad. Se construyen qqplots basados en que si el modelo representa adecuadamente a la respuesta, se debe cumplir $F(y_t|\tilde{\mu}_t, \tilde{\sigma}_t) \sim \text{Uniform}(0,1)$. También generaremos muestras de las distribuciones a posteriori de algunos parámetros de interés, como el coeficiente que expresa la tendencia temporal a largo plazo, asociado al año. Además, se comparan los intervalos de credibilidad de los parámetros con las estimaciones obtenidas en el modelo exploratorio.

3.3.1. Modelo bayesiano espacio-temporal

Se estima un modelo bayesiano para analizar conjuntamente las series de todos los lugares. Para ello, creamos una nueva variable respuesta que, en forma de vector, definimos de la siguiente manera: $(Y(s_1), Y(s_2), ..., Y(s_k))$. Con la misma idea, hacemos lo mismo para la matriz de covariables, con el día anterior, año, los armónicos, etc. Creamos otras 3 variables, que son las que van a distinguir un lugar de otro, un vector Alt, cuyas T primeras componentes son la repetición de $Altitud(s_1)$, sus T segundas la repetición de $Altitud(s_2)$, y así sucesivamente. De forma análoga se definen las variables con la latitud (Lat) y con la longitud (Lon). Esto hará que el modelo distinga un lugar de otro, gracias a los parámetros asociados a estas últimas 3 variables, y así tratar de explicar la variabilidad espacial con el fin de extender este modelo a cualquier punto de la Península. La expresión del modelo será la misma, tanto para la media como para la desviación típica, aunque incluirá el efecto aditivo de los 3 vectores geográficos creados. Una vez obtenido el modelo, compararemos los intervalos de credibilidad de los parámetros obtenidos con los del modelo individual, presentado en la subsección anterior.

3.4. Software desarrollado

Los procedimientos se han implementado en R, usando la version 4.1.2. El resultado ha sido la libreria de funciones descrita en el Anexo A.A. Además, está a disposicion de otros usuarios en GitHub: https://github.com/jorgenl21/TFG-Unizar/tree/612e00525529e2ea21860813e522e0a24317fece

Capítulo 4

Resultados de la modelización de la temperatura máxima diaria

En este capítulo se resumen los resultados de la aplicación de la modelización estadística, incluyendo la fase exploratoria y la inferencia del modelo bayesiano. Los resultados se completan con la información del Anexo A.B.

4.1. Base de datos

Se ha construido una base de datos con las series de temperatura máxima diaria en 40 observatorios españoles en la Península Ibérica, en el periodo 1965-2022. Los datos se han obtenido desde el repositorio del proyecto European Climate Assessment & Dataset (ECA), Klein Tank et al (2002) [8], https://www.ecad.eu/dailydata.

Comenzamos descargando los datos de la temperatura máxima diaria de España y Portugal, hasta 2022, de todas las estaciones metereológicas disponibles. De esto se desglosa una base de datos, que contiene el nombre de cada estación con su altitud, longitud y latitud. Para cada lugar, tenemos la temperatura máxima acompañada de la fecha y un indicativo de calidad. Puesto que la fecha de inicio del registro es distinta en cada lugar, se ha seleccionado como fecha de inicio general para todos los lugares el 1 de enero de 1955. En cuanto al indicativo de calidad, un valor de 0 indica que la temperatura máxima medida ese día es correcta, un valor de 1 considera como sospechosa esa medida, mientras que un valor 9 nos indica que el dato está perdido. Usamos estos datos para crear dos bases de datos. La primera contiene, en orden, las columnas año, mes y día. La base de datos dispone para cada lugar de la columna correspondiente a las temperaturas máximas diarias en la estación s_k, junto con la columna de sus indicadores de calidad. Es decir, con la notación que estamos utilizando, el dato de la fila t en la columna $2+2k$ corresponde con $Y_t(s_k)$. Otras columnas de esta base de datos expresan el día del año y el día desde el inicio de la serie. La primera modificación que hacemos a esta base es considerar como perdido (NA) todas las temperaturas con indicador de calidad dudosa, asociado 1 o 9. Posteriormente, eliminaremos las filas asociadas al 29 de febrero. Finalmente, la base de datos de trabajo descarta los lugares con gran cantidad de datos perdidos. La base de datos de información geográfica y climática contiene las variables, con el nombre de la estación correspondiente (en el mismo orden de aparición que la base anterior), junto con latitud, longitud y altitud y las variables climáticas presentadas en la sección 3.2.1.

4.2. Análisis exploratorio

Esta sección describe algunas características de la distribución y de la evolución de la temperatura máxima diaria, así como un análisis exploratorio de la estructura del predictor lineal para los submodelos

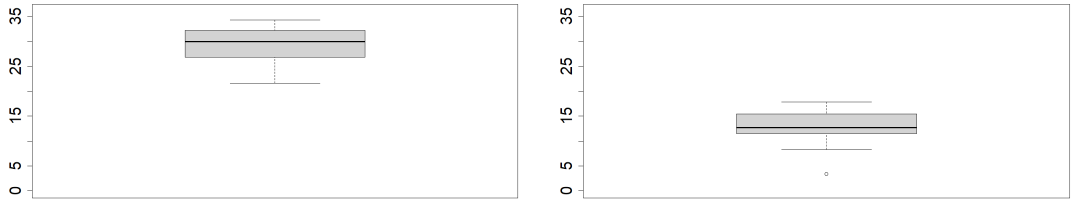


Figura 4.1: Diagrama de caja del valor medio de la temperatura máxima diaria en los 40 observatorios, obtenido en el periodo de referencia 1992-2022, para los meses de junio, julio y agosto (izd.) y para diciembre, enero y febrero (dch.)

de media y de varianza, que deben tenerse en cuenta en la posterior modelización bayesiana. El Anexo A.B contiene información complementaria sobre este análisis exploratorio.

4.2.1. Características de las series diarias de temperatura máxima

Los diagramas de cajas de la Figura 4.1 permiten comparar el diagrama de caja de los valores medios de verano (junio, julio, agosto) y los de invierno (diciembre, enero, febrero), como resúmenes de la distribución obtenidos en el periodo de referencia 1992-2022. El valor medio de verano varía entre observatorios de 20°C a 34°C. En invierno, figura derecha, el valor medio más elevado es 18°C y el mínimo es de unos 4°C. Este es un dato atípico y pertenece a la estación de montaña de Navacerrada, la de mayor altitud en la base de datos. Se puede apreciar que la variabilidad entre estaciones del año en un observatorio es de orden similar a la variabilidad entre observatorios en una misma estación del año.

La Figura 4.2 representa la media por mes de los últimos 31 años, en lugares de la Península con climas distintos: Badajoz, Málaga-aeropuerto y San Sebastián. Vemos que la mayor variabilidad a lo largo del año se presenta en Badajoz (interior). Málaga sigue un patrón parecido, pero con meses de invierno más cálidos (costa sur mediterránea). En San Sebastián aparecen valores mucho menores y menos variables (costa del mar Cantábrico).

