Modelos bayesianos para series temporales climáticas

Javier Torcal Villadangos

Trabajo de fin de grado de Matemáticas
Universidad de Zaragoza

Directores del trabajo: Jesús Asín Lafuente
y Jorge Castillo Mateo
26 de junio de 2022

Summary

The main objective of this work is to develop statistical models and fit the series of local daily maximum temperature (T.máx.) in three Spanish observatories. The interest lies not only in characterizing the effect of global warming in a daily scale on the mean temperature but also in reflecting that effect on its variability. The methodology employed is based on the work by Castillo-Mateo et al. [2] who suggested using the Bayesian framework, and more specifically hierarchical models, to make inferences about the distribution of T.máx. in Aragón.

Regarding the characteristics that our model must include, we highlight:

- Seasonality to capture the effect of variation in solar radiation depending on the day within year. This fact can affect the mean, the variance and even the serial correlation.
- Serial correlation to explain the dependence of the temperature with respect to the previous day, that is, the persistence. Its effect on the mean and variance will change influenced by seasonality.
- A linear trend to represent climate warming. It will be shown that it is not uniform throughout the year in both the mean and the variance.

The proposed model must allow to explain the mean and the variance of the data. Both must depend on covariates that express seasonality through harmonics, the short-term trend through an autoregressive term, and the long-term change through a linear trend. In particular, the T.máx. at day t will be denoted as Y_t for t varying in $1,\ldots,25500$. The predictors included in our model will be harmonic terms, sine and cosine functions, denoted as $S1_t = \text{sen}(2\pi t/365)$, $C1_t = \cos(2\pi t/365)$, $S2_t = \text{sen}(4\pi t/365)$, and $C2_t = \cos(4\pi t/365)$; an autoregressive term (AR) shown as Y_{t-1}; and a linear trend denoted $Year_t$. We will employ the notation $\mathbf{X}_t = (Year_t, Y_{t-1}, S1_t, C1_t, S2_t, C2_t)$ for the vector gathering all the covariates that model the mean value and the variance in the manner we can observe in the following equations:

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

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

Firstly, we start with an exploratory data analysis in order to decide which terms we should include in the model for the mean and the variance. At the beginning we must filter our series because they may contain not available values. After this step we proceed to extract all the amount of knowledge about our subject that we can discuss about them. Linear models will play an important role in the exploratory task. They will help us to identify the effects of the different potential predictors on the mean and the variance. We will seek a 'final' model for the mean which includes all the profitable terms for posterior inference. Moreover, we will take advantage of the residuals of this model so that we can use its squares to estimate the predictors for the variance.

According to Castillo-Mateo et al., we have decided to set our framework in the Bayesian paradigm due to the fact that it will allow us a more flexible fitting and inference. We also make use of R packages

that provide Bayesian inference such as **'jagsUI'** or **'bamlss'**. We finally decided to employ the last one because of the great range of functions that it provides and its capability to be user-friendly.

We focus in three Spanish observatories located in Soria, Zaragoza, and Bilbao. The chosen period is from 1951 to 2020. The data series are provided by the European database ECA&D [9]. We have estimated an average increase in the T.máx. in Soria of 1.57ºC, in Zaragoza of 2.07ºC and in Bilbao of 1.58ºC, the same day of the year but after a period of 60 years.

Nevertheless, there are several differences between this work and the one of Castillo-Mateo et al. They considered a spatio-temporal model including terms which vary because of the location such us the elevation and meanwhile we will restrict us just to time terms. Another difference we can observe is that we will take into consideration all the year long whereas they have focused in the warm days of summer period. Due to this fact, we will identify more variability among our data and we will consider the necessity of introducing interactions among the autoregressive term and seasonality and between the trend and seasonality both to explain better their effect during the year. Another feature to add is that we will care about modeling the heteroscedasticity of T.máx.

To summarize, Chapter 1 suggests us an introduction with the aim to aid the reader to understand the baseline and the main goals of the project. It will show the incentives that have motivate us to conduct this study and it will present a starting point of the model equations.

Furthermore Chapter 2 will present us an accurate revision of Bayesian basis that will lead us to the Bayesian approach and in addition it will present us general knowledge about hierarchical models, time series, autoregressive models and so on, with the incorporation of some easily comprehensible examples to illustrate this concepts. This overview is essential in the sense of setting a solid base that it will encourage us to operate in a Bayesian framework or within an autoregressive model.

In addition, Chapter 3 will establish the procedure to carry out in order to conduct, initially the exploratory analysis and latter the Bayesian inference. This chapter will describe the proposed methodology in order to ensure we will fulfill our objectives stated beforehand. At the end of this chapter it will be presented a brief description of the R packages chosen to make Bayesian inference.

Then Chapter 4 gathers the main results obtained by applying to our time series the method exposed in the previous chapter following a parallel work with the paper of Castillo-Mateo et al. We thoroughly try to express the conclusions that might arise from the auxiliary plots and tables supported by a closing chapter containing the supplementary material. Needless to say we will make great efforts on showing to such extent is essential to include the interaction effects or not considering a constant variance, therefore gaining in importance figures summarizing their respective posterior distributions.

Finally, in Chapter 5 we discuss about the achieved results in Chapter 4 and we put them into comparison with those of the pioneer works. Besides we would like to suggest forward guidelines to include interactions among model terms and to model the variance.

Índice general

Capítulo 1

Introducción

El incremento de las temperaturas es conocido como uno de los efectos más evidentes del cambio climático observado desde mitad del s. XX a nivel global y también en la Península Ibérica, Peña et al. [6]. Este cambio se refleja en el valor medio y puede aparecer en otros aspectos de la distribución de la temperatura diaria como la varianza. Es conveniente trabajar con la serie diaria y a escala local, escalas que se sitúan más próximas a aquellas donde se producen los principales efectos sobre la sociedad y la naturaleza.

El objetivo y alcance del trabajo final de grado (TFG) es plantear una modelización estadística para series diarias de temperatura máxima que sea capaz de dar una estimación del efecto del calentamiento en la posición central y en la variabilidad de su distribución. La metodología sigue la línea de Castillo-Mateo et al. [2], que plantean un modelo jerárquico bayesiano a partir del cual estiman el cambio en la distribución de la temperatura máxima diaria en series de la Comunidad Autónoma de Aragón.

Por tanto, como objetivo general del TFG, se plantea la construcción de modelos estadísticos para la temperatura máxima diaria en un observatorio, en el día t denotada Y_t, capaces de verificar las siguientes cualidades que explican gran parte de la variabilidad natural:

- La correlación serial, que refleja la inercia de la atmósfera a corto plazo.
- La estacionalidad, que muestra el efecto de la radiación solar variable según la época del año. Esta característica provoca que a lo largo del año cambie el valor medio, la variabilidad y, posiblemente, la correlación serial de la temperatura máxima diaria.
- El modelo estadístico debe incluir términos con los que se analice el cambio observado, que puede ser no homogéneo a lo largo del año, tanto en el valor medio como en la variabilidad.

Los modelos que se requieren extienden a los que se han estudiado en la asignatura de Regresión Lineal, donde se ha visto una introducción a los modelos lineales para una respuesta continua, haciendo depender el valor medio de la respuesta de un vector de covariables y considerando un error aleatorio con distribución gaussiana y varianza constante. También se estudian transformaciones de tipo Box-Cox y modelos con pesos para evitar el problema de la heterocedasticidad, pero no se han estudiado modelos para modelar la varianza.

El modelo permitirá modelar conjuntamente el valor esperado y la varianza, que deben depender de variables que expresan la estacionalidad, mediante términos seno y coseno de los primeros armónicos $S1_t = \text{sen}(2\pi t/365)$, $C1_t = \cos(2\pi t/365)$, $S2_t = \text{sen}(4\pi t/365)$, y $C2_t = \cos(4\pi t/365)$; la dependencia a corto plazo, mediante un término autorregresivo Y_{t-1}; y el cambio a largo plazo, mediante una tendencia temporal.

Denotaremos $\mathbf{X}_t = (Year_t, Y_{t-1}, S1_t, C1_t, S2_t, C2_t)$ al vector que contiene los términos considerados en la modelización del valor esperado y la varianza, como aparece en la siguiente expresión:

$$E[Y_t|\mathbf{X}_t] = \beta_0 + \sum_{i=1}^{2}(\beta_{i,s}Si_t + \beta_{i,c}Ci_t) + \left(\alpha_0 + \sum_{i=1}^{2}(\alpha_{i,s}Si_t + \alpha_{i,c}Ci_t)\right)Year_t + \left(\rho_0 + \sum_{i=1}^{2}(\rho_{i,s}Si_t + \rho_{i,c}Ci_t)\right)Y_{t-1}. \tag{1.1}$$

$$Var[Y_t|\mathbf{X}_t] = \exp\left\{\beta_0^\sigma + \sum_{i=1}^{2}(\beta_{i,s}^\sigma Si_t + \beta_{i,c}^\sigma Ci_t) + \left(\alpha_0^\sigma + \sum_{i=1}^{2}(\alpha_{i,s}^\sigma Si_t + \alpha_{i,c}^\sigma Ci_t)\right) Year_t + \left(\rho_0^\sigma + \sum_{i=1}^{2}(\rho_{i,s}^\sigma Si_t + \rho_{i,c}^\sigma Ci_t)\right) Y_{t-1}\right\}. \tag{1.2}$$

De forma que $t = 1, 2, \ldots, 25500$ son los días sucesivos considerados.

En este esfuerzo de modelización será primordial el paradigma bayesiano ya que permitirá un ajuste más versátil. Para ello utilizamos el software R que nos ofrece múltiples posibilidades con sus diferentes librerías.

Nuestro objeto de aplicación particular será la serie de temperatura máxima diaria registrada en observatorios situados en Soria, Zaragoza y Bilbao en el periodo 1951-2020, cuyos datos obtendremos de la base de datos europea ECA&D [9].

Una fase del análisis se centrará en la depuración de las series descargadas de esa base de datos ya que hay lagunas esporádicas en los registros. Después realizaremos un análisis exploratorio, lo que permitirá extraer información sobre qué variables predictoras podremos incluir en los modelos, ya que nuestro propósito es modelar la distribución con la intención de analizar aspectos del cambio observado.

En resumen, este TFG aborda una metodología en el marco de estimación bayesiano que extiende los modelos lineales revisados en el Grado de Matemáticas. En el siguiente capítulo se van a revisar los conceptos de inferencia bayesiana, modelos jerárquicos, series temporales y modelos autorregresivos, así como el trabajo de Castillo-Mateo et al. [2]. El Capítulo 3 presentará la metodología que se ha utilizado en el análisis exploratorio y en la estimación de los modelos. El Capítulo 4 muestra la aplicación sobre las series de temperatura máxima de observatorios españoles, utilizando los resultados obtenidos para analizar el cambio observado. Por último, se incluye un capítulo que resume las conclusiones del trabajo.

Capítulo 2

Fundamentos teóricos

Este capítulo revisa los conceptos básicos de la inferencia bayesiana e introduce un ejemplo de modelo bayesiano con distribución normal. En las siguientes secciones se plantea una breve introducción a los modelos jerárquicos, series temporales y modelos autorregresivoss. 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. Inferencia bayesiana

La inferencia bayesiana establece como paradigma que los parámetros de interés en un modelo corresponden a una distribución de probabilidad que puede estimarse dada la información conocida, es decir, la muestra de la respuesta observada. Esta distribución se denomina a posteriori y se obtiene a partir del concepto que establece la regla de Bayes. En esta sección se describen los elementos básicos de la inferencia bayesiana. De acuerdo con Gelman et al. [3], los tres pasos en el análisis de datos bayesiano se resumen en lo siguiente:

- Primero se establece el modelo de probabilidad completo, esto es, una distribución de probabilidad conjunta para todos los valores observables y no observables en el problema.
- A continuación, se condiciona en los datos observados. De esta forma se obtiene la *distribución a posteriori*, es decir, la distribución de probabilidad de los valores no observados dados los datos observados.
- Finalmente, se realiza la validación del modelo para evaluar su ajuste con las herramientas de diagnóstico que ofrece la estadística bayesiana.