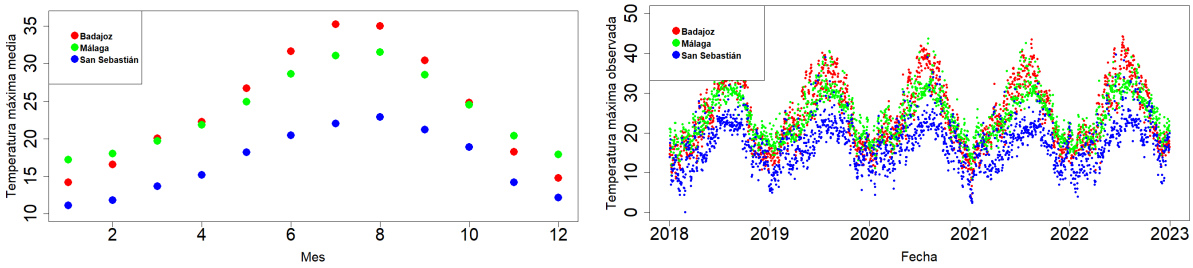


Figura 4.2: Izd.: Valor medio mensual en Badajoz, Málaga-aereopuerto y San Sebastián, calculado en el periodo 1991-2022. Dch.: serie diaria en 2018-2022 en los 3 observatorios.

En el panel derecho de la figura se representa la temperatura máxima diaria en los últimos 5 años, en los lugares elegidos anteriormente. Claramente, podemos apreciar un patrón en la distribución anual de las temperaturas máximas con épocas de máximos y mínimo comunes en los tres lugares, con los mayores valores presentes en verano, y los menores en invierno, pero el ciclo estacional no se puede considerar idéntico. Esta tendencia periódica anual, y los resultados vistos en la sección 2.2, nos sugieren la necesidad de incluir variables armónicas, de periodo 365, en el modelo para explicar el comportamiento de la temperatura máxima diaria. Además, observamos proximidad de puntos sucesivos, es decir, no se obser-

van saltos pronunciados en los valores sucesivos. Esto es un indicativo de la demanda de incluir variables asociadas a temperaturas máximas en los días previos (variables autorregresivas).

Se describe el cambio observado en la distribución durante el periodo de interés. Para ello se comparan los percentiles obtenidos en 1955-1966 y los obtenidos en 2011-2022. Definimos ahora los percentiles 90, 10 y 50 (la mediana) en una ventana móvil de 30 días análogamente a la media móvil de 30 días. Esta medida, en el caso del cuantil 0.9, podemos interpretarla para cada día como la temperatura que ha sido superada solo en el 10 % de los días en ese intervalo de 30 días. En el caso del cuantil 0.10, como la temperatura inferior al 90 % de los días. Mientras que para la mediana, consiste en la temperatura mayor al 50 %. Un aumento en el percentil 90 móvil 30 días a lo largo del tiempo es un indicativo de la presencia de temperaturas altas cada vez más extremas. Por otro lado, un aumento en el cuantil 0.10 indica que las temperaturas más bajas actuales son menos frías que 50 años atrás. Con respecto a la mediana, indicaría que las temperaturas medias están aumentando. Esto supondría un cambio general en la distribución de la temperatura máxima diaria, tendiendo hacia valores más cálidos. Vamos a plasmar este cambio mediante gráficas. En un observatorio se obtiene el percentil 90 móvil 30 días en cada día del año, correspondiente a los últimos 12 años (2011-2022), y los mismos cálculos para los primeros 12 años (1955-1966). Se obtiene la diferencia entre ambas (la correspondiente a los últimos doce años menos la respectiva a los doce primeros). Lo mismo se realiza para los cuantiles restantes.

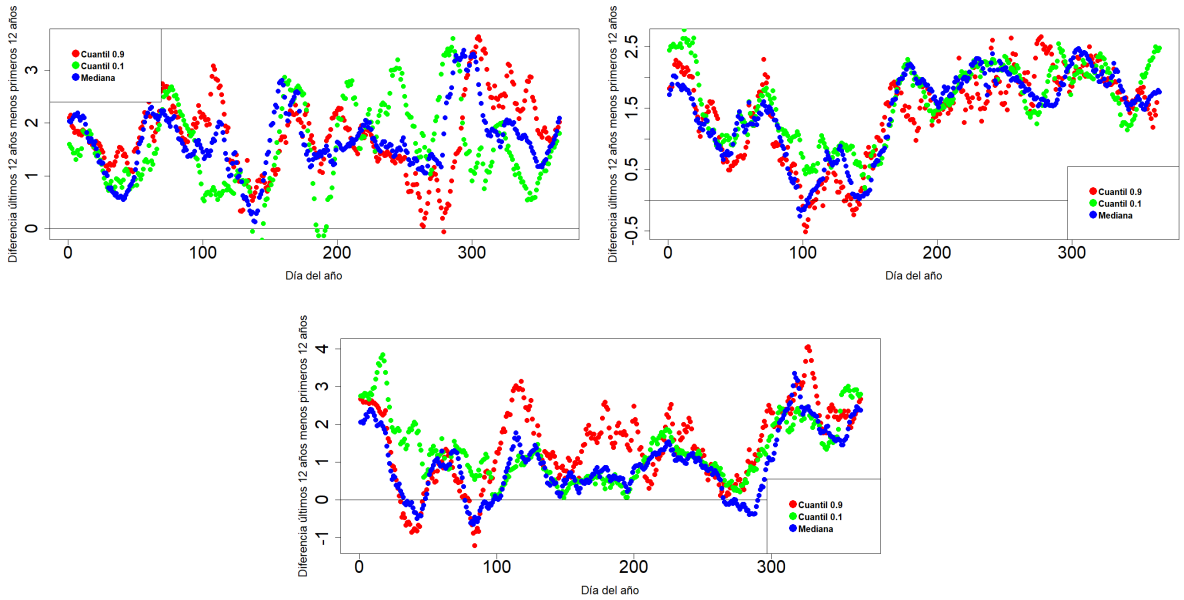


Figura 4.3: Cambio desde 1955-1966 a 2011-2022 en los percentiles móviles, con ventana de 30 días, en Badajoz, Málaga-aeropuerto y San Sebastián. Se representa en rojo el percentil 0.9, en azul el 0.5 y en verde el 0.1.

Representamos estos resultados en la Figura 4.3. En Badajoz, podemos apreciar que prácticamente todas las diferencias son positivas. Además observamos un patrón común, en la primera mitad de año y a finales de este. Esto nos muestra un cambio en la distribución de las temperaturas máximas diarias en este lugar.

En Málaga salvo en unos días de primavera, obtenemos una diferencia positiva. En general, se observa el mismo patrón en el incremento de los distintos cuantiles, a lo largo de todo el año. Esto nos muestra, que la distribución ha aumentado en los extremos y en la mediana de la misma manera. En San Sebastián, quitando unos días, de nuevo vemos un aumento, que ha llegado hasta a los 4°C en invierno en el percentil 0.1. Durante los primeros 100 días, la mediana y el cuantil 90 han cambiado siguiendo el mismo patrón. A partir de esas fechas, vemos el mismo patrón en la mediana y el cuantil 0.1 pero se desacopla el cambio en el 0.9.

Con lo obtenido en las tres gráficas, concluimos con la idea de que se observa un cambio en la

distribución de las temperaturas máximas diarias, a lo largo de los años. Además, estos incrementos en los cuantiles, parecen guardar cierta relación entre ellos pero no son coincidentes. Estos resultados, nos proporcionan la necesidad de incluir la variable año, en cualquier modelo que trate de explicar la distribución de la temperatura máxima diaria. Volveremos a ver estos resultados con el coeficiente de regresión de la variable año del módelo para el valor medio de la temperatura máxima diaria, y de la varianza.

4.2.2. Análisis exploratorio para el modelo para el valor medio y para la varianza

En esta sección explicaremos los resultados obtenidos con respecto al modelo 3.1 y 3.2. Se utiliza la función lm en R y se almacenan los resultados de cada lugar. Se ha trabajado con un nivel de significación de 0.05 para los test de tipo t que se han usado, por ejemplo en la selección del número de armónicos.

Respecto a la modelización del valor medio, en cada uno de los lugares obtenemos un R^2_{adj} elevado, lo que nos indica la buena calidad de los modelos. El R cuadrado ajustado más elevado es de 0.92, mientras que el más pequeño es de 0.71. En la Figura 4.4 se representa un mapa con los K lugares, el color del punto corresponde a una escala progresiva del azul al rojo (pasando por el blanco) que expresa el valor del R^2_{adj}. Podemos ver que la calidad del modelo es mejor en las zonas del centro de la Peninsula, mientras que en la costa es peor (sobre todo en la costa norte). Un modelo basado en longitud y latitud para explicar el R^2_{adj} de cada lugar logra explicar el 83 % de su variabilidad. En el Anexo A.B se ha incluido un análisis relativo a si R^2_{adj} del modelo de cada lugar se puede explicar mediante las variables geográficas (Latitud, altitud y longitud).

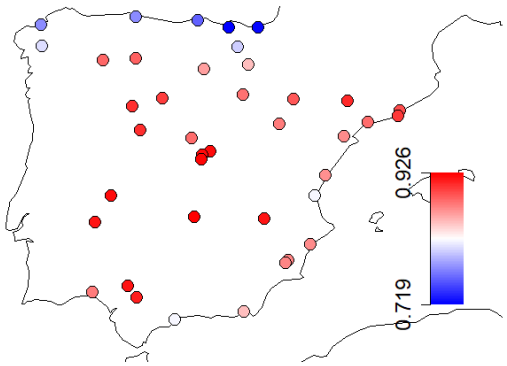


Figura 4.4: Mapa que representa la distribución de R^2_{adj} en los 40 observatorios de la Península.

El efecto del día anterior, ρ_0 es altamente significativo en cada modelo. La Figura 4.5 representa su valor estimado en cada observatorio peninsular. Se deduce que la temperatura máxima del día anterior tiene alto efecto sobre la del día t. En la parte derecha de la Figura 4.5, se representa un diagrama de puntos que enfrenta $\log(altitud)$ con ρ_0. Observamos que ρ_0 es positivo y mayor que 0.482, en todo el mapa peninsular. Los mayores valores se encuentran en el interior de la Península, mientras que en los lugares de la costa el día anterior tiene un menor efecto. El diagrama de puntos muestra cierta relación lineal entre las dos variables.

El efecto autorregresivo con retardo 2, γ_0, es decir, la variable hace dos días, se representa en la Figura 4.6. En este caso observamos que, salvo en 3 lugares, $\gamma_0(s_k)$ es altamente significativo, luego es necesario incluir información relativa a la sitación dos días previos. En este caso, observamos que en el interior de la Península hay un efecto negativo sobre la respuesta, encontrándose los mayores valores (en términos de valor absoluto) en el centro peninsular. También, se aprecia que los valores positivos se concentran en la costa. Dados los valores que proporciona la escala, concluimos que el efecto del día

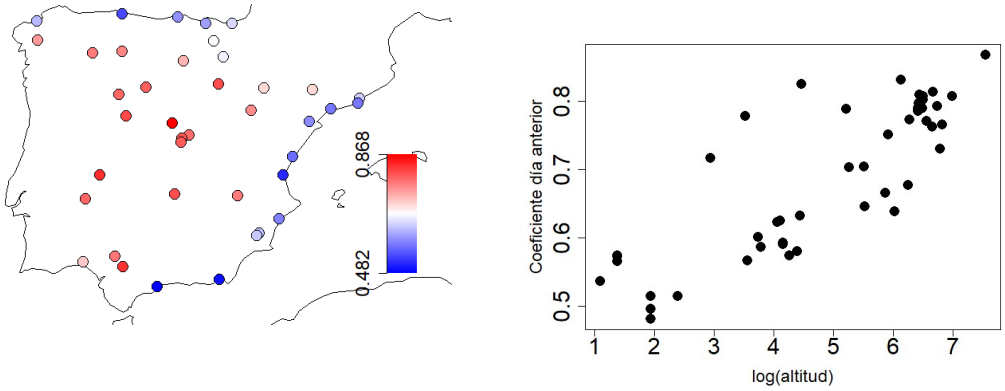


Figura 4.5: Izd.: Mapa que muestra la distribución del valor estimado de ρ_0 en los 40 observatorios. Dch.: ρ_0 estimado frente al logaritmo de la altitud.

anterior es mayor al de hace dos días, en todos los lugares (puesto que tanto la temperatura máxima ayer como la de anteayer se miden en las mismas unidades).

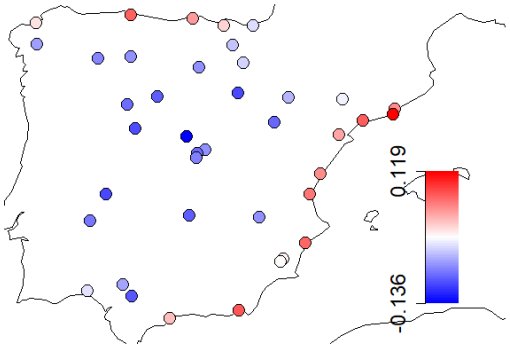


Figura 4.6: Mapa que muestra la distribución del valor estimado de γ_0.

Concluimos analizando el efecto del año en el modelo, α_0, es decir, el vector que contiene en la posición k a $\alpha_0(s_k)$. Este término es significativo en todos los lugares. Representamos en la Figura 4.7 este efecto en un diagrama de caja y un mapa. Todos los valores son positivos, a excepción de uno, lo que se traduce en que al aumentar los años aumenta el efecto de esta covariable sobre la temperatura máxima diaria, algo que no nos sorprende después de los resultados obtenidos en los percentiles móvil 30, ya que vimos que, en los últimos años, las temperaturas habían aumentado. No se aprecia ningún patrón geográfico en este parámetro ligado a longitud o latitud, pero podemos recalcar que en la zona del Ebro se agrupan valores más altos.

Variabilidad espacial de los parámetros estimados en los modelos locales

La variabilidad espacial que muestra cada coeficiente β_i, α_i, ρ_i y γ_i se trata de explicar mediante modelos lineales basados en las covariables geográficas. Además, repetimos el mismo proceso añadiendo las variables climáticas presentadas en la subsección 3.2.1.

También para cada coeficiente calculamos una medida adimensional mediante el cociente entre su media y su desviación típica. Esto nos dará información sobre si alguna covariable es irrelevante en los

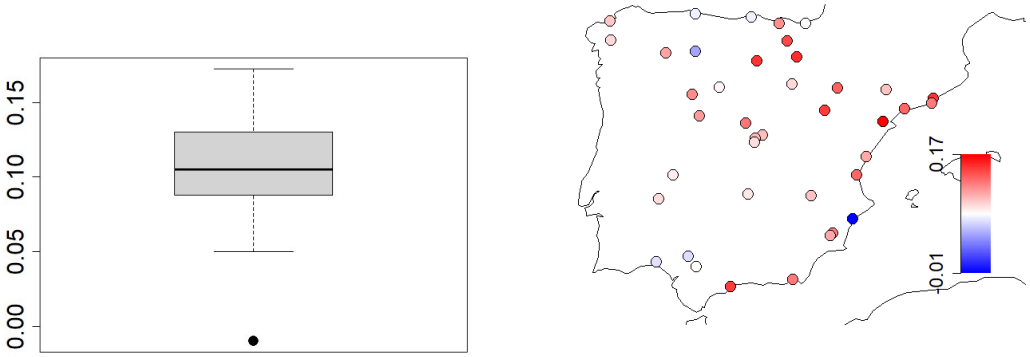


Figura 4.7: Izd.: Diagrama de caja de α_0 estimado en los 40 observatorios. Dch.: Mapa que muestra la distribución de α_0.

modelos, cuando esa medida tome valores próximos a 0. Los resultados se presentan en la Tabla 4.1. La media de los R^2_{adj} asociados a los modelos que explican la variabilidad espacial únicamente mediante las variables geográficas, es de 0.44. Su desviación típica de 0.26. Sus extremos son 0,86 de máximo, perteneciente al modelo del parámetro γ_0, y su mínimo de 0.04, asociado a $\alpha_{2,C}$. Veamos, ahora, el incremento obtenido, en términos de R^2_{adj}, obtenido al considerar variables climáticas en el modelo. El incremento medio es de 0.076 y su desviación típica de 0.089. El incremento máximo se alcanza en $\gamma_{2,S}$, con un valor de 0.30, seguido de un valor de 0.28, asociado a $\beta_{2,S}$. El resto de los incrementos no son tan reseñables. Además existen parámetros que no guardan relación con estas variables climáticas, obteniendo un valor de 0 en su incremento.