Una vez el modelo se considere adecuado, el siguiente paso consiste en explotarlo para realizar predicciones sobre los datos no observados o para describir la distribución de nuestros parámetros mediante herramientas gráficas o numéricas.

El objetivo final de la inferencia bayesiana será obtener la distribución a posteriori marginal de nuestros parámetros de interés. Se distinguen dos tipos de magnitudes a estimar, cantidades observables sobre las que se hará inferencia y cantidades potencialmente no observables. La forma de obtenerlo será primero calculando la distribución a posteriori conjunta de todos los parámetros y después integrando sobre los que resulten irrelevantes para obtener las distribuciones marginales deseadas.

Denotaremos θ al vector de cantidades no observables o parámetros de interés, y a los datos observados de los que disponemos e $\tilde{y}$ a las cantidades no observadas pero potencialmente calculables.

2.1.1. La regla de Bayes

La estimación de la distribución a posteriori $P(\theta|y)$ requiere un modelo que proporcione la distribución conjunta $P(\theta, y)$. En un caso discreto se trabajará con la función de masa de probabilidad conjunta

y en el caso continuo con la función de densidad conjunta. En cualquier caso se podrá escribir como producto de la *distribución a priori de los parámetros* $P(\theta)$ y la distribución de los datos dado θ, es decir la *función de verosimilitud*:

$$P(y,\theta) = P(\theta)P(y|\theta).$$

Usando la *regla de Bayes* sobre la densidad a posteriori de los parámetros se obtiene:

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

donde $P(y) = \sum_\theta P(\theta)P(y|\theta)$ es la suma sobre todos los valores posibles de θ o su expresión análoga $\int P(\theta)P(y|\theta)d\theta$ en el caso continuo.

Cuando se plantea un proceso de estimación sobre $P(\theta|y)$, una forma análoga a la expresión (2.1) consiste en la omisión de $P(y)$, factor que no depende de θ, luego que puede ser considerado constante en los siguientes pasos de la estimación. Esto nos conducirá a considerar la densidad a posteriori no normalizada:

$$P(\theta|y) \propto P(\theta)P(y|\theta). \tag{2.2}$$

Esto implica que es equivalente maximizar $P(\theta|y)$ a maximizar la expresión de (2.2).

Estas simples fórmulas resumen el núcleo técnico de la inferencia bayesiana.

La forma en que los datos y afectan a la inferencia a posteriori dado un modelo bayesiano es a través del término $P(y|\theta)$ en la regla de Bayes (2.2). Este término considerado como función de θ fijado y es la *función de verosimilitud.* De esta manera, se dice que la inferencia bayesiana sigue el principio de la verosimilitud.

2.1.2. Tipos de distribución a priori

Existen dos interpretaciones básicas en la elección de la distribución a priori. La interpretación poblacional, donde dicha distribución representa la población de los posibles valores de los parámetros para los cuales θ debe ser simulado.

La otra interpretación es subjetiva y expresa el nivel de conocimiento del problema. Debemos expresar nuestro conocimiento sobre θ pero también nuestra incertidumbre. Según esta interpretación distinguimos los tipos de distribuciones a priori: informativa cuando expresa nuestro conocimiento sobre los datos, débilmente informativa cuando juega el mínimo papel posible y no informativa o plana.

Habitualmente, la distribución a priori no necesita estar concentrada en torno al valor real ya que a menudo, si los datos son lo suficientemente informativos su aportación de la muestra sobre θ sobrepasará considerablemente a la de cualquier a priori razonable.

Distribuciones no informativas y débilmente informativas

Cuando no disponemos de información sobre los datos queremos que la distribución a priori tenga el mínimo efecto posible en la distribución a posteriori. Este tipo de distribuciones no informativas a veces se conoce como distribuciones a priori de referencia.

Un concepto relacionado son las distribuciones débilmente informativas, las cuales contienen algo de información para regularizar la distribución a posteriori dentro de unos límites razonables, incluyendo la mínima información posible.

El *principio de razón insuficiente* dictamina que la especificación uniforme de la distribución a priori es apropiada si no se conoce nada sobre θ. En la búsqueda de una distribución a priori no informativa surgen diversos problemas. El primero es que no existe una distribución a priori que siempre sea poco informativa en todas las situaciones. Además, en muchos problemas no hay una elección clara de distribución a priori no informativa ya que una distribución plana o uniforme en una parametrización puede no serlo en otra.

En el uso de distribuciones a priori no informativas proporcionales a una constante, puede ocurrir que la distribución sea *impropia*, es decir que su integral sea infinito. En otro caso se dice que es *propia*.

Sin embargo, existen casos de distribuciones a priori impropias que pueden llevar a distribuciones a posteriori propias.

Una propiedad en la búsqueda de distribuciones a priori no informativas es el *principio de invarianza de Jeffreys*, basado en la consideración de transformaciones uno a uno del parámetro $\phi = h(\theta)$. Este principio puede ser extendido a modelos multi-paramétricos. Este principio expresa que cualquier regla para determinar la densidad a priori $P(\theta)$ debe proporcionar un resultado equivalente si es aplicado a un parámetro transformado. Sea $\phi = h(\theta)$:

$$P(\phi) = P(\theta)\left|\frac{d\theta}{d\phi}\right| = P(\theta)|h'(\theta)|^{-1}. \tag{2.3}$$

Entonces, calcular $P(\phi)$ determinando primero $P(\theta)$ y aplicando (2.3) debe ser igual a obtenerlo directamente usando el modelo transformado: $P(y,\phi) = P(\phi)P(y|\phi)$.

Este principio sugiere definir la densidad a priori no informativa $P(\theta) \propto [J(\theta)]^{1/2}$, siendo $J(\theta)$ la información de Fisher de θ, se tiene:

$$J(\theta) = E\left(\left(\frac{d\log P(y|\theta)}{d\theta}\right)^2 \middle| \theta\right).$$

Para comprobar que dicha a priori es invariante a la parametrización, se evalúa $J(\phi)$ en $\theta = h^{-1}(\phi)$:

$$J(\phi) = -E\left(\frac{d^2\log P(y|\phi)}{d\phi^2}\right) = -E\left(\frac{d^2\log P(y|\theta = h^{-1}(\phi))}{d\theta^2}\left|\frac{d\theta}{d\phi}\right|^2\right) = J(\theta)\left|\frac{d\theta}{d\phi}\right|^2,$$

por tanto se llega a $J(\phi)^{1/2} = J(\theta)^{1/2}\left|\frac{d\theta}{d\phi}\right|$ como se buscaba.

Distribuciones a priori conjugadas

Cuando se verifica la propiedad de que la distribución a posteriori siga la misma forma paramétrica que la distribución a priori se dice que la distribución a priori es conjugada. La definimos formalmente:

Definición. Sea $\mathscr{F}$ una clase de distribuciones de muestreo $P(y|\theta)$ y sea $\mathscr{P}$ una clase de distribuciones a priori para θ, entonces la clase $\mathscr{P}$ es conjugada con $\mathscr{F}$ si $P(\theta|y) \in \mathscr{P}$ para todo $P(*|\theta) \in \mathscr{F}$ y $P(*) \in \mathscr{P}$.

Esta definición es amplia ya que si tomamos $\mathscr{P}$ la clase de todas las distribuciones entonces esta sería conjugada con cualquier distribución de muestreo que se use. Sin embargo, estamos interesados en las familias a priori conjugadas naturales que surgen de tomar todas las funciones de densidad que tienen la misma forma paramétrica. Un ejemplo habitual es que la distribución a priori beta es una familia conjugada con la verosimilitud binomial. Otro ejemplo es con la distribución a priori normal conjugada con una verosimilitud normal.

El uso de una familia conjugada es matemáticamente conveniente ya que entonces conoceremos la forma paramétrica que sigue la distribución a posteriori. Sin embargo, la selección de una distribución a priori para nuestro modelo basada exclusivamente en que sea conjugada con la verosimilitud de nuestros datos puede ser no adecuada, ya que podríamos necesitar una familia más realista. Los modelos conjugados suelen ser un buen punto de partida. Cuando este tipo de modelos no sea razonable, las mixturas de familias conjugadas pueden ser útiles.

Otra ventaja adicional a la computacional es que las distribuciones conjugadas pueden ser interpretadas como datos adicionales.

Como un caso usual, veamos que las distribuciones de la familia exponencial son distribuciones a priori conjugadas naturales. Sean y_i los datos, θ el vector de parámetros y $\mathscr{F}$ la clase de la familia exponencial, sus miembros son de la forma:

$$P(y_i|\theta) = f(y_i)g(\theta)exp\left(\phi(\theta)^T u(y_i)\right),$$

donde $\phi(\theta)$ llamado parámetro natural de $\mathscr{F}$ y $u(y_i)$ son vectores de igual dimensión que θ. Denotando $t(y) = \sum_{i=1}^{n} u(y_i)$, la verosimilitud correspondiente a la muestra de observaciones independientes e idénticamente distribuidas $y = (y_1, \ldots, y_n)$ es:

$$P(y|\theta) = \left(\prod_{i=1}^{n} f(y_i) \right) g(\theta)^n exp\left(\phi(\theta)^T t(y)\right).$$

Luego será proporcional como función de θ a:

$$P(y|\theta) \propto g(\theta)^n exp\left(\phi(\theta)^T t(y)\right).$$

Notar que $t(y)$ es estadístico suficiente para θ ya que la verosimilitud de θ depende de los datos y solo a través de $t(y)$.

Si la densidad a priori se especifica:

$$P(\theta) \propto g(\theta)^{\chi} exp\left(\phi(\theta)^T \nu\right).$$

Entonces la densidad a posteriori es del mismo tipo:

$$P(\theta|y) \propto P(\theta)P(y|\theta) \propto g(\theta)^{n+\chi} exp\left(\phi(\theta)^T (\nu + t(y))\right).$$

Con lo cual concluimos que dicha familia es conjugada.

2.1.3. La influencia de la distribución a priori en la distribución a posteriori

Como hemos visto, el proceso de estimación bayesiano consiste en pasar de la distribución a priori $P(\theta)$ a la distribución a posteriori $P(\theta|y)$, luego es lógico que surjan relaciones entre dichas distribuciones.

Consideramos la ecuación de la esperanza condicional, $E(\theta) = E(E(\theta|y))$, expresa que la media a priori es la media de todas las posibles medias a posteriori sobre la distribución de los datos.

La distribución a posteriori incorpora la información proveniente de los datos, por lo que es menos variable que la distribución a priori, como se encuentra en la fórmula de la varianza condicional:

$$Var(\theta) = E(Var(\theta|y)) + Var(E(\theta|y)).$$

Por lo tanto, la varianza a posteriori en media $E(Var(\theta|y))$ es más pequeña que la varianza a priori $Var(\theta)$, con una diferencia que depende de la distribución de los datos, $Var(E(\theta|y))$. Cuanto mayor sea la varianza de la media a posteriori, mayor será el potencial de reducir nuestra incertidumbre respecto a θ.

2.1.4. Inferencia a partir de la distribución a posteriori

La distribución a posteriori contiene toda la información disponible sobre el parámetro θ. Una ventaja clave del paradigma bayesiano es la flexibilidad con la cual se puede resumir la inferencia a posteriori.

- Una manera de resumir la distribución es mediante estimaciones puntuales. Medidas comunes para la localización son la media, la mediana o la moda. La variabilidad se suele resumir con medidas como la desviación típica, el rango intercuartílico o mediante cuantiles.

- Además de los resúmenes puntuales, es importante representar la incertidumbre a posteriori. Para ello, una herramienta útil son los intervalos de credibilidad centrales. En estadística bayesiana, los intervalos de probabilidad de los parámetros se dicen *intervalos de credibilidad* y consiguen una alta probabilidad de contener al parámetro desconocido. Si estos corresponden al $100(1-\alpha)\%$, el intervalo deja a su derecha e izquierda una probabilidad $\alpha/2$ a posteriori. Estos intervalos son fácilmente computables mediante muestreos con el ordenador.