Un modelo para explicar ρ_0 con variables geográficas explica el 74 % de la variabilidad espacial. La longitud y la altitud explican el 86.2 % de la variabilidad geográfica de γ_0. Por lo tanto, vemos que la variabilidad espacial de los términos de inercia ρ_0 y γ_0 se explican correctamente con las variables geográficas. Si añadimos las variables climáticas, logramos alcanzar el 89 % con ρ_0, y el 94 % con γ_0. En el caso de α_0 (year), solo llegamos a un 20 %. Respecto a la medida definida en el párrafo anterior , su mayor valor absoluto es de 6.36 y el siguiente de 3.35, que pertenecen al día previo y al año respectivamente. Según esta medida, son los que tienen más variabilidad espacial.

Modelos para la varianza

Se ajusta en cada lugar el modelo para explorar la varianza con el predictor lineal dado por la expresión 3.2. La Figura 4.2.2 muestra la distribución espacial del efecto de la tendencia, α_0 mediante un boxplot y un mapa. A diferencia de lo visto en el modelo exploratorio para el valor medio, este parámetro estimado de la tendencia temporal no es siempre positiva. Es más, la mediana es negativa, es decir, en la mayoría de esos observatorios la distribución de la temperatura diaria máxima es menos variable actualmente que en el inicio de sus series. No observamos ningún patrón geográfico en esta distribución ligado a longitud o latitud. No obstante, vemos un efecto negativo pronunciado en la costa norte, mientras que en la zona del Mediterráneo tiende a ocurrir lo contrario. Si intentamos explicar la variabilidad espacial de este parámetro mediante modelos geográficos se alcanza un R^2_{adj} de 0.34. Esto nos indica que, con los datos geográficos que poseemos, no se explica su variación y que faltaría incluir otras variables relativas a las características climáticas.

De manera general, estos modelos alcanzan un R^2_{adj} reducido. A partir de ellos, se han ensayado modelos alternativos para elevar este valor, mediante la inclusión de interacciones entre más variables del modelo, así como la inclusión de términos polinómicos, pero la mejora en el grado de explicación era pequeña. Esta situación nos indica la necesidad de incluir otras variables atmosféricas que ayuden a establecer las condiciones útiles para explicar la varianza. Las variables que tenemos son capaces de

	Término	Parámetro	$R^2_{adj,geo}$	R^2_{adj}	Medida	Incremento clima
1	Intercepto	β_0	0.2291	0.329	-2.32	0.10
2	Dia anterior	ρ_0	0.7403	0.89	6.33	0.15
3	c1	$\beta_{1,C}$	0.115	0.115	0.56	0.00
4	c2	$\beta_{2,C}$	0.2949	0.2949	-1.40	0.00
5	s1	$\beta_{1,S}$	0.3882	0.3882	0.15	0.00
6	s2	$\beta_{2,S}$	0.592	0.8714	0.82	0.28
7	Dos días atrás	γ_0	0.8624	0.943	-0.23	0.08
8	Coeficiente año	α_0	0.1211	0.2091	3.37	0.09
9	Dia anterior C1	$\rho_{1,C}$	0.7716	0.8928	-0.12	0.12
10	Dia anterior C2	$\rho_{2,C}$	0.7624	0.7609	-0.30	-0.00
11	Dia anterior S1	$\rho_{1,S}$	0.4205	0.4403	0.62	0.02
12	Dia anterior S2	$\rho_{2,S}$	0.6319	0.6319	0.15	0.00
13	Dos dias c1	$\gamma_{1,C}$	0.8342	0.9368	0.82	0.10
14	Dos dias c2	$\gamma_{2,C}$	0.5879	0.6534	0.68	0.07
15	Dos dias s1	$\gamma_{1,S}$	0.4803	0.5814	1.07	0.10
16	Dos dias s2	$\gamma_{2,S}$	0.3839	0.6919	-0.51	0.31
17	Año C1	$\alpha_{1,C}$	0.1127	0.1291	-1.04	0.02
18	Año C2	$\alpha_{2,C}$	0.04236	0.04236	1.30	0.00
19	Año S1	$\alpha_{1,S}$	0.2207	0.2207	-0.77	0.00
20	Año S2	$\alpha_{2,S}$	0.2272	0.3253	-0.48	0.10

Cuadro 4.1: La columna $R^2_{adj,geo}$ corresponde con el R^2_{adj} obtenido en los modelos que únicamente usan las variables geográficas. La columna R^2_{adj} corresponde al perteneciente a los modelos ampliados con variables climáticas. La columna Incremento clima se refiere a la diferencia de R^2_{adj} entre ambos modelos.

explicar la media, no son suficientes por sí solas para la varianza.

4.3. Modelos bayesianos jerárquicos

Se presenta en primer lugar un modelo bamlss en R para la serie de Zaragoza. Se ha estimado la distribución a posteriori de cada coeficiente de cada covariable. Se han obtenido sus medias a posteriori y con ellas se obtiene $\tilde{\mu}_t$ y $\tilde{\sigma}_t$, con los que se dispone de la distribución $P(\tilde{Y}_t|Y) \sim \mathcal{N}(\tilde{\mu}_t, \tilde{\sigma}_t)$. El modelo consta de un predictor lineal con la misma estructura tanto para la media como para la desviación típica, que coincide con la fórmula 3.1. Los resultados se resumen en la Tabla A.C.1 del Anexo A.C. Observamos que, en la estimación de la desviación tipica, todos los coeficientes asociados a las interacciones de la variable Dos días atrás con sus armónicos, presentan un intervalo de credibilidad que contiene al valor 0. Como hemos explicado anteriormente, procedemos a la eliminación de estas covariables de la fórmula de la desviación típica y esto nos lleva a la creación de un nuevo modelo, manteniendo la fórmula de la media. Una vez que tenemos el segundo modelo, lo comparamos con la función DIC. El DIC del primer modelo es de 122251.5, mientras que el del segundo es más pequeño, con un valor de 122246.1, con lo que nos quedamos con el segundo modelo. Atendiendo a los intervalos de credibilidad, sería factible considerar la eliminación de las variables que interaccionan el año con los armónicos, de la parte respectiva a la desviación típica. Se ha ensayado un nuevo modelo, eliminando estas covariables de la fórmula de la desviación típica, pero comparamos su DIC con el modelo anterior y obtenemos que este nuevo modelo es peor que el anterior, luego lo descartamos. No tenemos ninguna evidencia que nos lleve a eliminar más variables, por lo tanto el segundo modelo obtenido es el modelo definitivo.

La media a posteriori y el intervalo de credibilidad para cada parámetro de este modelo seleccionado se muestra en la Tabla 4.2.

La librería *bamlss* permite generar muestras de las distribuciones a posteriori de los parámetros del modelo. Tiene interés para valorar el cambio en la distribución a lo largo del tiempo, la distribución a

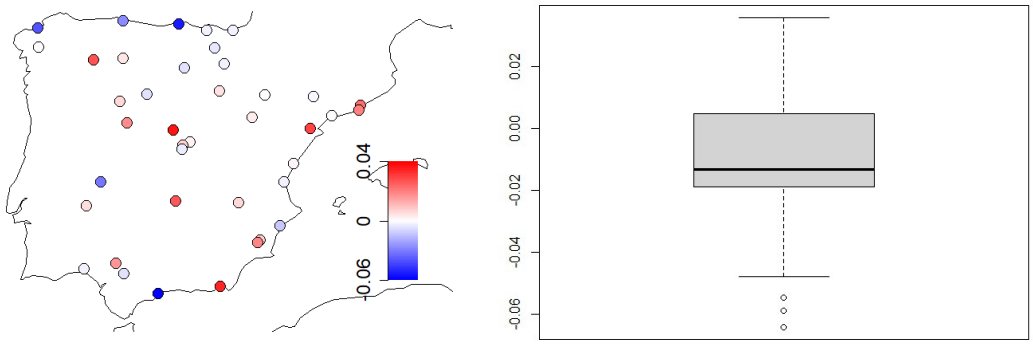


Figura 4.8: Valor estimado de α_0 en el modelo 3.2 en cada observatorio. Izd.: mapa peninsular. Dch.: diagrama de caja.

posteriori del efecto de la variable año. Esta distribución cambia en cada día del año, ya que la interacción del efecto del año con los armónicos provoca diferentes distribuciones a posteriori para la pendiente de esta variable en función del día del año. Para considerar este efecto, se va a representar la tendencia temporal en cuatro fechas: 21 de marzo, 21 de junio, 21 de septiembre y 21 de diciembre. La Figura 4.3 representa sus distribuciones a posteriori con un histograma, utilizando la misma escala. Vemos que, en las 4 fechas, la tendencia es positiva, es decir, cada año aumenta el valor de la media. Además, este crecimiento es mayor en el 21 de junio que en el 21 de septiembre. En ambas fechas, su distribución a posteriori no incluye al valor 0 y tiene mediana a posteriori de 0.019 y 0.015°C/año, respectivamente. Para los días 21 de marzo y de diciembre la distribución a posteriori es similar, con rango en valores positivos, pero con pendientes más reducidas que en las otras fechas; la mediana a posteriori es 0.0092 y 0.0102°C/año, respectivamente.

La Figura 4.10 representa de forma equivalente la distribución a posteriori para la tendencia temporal en el submodelo para la desviación típica. La pendiente es negativa en el 21 de junio. Dada la negatividad de esta distribución, vemos que a lo largo de los años, en el 21 de junio aumenta la media y disminuye la varianza, es decir, tendemos a situaciones más cálidas y con menor variabilidad. Esto podría ser una consecuencia del cambio climático.

El parámetro asociado al efecto de la temperatura máxima en el día anterior es positivo, de acuerdo con su intervalo de credibilidad. Obtenemos una muestra de la distribución a posteriori de este parámetro y se representa mediante un histograma y un gráfico de probabilidad normal en la Figura 4.11. El histograma sugiere que la distribución a posteriori es similar a una normal, aunque la cola inferior parece tener más peso que en una distribución normal. Para tener una evaluación de la normalidad usaremos el test de Shapiro-Wilk, que es una prueba comúnmente utilizada para evaluar la normalidad de una muestra de datos. La hipótesis nula es que los datos siguen una distribución normal. El p-valor es 0.237, claramente mayor que 0.05, por lo que no podemos rechazar la hipótesis nula. Por tanto no hay evidencia suficiente para afirmar que la distribución del parámetro no es normal.

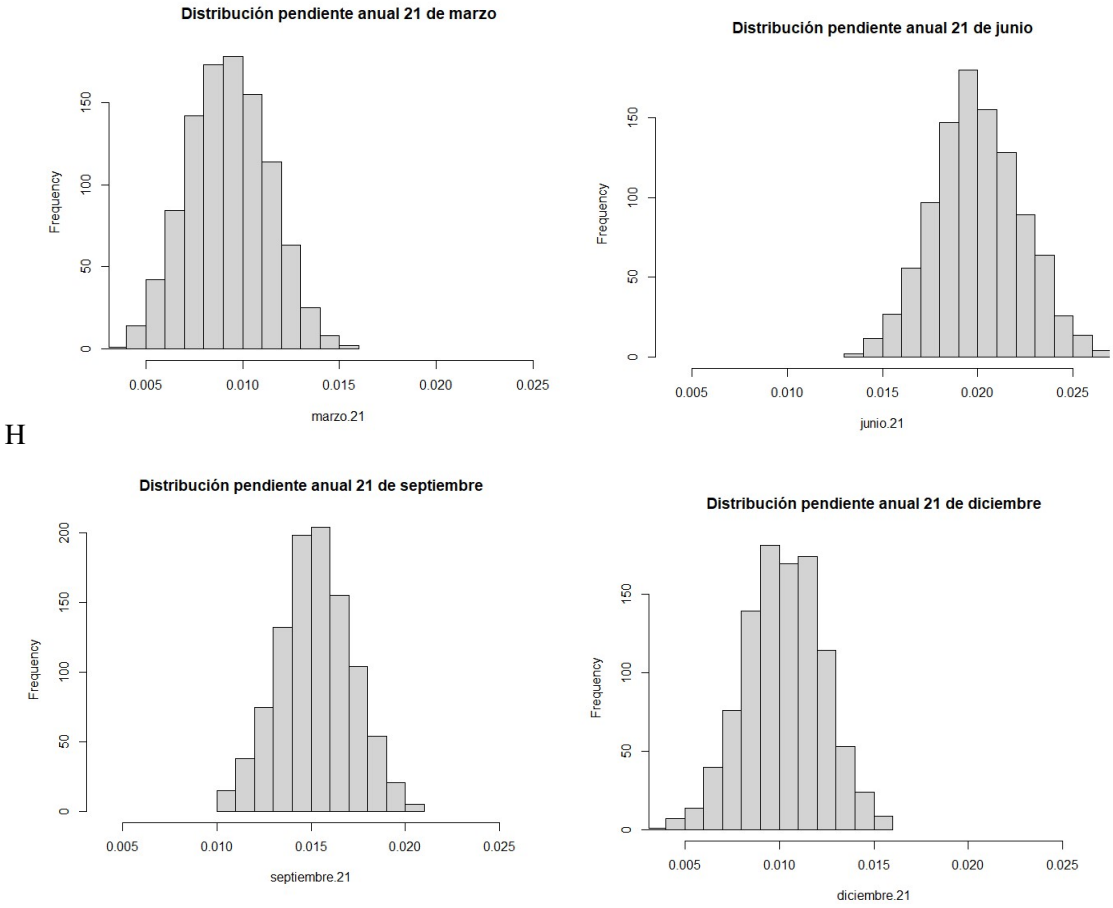
Comprobemos ahora la calidad del modelo. Como hemos dicho con anterioridad, calculamos con R una muestra de $F(Y_t(s)|\mu_t(s),\sigma_t(s))$, y veamos si sigue una distribucion uniforme, con mínimo 0 y máximo 1. Para ello, generamos de nuevo un histograma y un gráfico de cuántiles.