- Otra herramienta de resumen es la *región de mayor densidad de probabilidad* a posteriori, un conjunto de valores que incluye una probabilidad a posteriori $1-\alpha$. Coincidirá con el intervalo de credibilidad central cuando la distribución a posteriori sea simétrica. Su característica principal es que la densidad de probabilidad en esta región nunca es menor que fuera de ella. En la práctica se usa más el intervalo central porque es calculado directamente computacionalmente y debido a las interpretaciones de los cuantiles $\alpha/2$ y $1-\alpha/2$.

Predicción en la inferencia bayesiana

Se pueden realizar predicciones sobre las cantidades observables desconocidas, es decir sobre valores futuros de la respuesta o en general no registrados, lo que se denomina inferencia predictiva. La distribución del dato observable es:

$$P(y) = \int P(y,\theta)d\theta = \int P(\theta)P(y|\theta)d\theta.$$

Se llama distribución marginal de y o distribución a priori predictiva. Se dice a priori ya que no es condicionada a un proceso previo observado y predictiva debido a la observabilidad del dato.

Ahora veremos cómo inferir el valor de un nuevo dato desconocido $\tilde{y}$. La distribución de $\tilde{y}$ se llama distribución predictiva a posteriori. Esta vez se dice a posteriori ya que se condiciona en la muestra observada.

$$P(\tilde{y}|y) = \int P(\tilde{y},\theta|y)d\theta = \int P(\tilde{y}|\theta,y)P(\theta|y)d\theta = \int P(\tilde{y}|\theta)P(\theta|y)d\theta.$$

El último paso se basa en la independencia condicional de y e $\tilde{y}$ dado θ, la cual asumimos.

Algoritmos MCMC

A la hora de inferir distribuciones a posteriori surgen integrales multidimensionales cuya solución analítica es muy difícil de obtener. Es por ello que los métodos de Monte Carlo basados en cadenas de Markov (MCMC) resultan una herramienta fundamental para la inferencia bayesiana. Los métodos MCMC son un tipo de algoritmo utilizado para el muestreo de una distribución de probabilidad y se basan en la construcción de cadenas de Markov que tienen como distribución de equilibrio la buscada. Dado un valor inicial para los parámetros, se simulan valores sucesivamente de una densidad propuesta 'sencilla', que no tiene que ser necesariamente parecida a la densidad a posteriori, generalmente compleja y de gran dimensionalidad. Cada valor generado depende solo del anterior valor simulado, de ahí la noción de cadena de Markov.

Los modelos bayesianos en este trabajo son ajustados mediante métodos de simulación MCMC. Es decir, dado que la expresión explícita de la distribución a posteriori de los parámetros es desconocida, se generarán muestras de la distribución a posteriori mediante algoritmos MCMC para estimar cantidades de interés de la misma.

2.2. Modelo bayesiano con distribución normal

En esta sección se muestra un ejemplo de estimación en el marco bayesiano. Se va a revisar el modelo multi-paramétrico normal con media y varianza desconocidas. En la práctica, muchos modelos plantean el uso de aproximaciones normales donde la media y la varianza son los parámetros a estimar. Puede ser de interés considerar posibles transformaciones de sus parámetros, como en las distribuciones normales donde el inverso de la varianza tiene un papel destacado y se conoce como *precisión*.

Veamos un ejemplo donde los datos $y = (y_1,\ldots,y_n)$ siguen una distribución normal y se elige una distribución a priori no informativa. Asumiendo la independencia a priori de los parámetros de escala y forma, una distribución a priori poco informativa para μ y σ es una distribución uniforme en μ y en el parámetro transformado $\log(\sigma)$:

$$P(\mu,\sigma^2) \propto (\sigma^2)^{-1}.$$

Bajo dicha distribución a priori impropia, la distribución a posteriori conjunta es proporcional a la función de verosimilitud de la muestra multiplicada por el factor $(\sigma^2)^{-1}$:

$$P(\mu,\sigma^2|y) \propto P(\mu,\sigma^2)P(y|\mu,\sigma^2) \propto \sigma^{-n-2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$$
$$= \sigma^{-n-2}\exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2+n(\bar{y}-\mu)^2\right]\right), \tag{2.4}$$

donde $s^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1)$ es la cuasivarianza muestral e $\bar{y}$ la media muestral.

La distribución a posteriori marginal $P(\sigma^2|y)$ se obtiene integrando la distribución conjunta (2.4) sobre μ:

$$P(\sigma^2|y) \propto \int \sigma^{-n-2}\exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2+n(\bar{y}-\mu)^2\right]\right)d\mu$$
$$\propto \sigma^{-n-2}\exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right)\int \exp\left(-\frac{1}{2\sigma^2}n(\bar{y}-\mu)^2\right)d\mu$$
$$\propto \sigma^{-(n+1)/2}\exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right).$$

Notar que $\sigma^2|y$ sigue una distribución inversa-Gamma (IG):

$$\sigma^2|y \ \sim \ IG\left(\frac{n-1}{2},\frac{(n-1)s^2}{2}\right). \tag{2.5}$$

La distribución a posteriori condicional de μ conocida la varianza es normal:

$$\mu|\sigma^2,y \ \sim \ N(\bar{y},\sigma^2/n). \tag{2.6}$$

Entonces se factoriza la distribución a posteriori conjunta como: $P(\mu,\sigma^2|y) = P(\mu|\sigma^2,y)P(\sigma^2|y)$. Por tanto es fácil simular muestras de la distribución conjunta, primero simulando muestras de σ de (2.5) y después de μ de (2.6).

Ahora buscamos la forma analítica para la distribución a posteriori marginal de μ integrando la distribución a posteriori conjunta en este caso sobre σ^2:

$$P(\mu|y) = \int_0^\infty P(\mu,\sigma^2|y)d\sigma^2,$$

usando el cambio de variable $z = [(n-1)s^2+n(\mu-\bar{y})^2]/(2\sigma^2)$ resulta una integral gamma:

$$P(\mu|y) \propto \left[(n-1)s^2+n(\mu-\bar{y})^2\right]^{-n/2}\int_0^\infty z^{(n-2)/2}\exp(-z)dz \propto \left[1+\frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{-n/2}.$$

Luego bajo una distribución a priori no informativa uniforme para μ y $\log(\sigma)$, la distribución marginal a posteriori de μ es t de Student:

$$\left.\frac{\mu-\bar{y}}{s/\sqrt{n}}\right|y \ \sim \ t_{n-1}.$$

La distribución a posteriori predictiva para una observación futura $\tilde{y}$ se obtiene:

$$P(\tilde{y}|y) = \int\int P(\tilde{y}|\mu,\sigma^2,y)P(\mu,\sigma^2|y)d\mu d\sigma^2.$$

El primer factor es el valor de la observación futura en la distribución normal dados los valores de μ y σ^2 luego no depende de y. Para simular la distribución predictiva a posteriori, primero se simulan μ,σ^2 de su distribución conjunta a posteriori y posteriormente se simula $\tilde{y}|\mu,\sigma^2 \sim N(\mu,\sigma^2)$.

2.3. Modelos bayesianos jerárquicos

Los modelos jerárquicos, también llamados modelos multinivel, son usados cuando la información está organizada en diferentes niveles observacionales. Este tipo de modelos permite introducir dependencia a priori entre los parámetros. Esto se puede conseguir si usamos una distribución a priori en la cual los parámetros θ_j se consideran como una muestra de una distribución poblacional, es decir, se consideran un efecto aleatorio. Los parámetros de esta distribución poblacional se llaman *hiper-parámetros*.

En la práctica los modelos no jerárquicos con pocos parámetros no pueden ajustar grandes conjuntos de datos adecuadamente. En cambio, con un gran número de parámetros se tiende a un sobreajuste, el modelo explica bien nuestros datos pero en el caso de predicciones para nuevas observaciones tienden a ser deficientes.

Los modelos jerárquicos permiten el uso de suficientes parámetros para ajustar los datos mientras que evitan el sobreajuste introduciendo la distribución poblacional para estructurar la dependencia entre parámetros. Por ejemplo, consideramos un conjunto de experimentos $j = 1, \ldots, J$ con vector de datos y_j y vector de parámetros θ_j. Si no conocemos nada sobre los parámetros θ_j que pueda distinguirlos, estos pueden considerarse intercambiables, es decir, su distribución conjunta $P(\theta_1, \ldots, \theta_J)$ es invariante frente a permutaciones en sus índices y podemos considerarla dependiente de un hiper-parámetro ϕ.

$$P(\theta_1, \ldots, \theta_J | y_1, \ldots, y_J) = \int \prod_{j=1}^{J} P(\theta_j | \phi, y_j) P(\phi | y) d\phi.$$

Inferencia en modelos jerárquicos

No conocemos el valor de los hiper-parámetros, denotados por ϕ, luego debemos fijar una distribución a priori $P(\phi)$. De esta forma se incluye nuestra incertidumbre sobre ellos. La distribución a posteriori bayesiana adecuada es la del vector (ϕ, θ). La distribución a priori es:

$$P(\phi, \theta) = P(\phi) P(\theta | \phi),$$

y la distribución a posteriori conjunta correspondiente es:

$$P(\phi, \theta | y) \propto P(\phi, \theta) P(y | \phi, \theta) = P(\phi, \theta) P(y | \theta).$$

Hay que notar que la última igualdad se debe a que la distribución de los datos $P(y|\phi, \theta)$ depende directamente solo de θ ya que los hiper-parámetros afectan a y solo a través de θ.

Para crear una distribución de probabilidad conjunta para (ϕ, θ), debemos asignar una distribución a priori a ϕ. En caso de tener escaso conocimiento acerca de ϕ es aconsejable asignarle una distribución poco informativa, teniendo cuidado cuando sea impropia de que la distribución a posteriori resulte propia. Es habitual empezar con distribuciones relativamente no informativas y añadir información en la a priori si hay demasiada varianza a posteriori.

En cuanto a la distribución a posteriori predictiva, dos tipos pueden resultar de interés, la distribución de una futura observación $\tilde{y}$ correspondiente a un θ_j previamente considerado o la distribución de una futura observación $\tilde{y}$ correspondiente a un parámetro futuro $\tilde{\theta}$ simulado de la misma superpoblación. Sin embargo, hay que tener algunas precauciones en la utilización de este tipo de modelos. Un problema que puede surgir del elevado número de parámetros es la identificación de alguno de ellos o incluso en el caso de conjuntos de datos pequeños que el número de datos no pueda soportarlos.

Para comprobar la fiabilidad del modelo es recomendable realizar un análisis de sensibilidad empleando datos usados en la estimación o una validación usando datos externos a la muestra considerada. Hay medidas que pueden resultar útiles como el p-valor predictivo a posteriori, medidas de incertidumbre en la validación cruzada como MAE, CRPS o DIC. Otras herramientas de interés son los métodos de selección de variables.