Los resultados muestran que la mayoría de las probabilidades predichas están agrupadas cerca de 0 y 1, con muy pocos datos en los valores intermedios. Esto es un indicativo de que el modelo puede no estar capturando bien la variabilidad. Visto este resultado, y el resultado obtenido en la subsección 4.2.2, podemos concluir que sería sensato, en un modelo futuro, incluir diferentes variables atmosféricas con el fin de explicar la variabilidad, de manera más óptima.

Procedemos ahora a comparar los resultados obtenidos en el modelo local lineal sobre la media, correspondiente con el modelo local bayesiano que acabamos de presentar, con la parte respectiva a la media.

		Predictor lineal para μ		Predictor lineal para σ	
Coeficiente	Parámetro	media	IC 95%	media	IC 95%
Intercepto	β_0	-20.05	[-23.7468,-16.3505]	1.82	[0.9297,2.7291]
Día anterior	ρ_0	0.70	[0.6903,0.7146]	0.00	[0.0017,0.0075]
s1	$\beta_{1,S}$	2.27	[-2.6578,7.4676]	-0.90	[-2.094,0.3412]
s2	$\beta_{2,S}$	2.07	[-2.8897,6.9039]	0.81	[-0.4934,2.046]
c1	$\beta_{1,C}$	6.61	[1.7629,11.826]	-0.68	[-1.9742,0.6479]
c2	$\beta_{2,C}$	-2.78	[-7.7736,2.3065]	1.43	[0.119,2.6536]
Dos días atrás	γ_0	-0.04	[-0.0523,-0.0285]	0.01	[0.0047,0.0106]
Año	α_0	0.01	[0.0118,0.0156]	-0.00	[-0.001,-1e-04]
Día anterior s1	$\rho_{1,S}$	0.02	[0.0037,0.0376]	-0.00	[-0.0039,0.0024]
Día anterior s2	$\rho_{2,S}$	0.02	[0.0044,0.0363]	0.00	[-3e-04,0.0034]
Día anterior c1	$\rho_{1,C}$	-0.03	[-0.0466,-0.0132]	-0.02	[-0.0212,-0.0152]
Día anterior c2	$\rho_{2,C}$	0.01	[-0.0025,0.0297]	-0.01	[-0.0092,-0.0053]
Año s1	$\alpha_{1,S}$	-0.00	[-0.0047,5e-04]	0.00	[-1e-04,0.0011]
Año s2	$\alpha_{2,S}$	-0.00	[-0.003,0.002]	-0.00	[-0.0011,2e-04]
Año c1	$\alpha_{1,C}$	-0.01	[-0.0079,-0.0028]	0.00	[-1e-04,0.0012]
Año c2	$\alpha_{2,C}$	0.00	[-0.0013,0.0038]	-0.00	[-0.0013,0]
Dos días atrás s1	$\gamma_{1,S}$	0.02	[-6e-04,0.0323]		
Dos días atrás s2	$\gamma_{2,S}$	-0.03	[-0.0423,-0.0118]		
Dos días atrás c1	$\gamma_{1,C}$	0.05	[0.0373,0.0694]		
Dos días atrás c2	$\gamma_{2,C}$	-0.01	[-0.0209,0.0105]		

Cuadro 4.2: Representación de los resultados obtenidos con el modelo bamlss final, como la media a posteriori y los intervalos de credibilidad, para μ y σ .



H

Figura 4.9: Distribución a posteriori de la pendiente (ºC/año), para el submodelo del valor medio en las fechas 21 de marzo, 21 de junio, 21 de septiembre y 21 de diciembre, en Zaragoza.

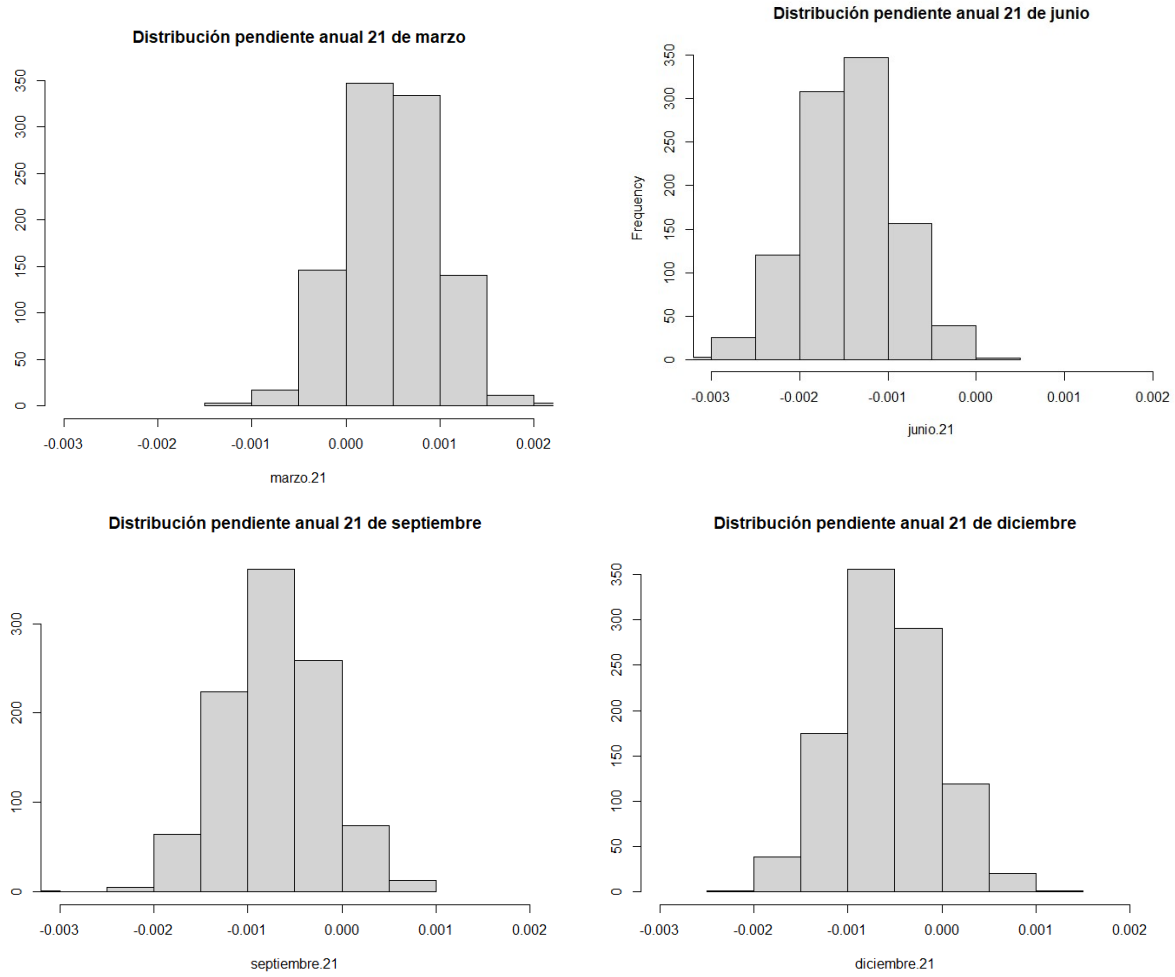


Figura 4.10: Distribución a posteriori de la pendiente (ºC/año), para el submodelo de la desviación tipica en las fechas 21 de marzo, 21 de junio, 21 de septiembre y 21 de diciembre, en Zaragoza.

Además, en la Figura 4.12 se representan los coeficientes estimados por el modelo exploratorio frente al valor medio a posteriori del respectivo parámetro, estimado por el modelo ajustado con bamlss. En la Tabla 4.3, vemos que todos los parámetros estimados del modelo lineal están contenidos en los respectivos intervalos de credibilidad. Esto nos indica que, desde el punto de vista práctico, ambos tipos de modelos son factibles para la media de la variable que queremos explicar. No obstante, la principal ventaja del bayesiano es que nos permite hacer inferencia.