Ejemplo de modelo normal con parámetros aleatorios

Se plantea a continuación un ejemplo de modelo jerárquico normal simple debido a su gran aplicabilidad, ver Figura 2.1. Se consideran diferentes experimentos aleatorios $i = 1,\ldots,n$, cada uno con m observaciones y_{ij}, $j = 1,\ldots,m$, que se distribuyen normalmente con medias θ_i particulares a cada experimento y varianza σ^2 constante.

$$y_{ij}|\theta_i,\sigma^2 \ \sim \ N(\theta_i,\sigma^2).$$

La media de cada experimento θ_i se simula a partir de una superpoblación distribuida normalmente con hiper-parámetros (μ,τ^2) para la localización y la escala, respectivamente. El hiper-parámetro para la localización μ también se distribuye normalmente con media y varianza (μ_0,σ_0^2) prefijadas:

$$\theta_i|\mu,\tau^2 \ \sim \ N(\mu,\tau^2), \quad \mu \ \sim \ N(\mu_0,\sigma_0^2).$$

Para la distribución a priori del parámetro de escala es usual proponer la distribución IG $(\varepsilon,\varepsilon)$ como distribución poco informativa alternativa a la uniforme. Esta distribución tiene la ventaja de ser condicionalmente conjugada dados los demás parámetros, es decir, la distribución condicional a posteriori $P(\tau^2|\theta,\mu,y)$ es también inversa-Gamma. Para conseguir que sea poco informativa, debemos fijar los parámetros $\varepsilon \leq 2$. En este ejemplo se toman ε_τ y ε_σ valores prefijados para:

$$\tau^2 \ \sim \ IG(\varepsilon_\tau,\varepsilon_\tau), \quad \sigma^2 \ \sim \ IG(\varepsilon_\sigma,\varepsilon_\sigma).$$

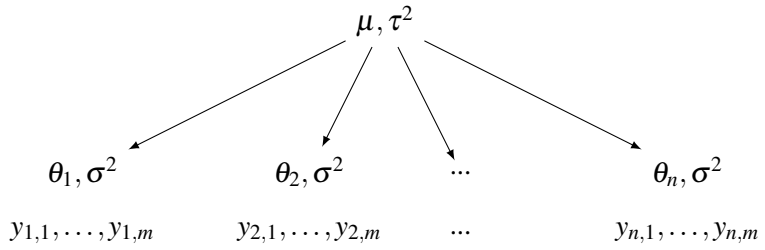

Figura 2.1: Esquema del modelo jerárquico

2.4. Series temporales y modelos autorregresivos

En esta sección se mostrarán los conceptos básicos acerca de las series temporales y de los términos autorregresivos de acuerdo con Brockwell y Davis [1].

Definición. Una *serie temporal* es un proceso estocástico compuesto de variables $\{Y_t,\ t = 0, \pm 1, \dots\}$ cuyos índices se mueven en el tiempo dentro de intervalos regulares.

Lo más habitual es considerar una escala temporal discreta, por ejemplo la escala diaria.

Las observaciones medidas a lo largo del tiempo suelen presentar una distribución que puede evolucionar. Además, estas observaciones poseen una estructura de *correlación serial*, es decir, el valor de una observación depende de las anteriores. Es importante recoger la correlación serial en el modelo ya que aporta información para nuestra predicción y no considerarla puede afectar a las propiedades en las estimaciones.

Es habitual modelizar las series temporales mediante una distribución normal multivariante. En este caso, la distribución de la serie queda caracterizada por los momentos de primer y segundo orden.

Definición. Dada $\{Y_t\}$ una serie temporal, que verifica $E(Y_t^2) < \infty$, llamamos:

1. *Función de medias* de $\{Y_t\}$ a $\mu_Y(t) = E[Y_t]$.
2. *Función de covarianzas* de $\{Y_t\}$ a $\gamma_Y(r,s) = Cov[Y_r, Y_s]$ para r, s enteros cualesquiera.
3. *Función de correlación* de $\{Y_t\}$ a $\rho_Y(r,s) = Cor[Y_r, Y_s]$ para r, s enteros cualesquiera.

Un tipo importante de serie temporal son las series *estacionarias*, las cuales tienen propiedades estadísticas similares a las series desplazadas $\{Y_{t+h},\ t = 0, \pm 1, \dots\}$ para cada entero h.

Definición. Se dice que una serie temporal $\{Y_t\}$ es *débilmente estacionaria* si cumple dos condiciones:

1. La función de medias $\mu_Y(t)$ es independiente de t, es decir, constante.
2. La función de covarianzas $\gamma_Y(h) = Cov[Y_t, Y_{t+h}]$ no depende de t, para cada retardo h.

Para este tipo de series se utiliza la notación $\gamma_Y(h) := \gamma_Y(h,0) = \gamma_Y(t+h,t)$ para la función de covarianzas debido a su independencia de t.

Definición. Dada una serie débilmente estacionaria $\{Y_t\}$ se llama *función de autocovarianza (ACF)* a $\gamma_Y(\cdot)$ y $\gamma_Y(h)$ es su valor con retardo h.

Una aproximación común para modelar series de datos correlados son las estructuras autorregresivas. Un *proceso autorregresivo (AR)* es un modelo de regresión que explica la variable Y_t en un instante t en función de su valor en los instantes anteriores $Y_{t-1}, Y_{t-2}, \dots$. El proceso AR(1) expresa dependencia solo del instante previo, se describe con la variable $Z_t = Y_t - \phi Y_{t-1}$, donde $Z_t \sim\ WN(0, \sigma^2)$, distribución normal con términos independientes, y se satisface la condición $|\phi| < 1$, necesaria para que el proceso tenga solución estacionaria.

Como se ha indicado en la sección de Introducción, en el manejo de las series temporales diarias de temperatura se incluirá la dependencia temporal mediante la inclusión como covariable del valor en el día previo, es decir, se incluye un término autorregresivo. Una alternativa es considerar una dependencia temporal estocástica expresada por un modelo que recoge la dependencia entre los residuos, por ejemplo un modelo jerárquico. También se puede plantear una dependencia de segundo orden (respecto a los dos días previos).

Además, estas estructuras autorregresivas se pueden incluir en modelos más generales, junto con otras variables que pueden variar con el tiempo, llamados *modelos dinámicos*. Un ejemplo es el utilizado en las expresiones (1.1) y (1.2) para explicar la temperatura máxima diaria en una localidad, donde para cada día t aparecen como covariables términos armónicos que ciclan en el año Si_t o Ci_t, el año $Year_t$ y el valor de la variable respuesta en el día previo Y_{t-1}.

2.5. Modelización de series para temperatura diaria

A la hora de proponer un modelo para series de temperatura se ha tomado en consideración trabajos previos sobre la materia. En específico, se ha tomado como referencia el trabajo de Castillo-Mateo et al. [2] sobre modelado espacio-temporal de temperaturas máximas diarias en los meses centrales del año (de mayo a septiembre), en el periodo 1956-2015, en la Comunidad Autónoma de Aragón.

Castillo-Mateo et al. proponen un modelo jerárquico bayesiano, de tipo espacio-temporal, con dos niveles para explicar la temperatura máxima diaria. Este modelo busca caracterizar patrones espaciales y detectar tendencias temporales. Se adoptan dos unidades temporales discretas, los años y los días dentro del año y un tratamiento continuo de la componente espacial. En particular, persigue explicar la persistencia temporal mediante la inclusión de un término autorregresivo a través de los años y otro para los días dentro de cada año. El modelo considera efectos fijos y aleatorios. Los efectos fijos describen la media global, la componente estacional sobre los días del año, la tendencia temporal a largo plazo o la influencia de la elevación. Los efectos aleatorios se usan para representar la dependencia espacial en los interceptos, los coeficientes de pendiente, los coeficientes de autorregresión y la varianza de la respuesta. Los parámetros del modelo son estimados mediante el uso de algoritmos MCMC. Finalmente, se emplea la distribución a posteriori predictiva para interpolar, es decir, para obtener muestras de las temperaturas en localizaciones no observadas dentro de la región.

Sea $Y_{t,l}(s)$ la variable que denota la temperatura máxima diaria para el día $l = 2, \ldots, L$ correspondiente al año $t = 1, \ldots, T$ en una localización $s \in D$ siendo D la región de estudio. El modelo separa los efectos aleatorios de los fijos en la media y también separa espacio, día y año en los diferentes efectos:

$$Y_{t,l}(s) = \mu_{t,l}(s) + \gamma_t(s) + \rho_y(s)(Y_{t,l-1}(s) - (\mu_{t,l-1}(s) + \gamma_t(s))) + \varepsilon_{t,l}(s).$$

Notar que se denota por $\mu_{t,l}(s)$ a la componente de efectos fijos y por $\gamma_t(s)$ a la componente de efectos aleatorios.

$$\mu_{t,l}(s) = \beta_0 + \alpha t + \beta_1 \sin(2\pi l/365) + \beta_2 \cos(2\pi l/365) + \beta_3 \text{elev}(s).$$

En esta fórmula se denota β_0 al intercepto global, α al coeficiente de la tendencia lineal global y los términos de seno y coseno se introducen para la estacionalidad. Por otra parte, elev(s) es la elevación en el lugar s.

Ahora nos centramos en $\gamma_t(s)$, cuya composición está descrita así:

$$\gamma_t(s) = \beta_0(s) + \alpha(s)t + \psi_t + \eta_t(s).$$

El término $\psi_t = \rho_\psi \psi_{t-1} + \lambda_t$ es de tipo AR(1), y proporciona autorregresión para los interceptos anuales. Los términos $\beta_0(s)$ y $\alpha(s)$ son el intercepto y el coeficiente de la tendencia lineal global respectivamente. Dichos parámetros espacialmente variables se modelan mediante procesos gaussianos o transformaciones de los mismos. No se incide en ello ya que se aleja del marco del trabajo.

Notar que se han introducido tres términos de error: $\lambda_t \sim N(0, \sigma_\lambda^2)$ en la escala anual, $\eta_t(s) \sim N(0, \sigma_\eta^2)$ para las ubicaciones dentro de los años y $\varepsilon_{t,l}^{(Y)}(s) \sim N(0, \sigma_\varepsilon^2(s))$ para las ubicaciones dentro de los días del año. Además, $\rho_Y(s)$ es un término autorregresivo que varía espacialmente y $\sigma_\varepsilon^2(s)$ es una varianza que cambia espacialmente.

Capítulo 3

Metodología propuesta para construir el modelo

En este capítulo se presentan las herramientas gráficas y numéricas utilizadas para explorar las fuentes de variación que afectan a la respuesta, así como la forma de la relación funcional. Para este objetivo, se usan como instrumentos exploratorios algunos modelos lineales. En segundo lugar, se presenta la estrategia de modelización en el ámbito bayesiano. La última sección se refiere al software utilizado.

3.1. Herramientas exploratorias

Inicialmente se realiza un trabajo de análisis exploratorio sobre nuestro conjunto de datos para determinar qué variables pueden explicar las series de temperaturas. En primer lugar, es primordial estudiar la calidad de los datos ya que las series pueden contener lagunas (datos perdidos) o estar en un formato que no deseamos. Para el manejo de las fechas se propone utilizar la librería **'lubridate'** [4] y se programará una función para facilitar el filtrado de los datos.

Entre las herramientas utilizadas para describir el comportamiento del valor medio se encuentran algunos elementos descriptivos usuales pero también se utilizan resultados de modelos lineales para explorar el efecto de potenciales variables predictoras. Se destacan las siguientes:

- Gráfico de temperatura máxima media anual frente al año correspondiente para mostrar la tendencia en el largo plazo, se incluye un suavizado loess.
- Gráfico de media móvil de temperatura máxima diaria para mostrar la estacionalidad y otros resultados del perfil estacional obtenidos mediante el ajuste de modelos lineales con términos armónicos hasta el orden 2.

  Dado el carácter estacional de la variable respuesta y con el objetivo de mostrar una evolución suave de los resúmenes, algunos gráficos propuestos emplean ventanas móviles de 15 o 30 días para suavizar los cambios entre los días consecutivos, gracias al uso de la librería **'zoo'** [12].
- Gráfico ACF para determinar la posible dependencia entre los residuos de un modelo para mostrar la correlación de los datos.
- Gráficos mostrando los coeficientes de los términos del modelo que cuantifican las interacciones entre el término AR(1) y la tendencia lineal con la estacionalidad, frente al día del año.

Además de ayudar a la elaboración de gráficos exploratorios por medio de sus residuos o valores ajustados, los modelos lineales darán una medida aproximada de la aportación de cada término a la variabilidad total. El uso de resúmenes numéricos como el valor del coeficiente que acompaña a cada término resultan de gran utilidad para ello. Dichos modelos lineales básicos se basan en suposiciones como que el error sigue una distribución gaussiana de varianza constante o la linealidad e independencia

en los datos. Por lo tanto, cuando se pretende hacer inferencia a partir de estos modelos es importante realizar test de hipótesis o gráficos para analizar si los residuos resultantes en los ajustes de los modelos satisfacen dichas condiciones. Aquí se propone obtener modelos lineales solo como una herramienta exploratoria adicional para promover los términos a considerar en el modelo bayesiano, por ello no se lleva a cabo una crítica de cada modelo lineal obtenido en esta fase del estudio.