4.3.1. Modelo global con predictores geográficos

Se presentan los resultados modelo bayesiano global referentes a los parámetros de mayor interés. Este modelo considera de forma conjunta la respuesta en los 40 observatorios, tal y como se ha explicado en la subsección 3.3.1. Por ello el proceso de estimación tiene un elevado coste computacional. Su convergencia ha requerido alrededor de 14 horas (52642.58 segundos), en un ordenador con procesador AMD A8-7410 APU de cuatro núcleos con una velocidad base de 2.2 GHz.

La Tabla 4.3.1 muestra los intervalos de credibilidad para los parámetros del modelo. De los 20 parámetros, solo 2 tienen un IC que contiene al 0 en el submodelo de mu, y solo 4 en la sigma, es decir, los términos son necesarios y por lo tanto cualquier modelo más sencillo estará desconsiderando una variante de variación. Además, α_0 es positivo en media y negativo en desviación típica. Esto resulta en un incremento general anual en la media.

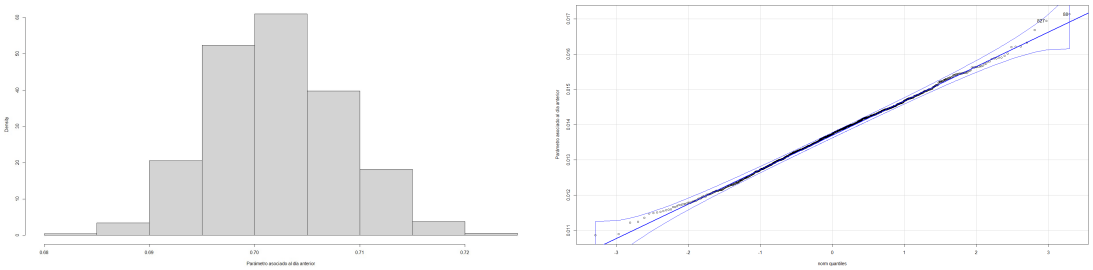


Figura 4.11: Distribución a posteriori de ρ_0 para el submodelo del valor medio, en Zaragoza.

Coeficiente	Parámetro	Exploratorio	media a posteriori	IC 95 %
Intercepto	β_0	-19.9585	-20.0494	[-23.7468,-16.3505]
Día anterior	ρ_0	0.7041	0.7021	[0.6903,0.7146]
s1	$\beta_{1,S}$	2.2998	2.2687	[-2.6578,7.4676]
s2	$\beta_{2,S}$	2.0050	2.0684	[-2.8897,6.9039]
c1	$\beta_{1,C}$	7.6313	6.6055	[1.7629,11.826]
c2	$\beta_{2,C}$	-3.3645	-2.7769	[-7.7736,2.3065]
Dos días atrás	γ_0	-0.0460	-0.0403	[-0.0523,-0.0285]
Año	α_0	0.0137	0.0137	[0.0118,0.0156]
Día anterior s1	$\rho_{1,S}$	0.0241	0.0201	[0.0037,0.0376]
Día anterior s2	$\rho_{2,S}$	0.0167	0.0203	[0.0044,0.0363]
Día anterior c1	$\rho_{1,C}$	-0.0298	-0.0299	[-0.0466,-0.0132]
Día anterior c2	$\rho_{2,C}$	0.0128	0.0140	[-0.0025,0.0297]
Dos días atrás s1	$\gamma_{1,S}$	0.0173	0.0164	[-6e-04,0.0323]
Dos días atrás s2	$\gamma_{2,S}$	-0.0256	-0.0271	[-0.0423,-0.0118]
Dos días atrás c1	$\gamma_{1,C}$	0.0580	0.0535	[0.0373,0.0694]
Dos días atrás c2	$\gamma_{2,C}$	-0.0053	-0.0051	[-0.0209,0.0105]
Año s1	$\alpha_{1,S}$	-0.0021	-0.0020	[-0.0047,5e-04]
Año s2	$\alpha_{2,S}$	-0.0005	-0.0006	[-0.003,0.002]
Año c1	$\alpha_{1,C}$	-0.0059	-0.0053	[-0.0079,-0.0028]
Año c2	$\alpha_{2,C}$	0.0016	0.0013	[-0.0013,0.0038]

Cuadro 4.3: Comparación entre las estimaciones del modelo exploratorio para μ y del modelo bayesiano.

En segundo lugar, se valora si los intervalos de credibilidad de la distribución a posteriori de los parámetros son compatibles con los estimados en el modelo bayesiano local de Zaragoza. Para ello, comparamos la Tabla 4.3.1 que se muestra a continuación, con la Tabla 4.2. Vemos compatibilidad variada entre los intervalos de ambos modelos. Se observa que muchos de estos intervalos tienen intersección vacía, aunque en algunos casos, estos intervalos son colindantes. Destacamos que el intercepto (β_0), tanto para μ como para σ, es compatible en ambos modelos. Para ρ_0, se tiene que no es compatible para el caso μ, pero si lo es para σ. Lo mismo ocurre con γ_0. Los intervalos de credibilidad de α_0 de ambos modelos para μ tienen instersección vacía, pero son colindantes.

El modelo incluye términos suavizados dúctiles sobre la posición geográfica que solo afectan aditivamente, es decir, solo afectan al término independiente β_0 y β_0^{σ}. Estos términos son relevantes pero también será necesario incluir interacciones con la pendiente temporal y con los términos autorregresivos.

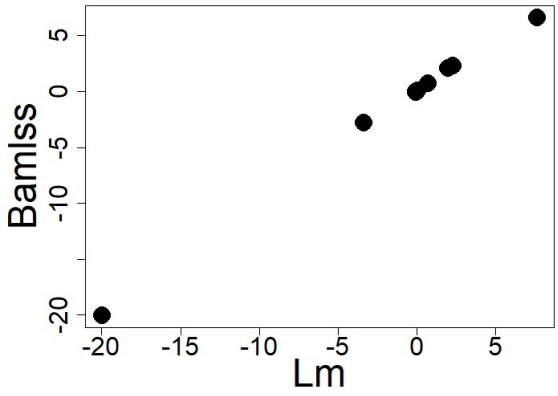


Figura 4.12: Valor medio a posteriori frente a los parámetros estimados por el modelo exploratorio, del valor medio, en Zaragoza .

Coeficiente		IC 95 % μ	IC 95 % σ
Intercepto	β_0	[-23.5851,-22.4368]	[1.2148,1.459]
Día anterior	ρ_0	[0.5065,0.5113]	[0.0041,0.0051]
c1	$\beta_{1,C}$	[8.4287,10.1738]	[-0.1091,0.2181]
c2	$\beta_{2,C}$	[-9.9269,-8.3243]	[-0.7909,-0.445]
s1	$\beta_{1,S}$	[5.9376,7.5341]	[0.1168,0.4636]
s2	$\beta_{2,S}$	[5.5693,7.1383]	[0.8525,1.194]
Dos días atrás	γ_0	[-4e-04,0.0038]	[0.0079,0.0089]
Año	α_0	[0.0157,0.0162]	[-3e-04,-2e-04]
Día anterior c1	$\rho_{1,C}$	[0.0317,0.0385]	[-0.0261,-0.0247]
Día anterior c2	$\rho_{2,C}$	[-0.0452,-0.039]	[-0.0014,-2e-04]
Día anterior s1	$\rho_{1,S}$	[0.0367,0.0429]	[0.0023,0.0037]
Día anterior s2	$\rho_{2,S}$	[-0.0067,-7e-04]	[0.0014,0.0028]
Dos días atrás c1	$\gamma_{1,C}$	[0.0595,0.0664]	[-8e-04,6e-04]
Dos días atrás c2	$\gamma_{2,C}$	[0.0548,0.0609]	[-0.0048,-0.0036]
Dos días atrás s1	$\gamma_{1,S}$	[-0.0021,0.0043]	[-0.0011,3e-04]
Dos días atrás s2	$\gamma_{2,S}$	[0.0087,0.0147]	[-9e-04,4e-04]
Año c1	$\alpha_{1,C}$	[-0.0078,-0.0069]	[2e-04,4e-04]
Año c2	$\alpha_{2,C}$	[0.0042,0.005]	[3e-04,4e-04]
Año s1	$\alpha_{1,S}$	[-0.005,-0.0042]	[-2e-04,0]
Año s2	$\alpha_{2,S}$	[-0.0032,-0.0024]	[-6e-04,-5e-04]

Cuadro 4.4: Intervalos de credibilidad al 95 % para los parámetros del modelo global.