Se construyen diferentes modelos lineales para comprobar la importancia de la inclusión de los diferentes términos así como la existencia de las interacciones entre algunos de ellos. En este proceso se analizará si existe la necesidad de incluir términos de tipo autorregresivo AR(1). Los modelos lineales poseen medidas como el grado de ajuste R^2 o el R^2_{adj} que dan una idea sobre la variabilidad explicada por el modelo sobre la variabilidad total de los datos.

Por último, se probará el modelo completo incluyendo todos los términos predictores y sus interacciones que se han mostrado relevantes en el análisis exploratorio. Uno de los principales problemas en los modelos de regresión lineal múltiple es la colinealidad entre las diferentes covariables lo que puede resultar en problemas para nuestra estimación. Por todo ello es importante realizar procedimientos de selección de variables como el procedimiento paso a paso de tipo 'backward', que consiste en empezar tomando todas las variables e ir eliminando las menos relevantes hasta optimizar medidas basadas en el log-verosimilitud como el AIC o BIC.

Debido a la homocedasticidad (varianza constante) requerida por los modelos lineales, ante una varianza cambiante no podemos modelar su variabilidad apropiadamente con lo propuesto hasta ahora. Esto conlleva que aunque sean modelos adecuados para representar el valor medio, en los ajustes finales nos preocupemos de modelar también la varianza.

Para explorar la dependencia de la varianza en función de las variables predictoras, se realizará un estudio utilizando como nueva respuesta el cuadrado de los residuos del modelo lineal final para la media. Los gráficos que emplearemos para estudiar el comportamiento de la varianza son:

- Gráfico de la nueva respuesta frente al año con suavizado loess.
- Gráfico de la nueva respuesta respecto el día del año, considerando una ventana móvil para su suavizado.
- Gráfico de la nueva respuesta frente a su valor en el día previo, incluyendo un suavizado loess.

Una vez comprobada la influencia de las diferentes covariables sobre la varianza se procede a ajustar un modelo lineal donde la variable respuesta es el cuadrado de los residuos procedentes del modelo lineal final para la media. Aplicando un procedimiento paso a paso se seleccionan sus términos relevantes dando lugar a una propuesta de modelado para la varianza en los modelos bayesianos.

3.2. Ajuste del modelo bayesiano

3.2.1. Estrategia de modelización

Una vez concluimos la fase exploratoria continuamos el ajuste de modelos dentro del paradigma bayesiano. Vamos a ajustar varios, primero modelizando tanto la media como la varianza pero sin término autorregresivo, después modelos basados en el mejor modelo lineal propuesto con varianza constante y, por último, el modelo completo con media y varianza dependientes de las covariables e incluyendo término de tipo AR(1). El objetivo es demostrar la mejora del ajuste obtenido al incluir el término autorregresivo y al modelizar la varianza, en un grado que no se puede obviar. El criterio con el que mostraremos la bondad de ajuste del modelo es el DIC, el mejor modelo será el que lo minimice.

DIC

El Criterio de Información de la Desviación o DIC es una versión generalizada del AIC y BIC que se utiliza con la estimación MCMC. El DIC mide la bondad de ajuste a partir de la desviación definida como $D(\theta) = -2\log(P(y|\theta))$. Se considera el promedio de la desviación:

$$\bar{D}(\theta) = -2\int \log(P(y|\theta))d\theta.$$

Para penalizar la complejidad del modelo, el DIC calcula el número efectivo de parámetros denotado pD, como $pD = \bar{D}(\theta) - D(\bar{\theta})$ y el DIC se define:

$$DIC = \bar{D}(\theta) + pD = D(\bar{\theta}) + 2pD$$

El número efectivo de parámetros pD tiene en cuenta simultáneamente el tamaño muestral, la covarianza de los parámetros, el número de parámetros y el tamaño de los efectos de las variables.

Convergencia del muestreo MCMC

Es importante asegurarse de la convergencia de las cadenas MCMC en el ajuste del modelo propuesto. Para ello se emplean los siguientes métodos con el apoyo de la librería **'coda'** [7].

- Herramientas visuales como *trace plot* o gráfico de trazos que muestra el trazo de las iteraciones de las diferentes cadenas MCMC. Es adecuado considerar varias cadenas que comiencen en puntos diferentes y ver si convergen a una misma distribución.
- Medidas estadísticas como el *factor de reducción potencial de escala* $\hat{R}$ que da una estimación de la convergencia midiendo la varianza entre las cadenas y la varianza dentro de cada cadena. En el caso ideal el valor de $\hat{R}$ para cada parámetro debe estar cerca de 1 y al menos ser menor que 1,1. En caso de no converger se debería aumentar el número de iteraciones.
- *Tamaño de muestra efectivo* de cada parámetro es el número efectivo de sorteos en la simulación. Es deseable que sea similar al número de sorteos a posteriori solicitados.

Inferencia sobre los parámetros y explotación del modelo

Se estudia la distribución a posteriori de cada parámetro para valorar si el cero se sitúa próximo a su masa de probabilidad. Los intervalos de credibilidad y la media de las distribuciones a posteriori marginales son las herramientas usuales. Cuando el intervalo de credibilidad considerado contiene al cero se considera factible que el efecto correspondiente no tenga influencia sobre la respuesta.

También serán ilustrativos los gráficos de densidad de las distribuciones a posteriori de los parámetros. Se hará énfasis en comparar las distribuciones marginales a posteriori de los parámetros de mayor interés como la tendencia lineal o la dependencia del día previo y se estudiará el efecto de las interacciones entre variables explicativas.

Una de las utilidades del modelo es la de obtener inferencia sobre resúmenes de la variable respuesta, que no podrían hacerse solo a partir de los datos. En particular, el modelo se va a explotar para estimar aspectos del calentamiento referidos a días concretos del año. Como las muestras de datos contienen las temperaturas de diferentes décadas utilizamos el modelo final para discutir el efecto del cambio climático en la temperatura máxima en media. Mediante la simulación de nuevos datos con el modelo en dos décadas distantes hacemos la diferencia de temperaturas en cada día del año y calculamos el intervalo de credibilidad para ver si hay evidencia de un cambio climático y en tal caso en qué nivel se ha producido o si es homogéneo en todas las estaciones del año.

3.2.2. Software disponible

No menos importante que la justificación teórica utilizada es la elección del software estadístico a utilizarse para la inferencia de los parámetros del modelo. Encontramos una diversidad amplia de paquetes en R que proporcionen estimaciones en el paradigma bayesiano. En el primer acercamiento a la modelización bayesiana se utilizó el paquete **'jagsUI'** [5] y finalmente se ha decidido el uso de **'bamlss'** [10] para la presentación de resultados. Se describen brevemente algunas características de ambas librerías.

‘jagsUI’

Su principal ventaja es que requiere explicitar en el ‘script’ todos los elementos que definen el modelo así como parámetros, verosimilitud, distribución a priori. Este método de trabajo constructivo favorece la comprensión de los modelos bayesianos desde la aportación de distribuciones a priori a cada término hasta la forma de estructurar el modelo. Su desventaja es una mayor dificultad en su uso y una mayor carga computacional. Además para la modelización de la varianza deberíamos expresar su verosimilitud resultando más trabajoso.

En las primeras fases de este proyecto se trabajó con esta librería, desarrollando ‘scripts’ con la intención de estimar modelos para el valor medio.

‘bamlss’

Su objetivo general es proporcionar una infraestructura para estimar modelos de regresión dentro del esquema bayesiano donde los parámetros pueden capturar la localización, escala y forma, como se indica en Umlauf et al. [10]. Una de las principales ventajas respecto a otras librerías es que su uso resulta intuitivo debido a su estructura de “lego". Este paquete permite un alto grado de flexibilidad debido a la elevada capacidad de elección de funciones en cada etapa de la construcción del modelo como se detallará a continuación, ver Figura 3.1.

La descripción del modelo se realiza mediante la introducción de una fórmula que incluye las variables predictoras a obtener de los datos, del conjunto de datos y de la familia que contiene información sobre la distribución de la variable respuesta.

En cuanto a las herramientas de estimación, primeramente permite incluir una función optimizadora entre varias disponibles que calcule una estimación de la moda a posteriori. Después de este elemento opcional, **‘bamlss’** incluye diversas funciones de muestreo para una inferencia bayesiana completa utilizando algoritmos MCMC que aprovechan la estimación del optimizador como valores iniciales.

A continuación se puede realizar un paso de post-procesado para calcular las estadísticas de la estimación o sus resultados. Su carácter optativo es una ventaja en cuanto a los grandes conjuntos de datos que llevarían a un gasto excesivo de tiempo y posibles problemas de memoria computacional.

Por último, esta librería aporta funciones para la explotación de la distribución a posteriori obtenida de nuestro modelo permitiendo la elaboración de gráficas de los resultados obtenidos, de descripción de los residuos, de distribuciones marginales de los términos de interés; la realización de resúmenes o la predicción nuevos valores.

Otra ventaja es la inclusión de opciones para la disminución del uso de memoria y una reducción considerable en el tiempo de computación respecto a **‘jagsUI’**.

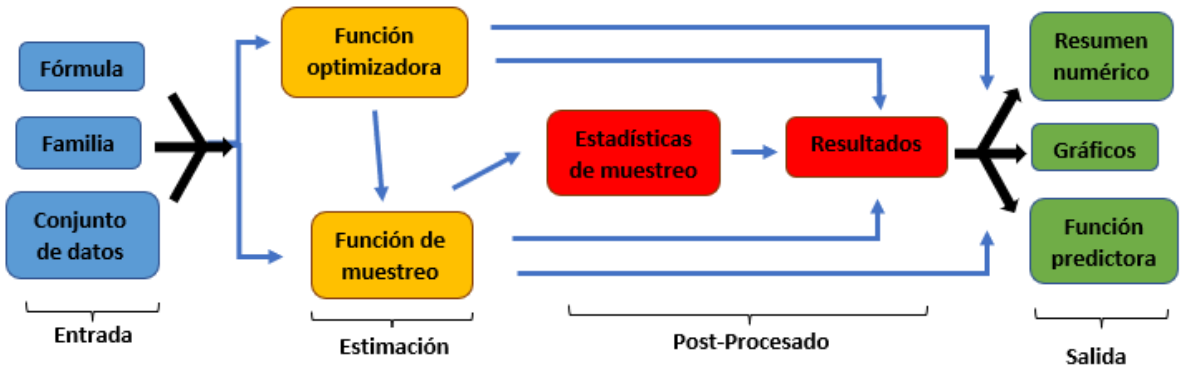

Figura 3.1: **Arquitectura ‘bamlss’.** Las líneas azules representan caminos opcionales. Se puede usar primero la función de muestreo y luego la optimizadora o directamente la función de muestreo.

Capítulo 4

Resultados de la modelización de la temperatura máxima diaria

En este capítulo se aplicará la metodología propuesta en el Capítulo 3 para modelar tres series temporales siguiendo la línea de trabajo propuesta por Castillo-Mateo et al. [2]. Vamos a considerar las series de localizaciones con climas diversos para estudiar la adecuación del modelo a las variaciones del clima y para determinar las posibles adaptaciones a realizar en el modelo.

En la primera sección del capítulo se hará un análisis exploratorio, incluyendo el trabajo descriptivo usando los modelos lineales. En la segunda sección se propondrán los modelos bayesianos empleando el software **'bamlss'** y se explotará el modelo final.

Debido al elevado número de gráficos y tablas con resúmenes obtenidos, se mostrarán los más informativos en el cuerpo del trabajo, mientras que los resultados completos aparecerán en los anexos. El código de R desarrollado para obtener los resultados se puede encontrar en el repositorio de GitHub: `https://github.com/JavierTorcal/TFG-Bayesian-models-for-climate-time-series`.

4.1. Análisis exploratorio

Las series temporales se obtienen de la base de datos ECA&D [9]. El proyecto ECA&D (European Climate Assessment `https://www.ecad.eu/`) es una iniciativa dirigida por los Institutos de Metereología europeos para disponer de una base de datos de referencia para la Climatología.

Primero se depuran los datos debido a que las series contienen algunos ausentes y en tal caso están representados con valor -9999. Se ha decidido emplear tres series de ciudades españolas con climas variados como tienen Soria, Zaragoza y Bilbao. Como se muestra en el Cuadro 4.1, Soria tiene el clima más frío debido a que se encuentra en una meseta a 1,000 metros de elevación. En cambio, Zaragoza se sitúa a menor elevación, a unos 200 metros, y se localiza en el valle del Ebro, entonces sus temperaturas máximas son notablemente más altas. Cuando se comparan los rangos intercuartílicos, ambas series temporales son más extremas que la de Bilbao, donde se aprecia la influencia del Mar Cantábrico que suaviza las temperaturas tanto en las épocas cálidas del año como en las frías.

Ciudad	Mín.	Q1	Mediana	Media	Q3	Máx.	IQR
Soria	-7.70	10.00	16.10	16.95	24.00	38.00	14.00
Zaragoza	-3.40	13.80	20.20	20.72	27.60	44.50	13.80
Bilbao	-4.00	14.60	19.20	19.13	23.40	41.90	8.80

Cuadro 4.1: Resumen numérico de las series temperaturas máximas diarias (ºC) en 1951-2020.

Para la adecuación de nuestras series, se ha diseñado una función en R recogida en el Anexo A.F.1 que se ayuda de la librería **'lubridate'** para conseguir que las fechas tengan un formato conveniente. Dicha función filtrará los datos correspondientes al periodo 1951-2020 y eliminará los 29 de Febrero de los

años bisiestos para disponer de 7 décadas de 3,650 días cada una, dando un total de 25,550 temperaturas máximas. El número de datos perdidos es 3, 61 y 64, en Zaragoza, Soria y Bilbao respectivamente, luego no supondrán un inconveniente. También incluirá en nuestros conjuntos de datos el valor de las covariables a utilizar como los armónicos correspondientes a cada día del año o el valor de la temperatura en el día previo.

Se exploran cuáles son los elementos que van a explicar la temperatura máxima de cada localidad. La tendencia térmica a lo largo de los años aparece reflejada en los suavizados (loess) aplicados sobre la serie de la media anual de T.máx., ver Figura 4.1. Como nos sugiere la figura, incluiremos en nuestros modelos una tendencia térmica sobre la temperatura máxima diaria, como es esperable de acuerdo a la evidencia científica sobre el calentamiento del planeta.

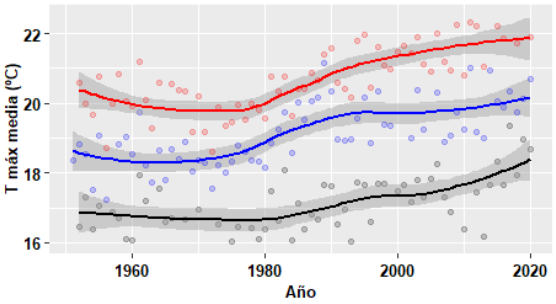

Figura 4.1: Valor medio anual de T.máx, la línea representa el suavizado loess, en Soria (negro), Bilbao (azul) y Zaragoza (rojo).

4.1.1. Modelos lineales para la media

Los modelos lineales nos resultarán útiles para medir la aportación a la variabilidad de cada término. Inicialmente se han considerado modelos con diferentes covariables pero aquí se resumen los resultados de los tres más interesantes por su proceso constructivo y que servirán para discutir los resultados del exploratorio para la inclusión de covariables.

M1 Modelo básico incluyendo la tendencia lineal con 2 armónicos para la estacionalidad.

$$Y_t = \beta_0 + \sum_{i=1}^{2}(\beta_{i,s}Si_t + \beta_{i,c}Ci_t) + \alpha_0 Year_t + \varepsilon_t, \quad \varepsilon_t \sim N(0,\sigma^2). \tag{4.1}$$

M2 Modelo AR(1) conteniendo las covariables anteriores e incluyendo el término Y_{t-1}.

$$Y_t = \beta_0 + \sum_{i=1}^{2}(\beta_{i,s}Si_t + \beta_{i,c}Ci_t) + \alpha_0 Year_t + \rho_0 Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0,\sigma^2). \tag{4.2}$$

M3 Modelo final conteniendo las covariables del último modelo pero incluyendo las interacciones de estacionalidad con tendencia lineal y estacionalidad con Y_{t-1}. Mediante un procedimiento paso a paso ('stepwise') se seleccionan las variables para obtener el modelo que minimiza su AIC. Dependiendo de las características particulares de cada ciudad el modelo se reduce a diferentes covariables. En el caso de Soria se obtiene:

$$Y_t = \beta_0 + \sum_{i=1}^{2}(\beta_{i,s}Si_t + \beta_{i,c}Ci_t) + \alpha_0 Year_t + (\rho_0 + \rho_{1,s}S1_t + \rho_{1,c}C1_t)Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0,\sigma^2). \tag{4.3}$$

En Bilbao y Zaragoza, el procedimiento 'stepwise' reduce la ecuación a:

$$Y_t = \beta_0 + \sum_{i=1}^{2}(\beta_{i,s}Si_t + \beta_{i,c}Ci_t) + (\alpha_0 + \alpha_{1,s}S1_t + \alpha_{1,c}C1_t)Year_t + \left(\rho_0 + \sum_{i=1}^{2}(\rho_{i,s}Si_t + \rho_{i,c}Ci_t)\right)Y_{t-1} + \varepsilon_t. \tag{4.4}$$

La T.máx. diaria es fuertemente dependiente de los cambios en el nivel de radiación solar propios de cada estación del año. Para recoger dicha estacionalidad, incluiremos las funciones armónicas como covariables tal como se muestra en el modelo M1 y se comparará con un modelo conteniendo el primer armónico. En la Figura 4.2 se representa la media móvil de T.máx. en Soria, con una ventana de 30 días para suavizar los cambios entre días consecutivos y su valor medio en cada día del año (línea negra). La figura incluye el perfil obtenido del ajuste del modelo M1 con dos armónicos (línea verde), y el obtenido con 1 armónico (rojo). Se observa que la línea verde que representa el ajuste del modelo M1 se adapta mejor al perfil empírico de la estacionalidad de la media de T.máx. (en negro) que la línea roja del modelo con un solo armónico. Los gráficos análogos pertenecientes a las series de Zaragoza y Bilbao muestran un comportamiento equivalente y se pueden consultar en el anexo, ver Figuras A3 y A4.

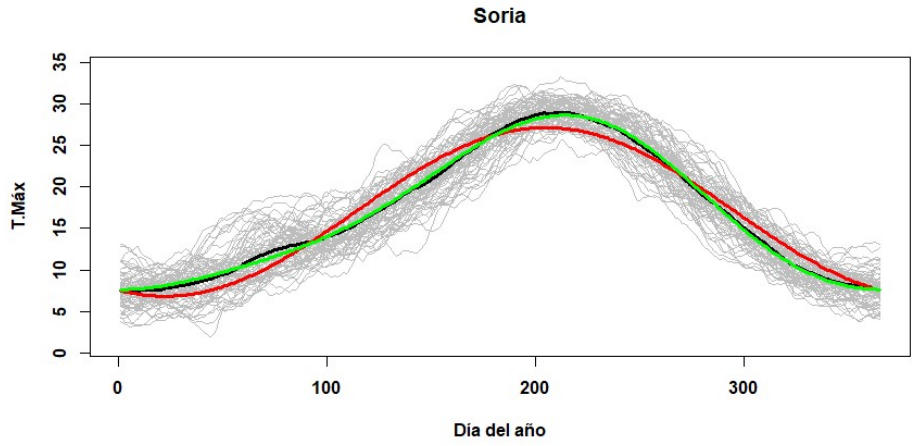

Figura 4.2: Se representa la media móvil de T.máx., de ventana 30 días, y su valor medio en cada día del año (línea negra), el perfil obtenido del ajuste del modelo M1 con dos armónicos (línea verde), y el obtenido con 1 armónico (rojo), en Soria.

Se va a estudiar la correlación serial de los datos de las series diarias. En la Figura 4.3, fila superior, se muestra el gráfico ACF aplicado sobre los residuos del Modelo M1, que incluye la tendencia lineal y dos armónicos para la estacionalidad. Se observa la clara correlación serial entre los residuos de los días consecutivos. Esto motivará el uso de modelos con término autorregresivo. Se comprueba cuando se analiza el gráfico ACF sobre los residuos del modelo M2 (fila inferior de la figura), que con la inclusión del término AR(1) parece expresarse correctamente la dependencia del día anterior ya que la autocorrelación de ordenes 1, 2 y 3 se sitúa dentro de los límites de confianza (línea roja disjunta).

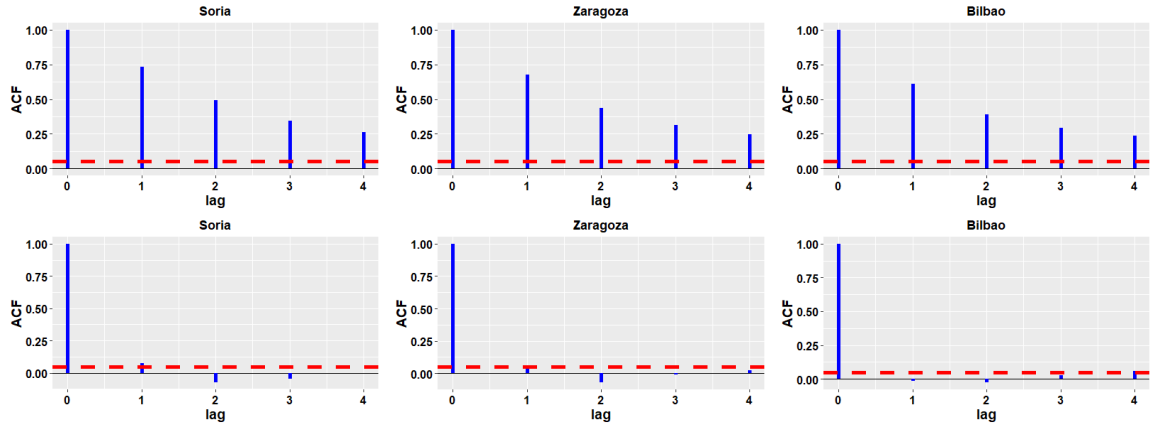

Figura 4.3: Gráfico ACF para residuos de los Modelo lineales (4.1) en la fila superior y (4.2) en la fila inferior de Soria, Zaragoza y Bilbao.

Se van a analizar las interacciones entre la tendencia lineal y los armónicos que expresan la estacionalidad y entre el efecto del día anterior y los armónicos. Para ello vamos a representar en la Figura 4.4 los efectos de las interacciones recogidas en el modelo M3 (4.3) en Soria y (4.4) en Zaragoza y Bilbao. En la izquierda se representa la tendencia estimada en cada día del año, considerando la interacción de la pendiente con los armónicos, en línea roja Zaragoza y en azul Bilbao. Se deduce que en verano es cuando existe una mayor tendencia térmica mientras que en invierno es menor con una diferencia en torno a 0.01ºC/año. Notar que que tras realizar la selección de variables no se incluía la interacción entre la tendencia lineal y la estacionalidad en el modelo de Soria. En la derecha de la figura, se representa el coeficiente de Y_{t-1} en cada día del año, la línea negra corresponde al modelo de Soria. En el caso del término Y_{t-1}, destaca Bilbao donde en verano hay una clara menor dependencia de la T.máx. del día anterior, debido a la influencia amortiguadora del mar sobre las temperaturas. Como se muestra es importante la inclusión de estas interacciones ya que con un coeficiente fijo no ajustaría adecuadamente.