Capítulo 5

Conclusiones

En este trabajo se han explorado conceptos requeridos para el desarrollo de modelos bayesianos autorregresivos, que son útiles para representar series de temperatura máxima diaria. Se han considerado, en particular, modelos inspirados en la modelización espacio-temporal de Castillo-Mateo et al. [1].

En este TFG se han abordado varios retos e innovaciones respecto a los artículos de referencia:

1. Modelización de media y varianza simultáneamente: En este trabajo, se ha abordado la modelización de la media y la varianza de manera simultánea, lo que representa una complicación respecto a los modelos de regresión habituales que tratan estas dos componentes por separado, o incluso consideran constante a la varianza.

2. Análisis durante todo el año: A diferencia de los trabajos de referencia [1] y [2], que se centran solo en el verano, este estudio aborda el análisis de la temperatura durante todo el año. Esto conlleva una mayor complejidad para reflejar una estacionalidad más marcada y tendencias temporales cambiantes a lo largo del año, es decir, interacciones entre armónicos y tendencia temporal.

3. Análisis geográfico amplio: En este estudio, se ha abordado toda la Península, y no solo la Cuenca del Ebro como en el caso de [1] y [2]. Esto implica una mayor variabilidad geográfica y climática, incluyendo comportamientos distintos en la costa y el interior. Además, aquí, entrarían en juego variables atmosféricas que, a una escala local, como la de la Cuenca del Ebro, pueden considerarse sin variabilidad espacial, pero que a gran escala geográfica cobren mayor importancia. La cercanía a la costa puede ser necesaria para explicar la variabilidad espacial-regional a nivel peninsular, así puede ser informativa la distancia al mar, ya que la costa afecta directamente al clima.

Respecto a los resultados obtenidos, merece la pena resaltar los siguientes:

1. Se ha demostrado la gran dependencia, en toda la Península entre la temperatura máxima del día actual y la del previo. Además, la variación espacial de esta correlación se puede explicar en gran parte mediante la altitud, con mayor relación entre estas dos variables en el interior de la Península y algo menor en las zonas de la costa.

2. La temperatura máxima registrada dos días atrás tiene efecto significativo en la mayoría de los lugares. De nuevo, muestra un mayor nivel de correlación en el interior peninsular. Este resultado, y el anterior, refuerzan la necesidad de una estructura autorregresiva.

3. Se ha identificado la existencia de épocas del año donde la evolución en el largo plazo presenta incremento sobre la media y descenso sobre la varianza, es decir, se tiende a temperaturas máximas más elevadas y menos variables, siendo un claro indicativo de cambio climático.

4. Las variables temporales y climáticas que se han utilizado explican la variabilidad geográfica de la varianza en menor medida que lo que se ha identificado para el valor medio.

5. La media en la temperatura máxima diaria se explica satisfactoriamente con el modelo propuesto. Además, la geografía está relacionada con la calidad del modelo, como se ha mostrado en la Figura 3.1, que representa la distribución de R^2_{adj} sobre la Península. No obstante, sigue habiendo un cierto margen de mejora en algunos lugares, especialmente en la Costa Cantábrica.

Hemos llegado a la conclusión de que el modelo espacio-temporal es adecuado para representar la variabilidad de la temperatura diaria. Existen opciones de mejora respecto a la representación de la variabilidad espacial de los coeficientes de la expresión 3.1, sería conveniente introducir otras variables con el fin de explicar la distribución geográfica de algunos parámetros (como α_0). Para mejorar el submodelo de la desviación típica, sería útil obtener información atmosférico-temporal, esas variables podrían ser el viento, la presión atmosférica o la humedad. Respecto a la variabilidad espacial, para algunos parámetros no es suficiente solo con las variables geográficas y climáticas que poseemos. Una idea intuitiva sería incluir la distancia al mar, dado que la costa tiende a suavizar las temperaturas. También sería factible introducir el nivel de contaminación local, e incluso, indagar más en la geografía local, debido a que la presencia de valles, montañas, ríos, lagos, etc., es capaz de afectar a la temperatura.

Bibliografía

[1] J. CASTILLO-MATEO, M. LAFUENTE, J. ASÍN, A. C. CEBRIÁN, A. E. GELFAND, J. ABAURREA (2022). Spatial modeling of day-within-year temperature time series: an examination of daily maximum temperatures in Aragón, Spain. *Journal of Agricultural, Biological and Environmental Statistics*, 27, 487–505. `https://doi.org/10.1007/s13253-022-00493-3`.

[2] J. CASTILLO-MATEO, J. ASÍN, A. C. CEBRIÁN, A. E. GELFAND, J. ABAURREA (2023). Spatial quantile autoregression for season within year daily maximum temperature data. *Annals of Applied Statistics*. `https://doi.org/10.1214/22-AOAS1719`.

[3] A. GELMAN, J. CARLIN, H. STERN, D. DUNSON, A. VEHTARI, D. RUBIN (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC.

[4] G. GROLEMUND, H. WICKHAM (2011). Dates and Times Made Easy with **'lubridate'**. *Journal of Statistical Software*, 40(3), 1–25.

[5] D. PEÑA-ANGULO, J.C. GONZALEZ-HIDALGO, L. SANDONÍS, S. BEGUERÍA, M. TOMAS-BURGUERA, J.A. LÓPEZ-BUSTINS, M.LEMUS-CANOVAS, J. MARTIN-VIDE (2021). Seasonal temperature trends on the Spanish mainland: A secular study (1916–2015). *International Journal of Climatology*, 41(5), 3071–3084.

[6] H. RIEBL (2022). **'lmls'**: Gaussian Location-Scale Regression. *R package version 0.1.0*. `https://CRAN.R-project.org/package=lmls`.

[7] D. M. STASINOPOULOS, R. A. RIGBY (2008). Generalized additive models for location scale and shape (GAMLSS). *Journal of Statistical Software*, 23, 1-46.

[8] K. TANK, A.M.G. WIJNGAARD AND OTHERS (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology*, 22, 1441–1453.

[9] N. UMLAUF, N. KLEIN, A. ZEILEIS (2018). BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *Journal of Computational and Graphical Statistics*, 27(3), 612-627.

[10] N. UMLAUF, N. KLEIN, T. SIMON, A. ZEILEIS (2021). **'bamlss'**: A Lego Toolbox for Flexible Bayesian Regression (and beyond). *Journal of Statistical Software*, 100(4), 1–53.

[11] H. WICKHAM (2016). **'ggplot2'**: Elegant Graphics for Data Analysis. *Springer-Verlag New York*.

[12] A. ZEILEIS, G. GROTHENDIECK (2005). **'zoo'**: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software*, 14(6), 1–27.