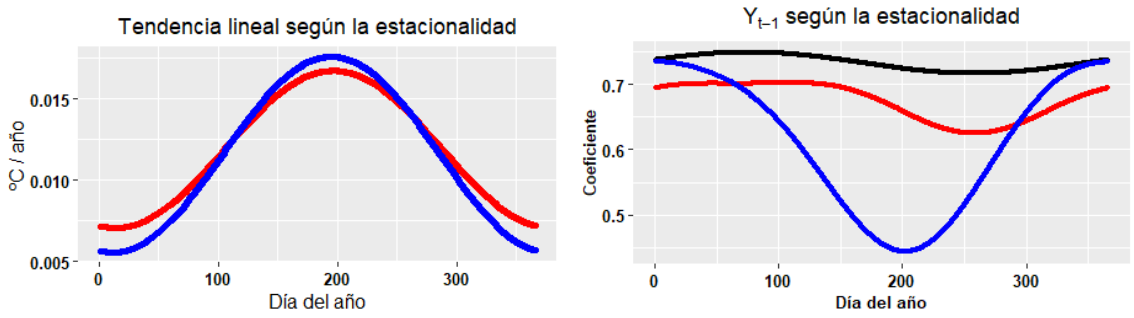

Figura 4.4: Izda., tendencia estimada en cada día del año, considerando la interacción de la pendiente con los armónicos. Dcha., coeficiente de Y_{t-1} en cada día del año. La línea negra corresponde al modelo de Soria, la roja al de Zaragoza y la azul al de Bilbao.

4.1.2. Análisis exploratorio de la varianza

De forma análoga al estudio exploratorio sobre la respuesta Y_t para identificar posibles efectos en la media, vamos a utilizar los residuos de los Modelos (4.3) y (4.4) para explorar la dependencia de la varianza frente al año, al día del año y al residuo del día anterior. Para ello utilizaremos como variable respuesta el cuadrado de dichos residuos.

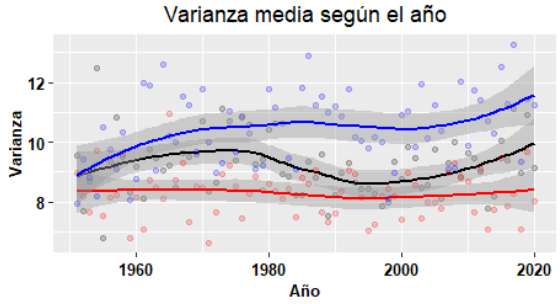

Figura 4.5: Varianza obtenida del cuadrado de los residuos de los Modelos (4.3) y (4.4) respecto al año y suavizado loess. En rojo Zaragoza, en negro Soria y en azul Bilbao.

En la Figura 4.5 se observa el suavizado loess del cuadrado de los residuos del modelo M3 respecto al año. Se representa en rojo Zaragoza, en negro Soria y en azul Bilbao. Destaca que su valor en Zaragoza es aproximadamente constante lo que es un indicio de que no hay tendencia en ella. En el caso de Bilbao ha habido una clara tendencia positiva a lo largo de los años.

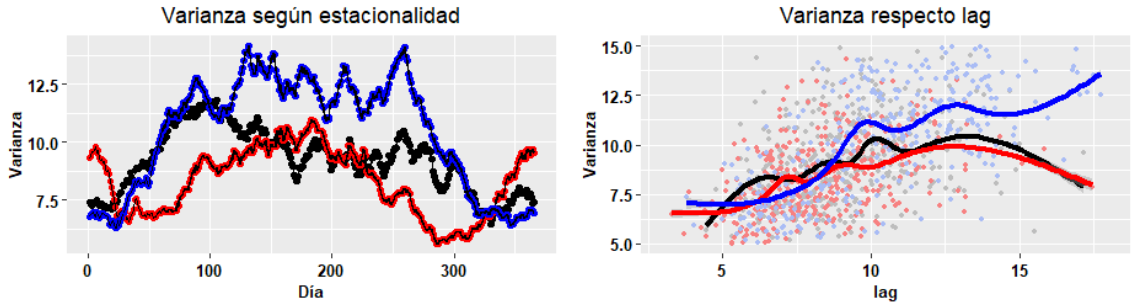

Figura 4.6: Izda., varianza obtenida de una ventana móvil de 15 días de los residuos al cuadrado del modelo M3, en Soria (4.3) (trazo negro), (4.4) en Zaragoza (rojo) y Bilbao (azul). Dcha., varianza obtenida de los residuos al cuadrado del modelo M3 para cada valor de su retardo y suavizado loess.

La Figura 4.6, izquierda, muestra frente a cada día del año la varianza obtenida de una ventana móvil de 15 días de los residuos al cuadrado del modelo M3, en (4.3) Soria (trazo negro), (4.4) en Zaragoza (rojo) y Bilbao (azul). Se encuentra una clara heterocedasticidad, que refleja un cambio estacional de la varianza, mayor en los días centrales de verano y menor durante el otoño e invierno. En la parte derecha de la figura se representa la varianza obtenida de los residuos al cuadrado del modelo M3 para cada valor del retardo del residuo, con su suavizado loess. Aparece un cambio en la varianza ligado al término AR(1), que tiende a crecer cuanto mayor es el residuo del día previo.

Para estudiar con qué variables se podrá explicar la varianza de los modelos bayesianos ajustaremos un modelo lineal para cada localidad, de forma análoga a lo realizado para la media. En este caso, la variable respuesta serán los residuos al cuadrado de los Modelos (4.3) y (4.4) y se usará un enlace logarítmico con el predictor lineal para asegurar que los valores ajustados sean positivos.

En el caso de Zaragoza debido a lo mostrado en la Figura 4.5 no consideramos la variable de la tendencia lineal para el modelado de su varianza. Tras realizar el procedimiento paso a paso ('stepwise') de selección de variables, también se rechaza la tendencia lineal y sus interacciones en Soria.

El modelo resultante para la varianza de Soria y Zaragoza es el siguiente:

$$E[\varepsilon_t^2] = \exp\left\{\beta_0^\sigma + \sum_{i=1}^{2}(\beta_{i,s}^\sigma Si_t + \beta_{i,c}^\sigma Ci_t) + \left(\rho_0^\sigma + \sum_{i=1}^{2}(\rho_{i,s}^\sigma Si_t + \rho_{i,c}^\sigma Ci_t)\right)\varepsilon_{t-1}^2\right\}. \tag{4.5}$$

En el modelo de Bilbao, no se prescinde ningún término:

$$E[\varepsilon_t^2] = \exp\left\{\beta_0^\sigma + \sum_{i=1}^{2}(\beta_{i,s}^\sigma Si_t + \beta_{i,c}^\sigma Ci_t) + \left(\alpha_0^\sigma + \sum_{i=1}^{2}(\alpha_{i,s}^\sigma Si_t + \alpha_{i,c}^\sigma Ci_t)\right)Year_t + \left(\rho_0^\sigma + \sum_{i=1}^{2}(\rho_{i,s}^\sigma Si_t + \rho_{i,c}^\sigma Ci_t)\right)\varepsilon_{t-1}^2\right\}. \tag{4.6}$$

En el anexo, se puede consultar el Cuadro A1, donde se recoge el valor de los coeficientes asociados a los parámetros de los modelos finales para la media y la varianza en cada uno de los observatorios. No se incluyen intervalos de confianza ya que no se está realizando inferencia.

4.2. Modelos bayesianos

Vamos a ajustar el modelo final para cada ciudad en su versión bayesiana mediante el paquete estadístico **'bamlss'**. En este caso, vamos a modelar la media mediante las covariables y también la varianza mediante los términos estudiados previamente para reflejar la heterocedasticidad identificada en la sección anterior. Se utiliza la familia gaussiana para la variable respuesta y en el caso de la varianza un enlace logarítmico. A continuación se muestran los modelos sobre los que se mostrarán resultados:

MB1 **Primer modelo bayesiano:** Incluye solo la tendencia lineal y la estacionalidad como variables predictoras y sus interacciones para modelar la media y la varianza.

MB2 **Segundo modelo bayesiano:** Basado en el modelo lineal final para la media pero considerando la varianza constante.

MBF **Modelo bayesiano final:** Utiliza todas las covariables que han resultado seleccionadas en el exploratorio de los modelos lineales tanto para la media como para la varianza.

En el Cuadro 4.2 se compara el DIC de los tres modelos para determinar cual es el menor, en cada observatorio. Notar que el mejor modelo es MBF por tener el menor DIC en las tres ciudades. Considerando estos resultados, se explotará el modelo bayesiano final.

	Soria			Zaragoza			Bilbao		
	MB1	MB2	MBF	MB1	MB2	MBF	MB1	MB2	MBF
DIC	148,504	128,913	128,594	142,154	126,552	125,974	144,264	132,072	130,094

Cuadro 4.2: DIC de los modelos bayesianos propuestos.

En el Cuadro 4.3 aparece la media a posteriori y el intervalo de credibilidad a posteriori al 95 % de los parámetros del MBF en Soria. Se resaltan en negrita aquellos intervalos de credibilidad que no contienen el valor 0. Las distribuciones a posteriori de los parámetros de los modelos de Zaragoza y Bilbao se recogen en el anexo, ver Cuadros A.D.1 y A.D.1.

Modelo bayesiano final Soria							
μ	Media	2.5 %	97.5 %	σ	Media	2.5 %	97.5 %
$Year_t$	0.007	**0,006**	**0,009**				
$S1_t$	-1.179	**-1,391**	**-0,971**	$S1_t$	-0.0367	-0.0848	0.0163
$C1_t$	-2.664	**-2,875**	**-2,455**	$C1_t$	-0.0830	**-0,1368**	**-0,0326**
$S2_t$	0.567	**0,489**	**0,649**	$S2_t$	0.0729	**0,0361**	**0,1053**
$C2_t$	0.048	-0.031	0.128	$C2_t$	-0.0113	-0.0460	0.0253
Y_{t-1}	0.730	**0,722**	**0,738**	Y_{t-1}	0.0069	**0,0049**	**0,0089**
$Y_{t-1} : S1_t$	0.014	**0,002**	**0,026**	$Y_{t-1} : S1_t$	0.0061	**0,0030**	**0,0088**
$Y_{t-1} : C1_t$	0.002	-0.009	0.013	$Y_{t-1} : C1_t$	0.0038	**0,0012**	**0,0067**
				$Y_{t-1} : S2_t$	-0.0023	**-0,0041**	**-0,0005**
				$Y_{t-1} : C2_t$	-0.0022	**-0,0040**	**-0,0004**

Cuadro 4.3: Resumen de la distribución a posteriori de los parámetros del modelo bayesiano final para μ y σ, incluyendo media e intervalo de credibilidad al 95 %, en Soria. Los intervalos de credibilidad que no contienen al cero se resaltan en negrita.

En el anexo se comprueba la convergencia correcta de las cadenas MCMC asociadas al modelo MBF. Se recoge el tamaño de muestra efectivo para cada parámetro, ver Tabla A8 y se muestran los 'trace plot' para cada parámetro con su valor $\hat{R}$ correspondiente, ver Figuras A11, A12 y A13.

4.2.1. Explotación del modelo final

El modelo bayesiano final introduce efectos que expresan interacciones entre diferentes términos. En particular, los términos $Year_t$ para la tendencia lineal e Y_{t-1} para la dependencia del día previo se interaccionan con los armónicos para variar el valor de su efecto según el día del año. Esto aporta mayor flexibilidad al modelo ya que con un término único no se expresa adecuadamente dicho efecto, como se verá a continuación en las Figuras 4.7 para Soria y 4.8 para Zaragoza y Bilbao.

Vamos a mostrar en la Figura 4.7 el efecto de las interacciones para la media y la varianza en el modelo MBF en Soria. Se representa la función de densidad de la distribución a posteriori del coeficiente de Y_{t-1} en cuatro fechas distribuidas por todas las estaciones del año, 20 de Abril (verde), 29 de Junio (rojo), 23 de Septiembre (amarillo) y 29 de Diciembre (azul). A la izquierda se representa el efecto en el

submodelo de la media y a la derecha en el submodelo para la varianza. En gris se representa el intervalo de credibilidad y la media a posteriori del coeficiente de Y_{t-1}.

Se encuentra que las medias a posteriori del efecto de Y_{t-1} para el 23 de Septiembre y para el 20 Abril no se sitúan dentro del intervalo de credibilidad al 95 % de la distribución a posteriori del coeficiente ρ_0 de Y_{t-1}. Es clara la necesidad de la interacción con los armónicos para expresar adecuadamente el cambio en el efecto del día anterior según la estación del año.

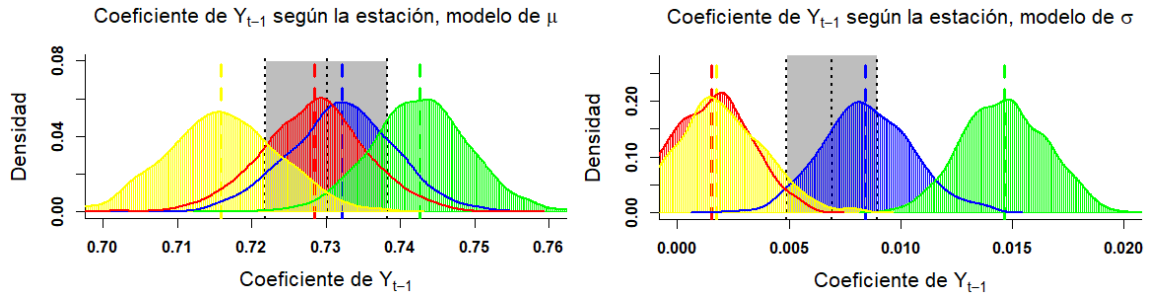

Figura 4.7: Izda., distribución a posteriori del coeficiente de Y_{t-1} en el submodelo de la media para el 20 de Abril (verde), 29 de Junio (rojo), 23 de Septiembre (amarillo) y 29 de Diciembre (azul), en el modelo de Soria. Dcha., gráfico equivalente para el submodelo para la varianza. En gris se representa el intervalo de credibilidad y la media a posteriori del coeficiente de Y_{t-1}.

La Figura 4.8 representa los efectos en esas 4 fechas para los modelos bayesianos de Bilbao y Zaragoza, mostrando sus distribuciones a posteriori mediante diagramas de cajas puesto que aparecen más interacciones. Esta figura ilustra la necesidad de incluir en el modelo todas las interacciones representadas para ambos términos Y_{t-1} y $Year_t$. En todas las gráficas existe al menos un día donde la línea central de la caja del diagrama, que representa la mediana, se encuentra fuera de la banda gris que señala el intervalo de credibilidad al 95 % de los parámetros ρ_0 y ρ_0^{σ} de Y_{t-1} o α_0 de $Year_t$.

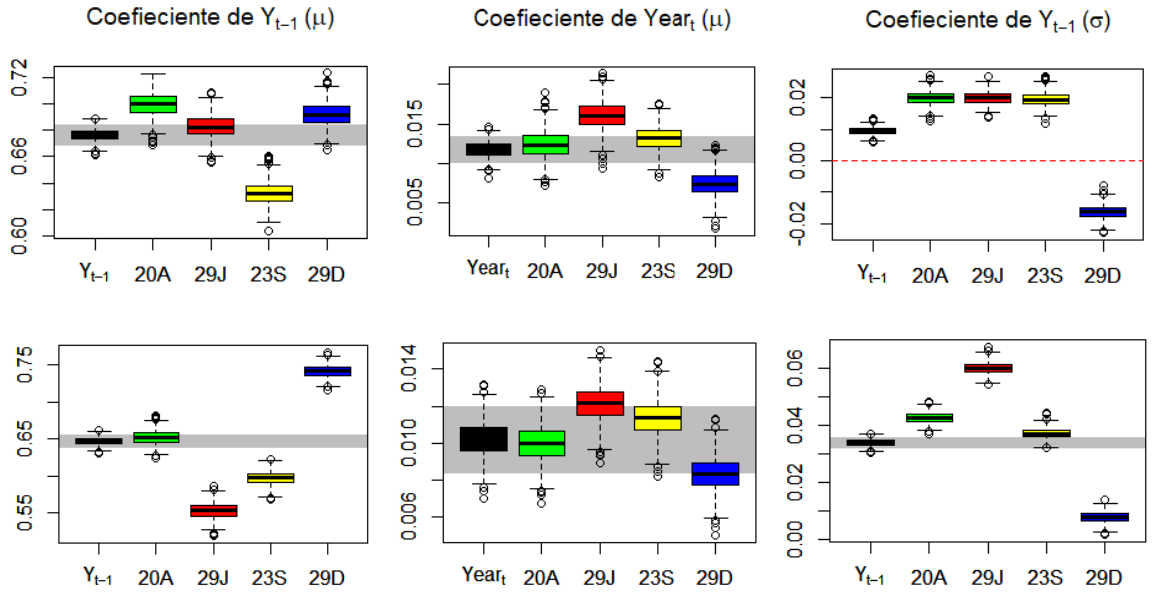

Figura 4.8: Box-plot de las distribuciones a posteriori del efecto asociado a 4 fechas del año estimado a partir de los parámetros con interacciones, en verde para el 20 de Abril, en rojo para el 29 de Junio, en amarillo para el 23 de Septiembre, en azul para el 29 de Diciembre, en Zaragoza (fila superior) y Bibao (fila inferior). En negro se representa el efecto sin interacción, en gris su intervalo de credibilidad.

Podemos comparar los resultados mostrados en las Figuras 4.7 y 4.8 con los obtenidos por Castillo-Mateo et al. [2]. El coeficiente de autocorrelación en el tiempo de verano que hemos obtenido en Zarago-

za es mayor que en Bilbao que se encuentra al noroeste y menor que en Soria que se encuentra al oeste, esto corresponde con el resultado para la región de Aragón que muestra un gradiente negativo hacia el Noroeste. El valor de la media a posteriori del coeficiente de autorregresión en Zaragoza es de 0.66 en el trabajo referido y en nuestro modelo 0.68 (considerando el día 29 de Junio), ver Cuadro A6.

Finalmente, se estima el efecto del calentamiento aprovechando la capacidad predictiva del modelo MBF. Para obtener estos resultados se ha programado una función en R recogida en el Anexo A.F.2. Dado que nuestras series contienen datos de 70 años, se va a estimar la distribución a posteriori de la diferencia de temperaturas diarias entre el periodo inicial (1952-1960) y el periodo final (2012-2020).

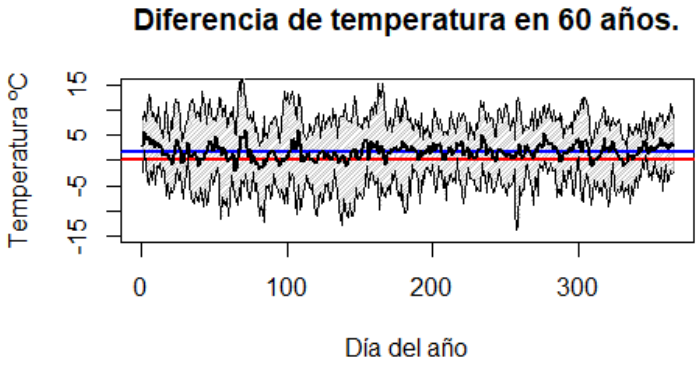

Figura 4.9: Intervalo de credibilidad (en gris) y media a posteriori (negro) de la diferencia de T.máx entre los periodos inicial y final frente al día del año, en Soria. La línea horizontal roja corresponde a 0ºC y la azul al incremento medio de la T.máx.

La Figura 4.9 muestra el incremento de temperaturas máximas el mismo día entre ambos periodos en Soria. La línea negra muestra la media a posteriori de la diferencia y se encuentra casi todos los días del año por encima de 0ºC (línea roja). La línea azul muestra la media del incremento en la T.máx. diaria tras 60 años cuyo valor es de 1.57ºC. Se observa que el intervalo de credibilidad (en gris) contiene a 0ºC debido a la gran variabilidad de T.máx. en un día cualquiera. Las figuras análogas para Zaragoza y Bilbao se pueden consultar en el anexo, ver figura A10. Resulta interesante que en Zaragoza con un incremento de T.máx. más pronunciado, el intervalo de credibilidad no contiene a 0ºC en bastantes días.

	Soria	Zaragoza	Bilbao
Incremento medio de T.máx. (ºC)	1.571	2.066	1.579
Probabilidad a posteriori de mayor T.máx.	0.6263	0.7153	0.6817

Cuadro 4.4: Valor medio de la distribución a posteriori del incremento de la T.máx en cada observatorio entre 1952-1960 y 2012-2020, y la probabilidad a posteriori de que un día del año haga más calor que el mismo día hace 60 años.

En el Cuadro 4.4 se muestra el valor medio a posteriori del incremento en la T.máx. diaria para cada observatorio y la probabilidad a posteriori de que un día del año haga más calor que el mismo día hace 60 años. Este último estadístico se obtiene con la proporción de elementos positivos de la matriz que contiene las diferencias de T.máx. entre ambos periodos a partir de las simulaciones MCMC. Las probabilidades estimadas son muy superiores a 0.5, valor que correspondería a comparar dos periodos en situación estacionaria. La mayor evidencia de calentamiento se encuentra en Zaragoza, en valor absoluto del incremento y en la probabilidad de sufrir mayores temperaturas en el último periodo.

En cuanto a la tendencia al calentamiento para la T.máx en los meses de Mayo a Septiembre, Castillo-Mateo et al. obtienen 0.42ºC/década en Zaragoza mientras que en nuestro trabajo el incremento medio a posteriori de la T.máx. en Zaragoza se ha estimado en dichos meses en 0.49ºC/década. No resulta extraña una variación en los resultados debido a la diferencia de periodos considerados.

Capítulo 5

Conclusiones

En este trabajo se han revisado los conceptos necesarios para plantear modelos bayesianos autorregresivos que han servido para representar las series de temperatura máxima diaria. En particular, se han considerado modelos locales que se inspiran en la modelización espacio-temporal de Castillo-Mateo et al. [2].

La metodología propuesta incluye una fase exploratoria donde se utilizan gráficas descriptivas, así como el uso de modelos de regresión lineal para explorar qué efectos pueden ser relevantes en la modelización, tanto de la media como de la varianza. Las herramientas exploratorias en este trabajo serán útiles para plantear las características de la temperatura diaria en otros observatorios peninsulares con climas diferentes.

Los modelos bayesianos se han ajustado utilizando la librería de R **'bamlss'**, que permite definir modelos para los parámetros de posición y escala en distintas distribuciones. Se han definido modelos con distribución gaussiana, donde tanto la media como la varianza se hacen depender de covariables adecuadas.

La modelización en Soria, Zaragoza y Bilbao ha identificado componentes habituales de la variabilidad de la temperatura máxima diaria, en particular, se ha reflejado el comportamiento estacional, la correlación serial y la tendencia temporal. Estas componentes se encuentran también en el modelo espacio-temporal para las series de Aragón de Castillo-Mateo et al. [2], donde se consideran únicamente las subseries asociadas a los meses de Mayo a Septiembre. Aunque los resultados no son directamente comparables con los obtenidos aquí para las series completas, se pueden establecer algunas similitudes. La variabilidad espacial aparece expresada en los modelos locales con distintas componentes y también con distintos valores estimados, por ejemplo en el caso de la tendencia.

El análisis de carácter anual ha permitido identificar algunos comportamientos estacionales que no aparecen cuando se consideran únicamente los meses más cálidos. En el caso del modelo para el valor medio se debe incluir términos que reflejen que la correlación serial varía a lo largo del año. La mayor diferencia es la necesidad de modelar la varianza para reflejar la menor variabilidad en invierno, además de una mayor variabilidad ligada a la situación cálida en el día previo.

Este trabajo deja abierta una línea a desarrollar, puesto que se ha identificado que la modelización espacio-temporal de series a lo largo de todo el año requerirá completar la propuesta de Castillo-Mateo et al. [2] modelando la varianza.

Bibliografía

[1] P.J. BROCKWELL, R.A. DAVIS (2002). *Introduction to time series and forecasting*. Springer.

[2] 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*. https://doi.org/10.1007/s13253-022-00493-3.

[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] K. KELLNER (2021). **'jagsUI'**: A Wrapper Around **'rjags'** to Streamline **'JAGS'** Analyses. *R package version 1.5.2*. https://CRAN.R-project.org/package=jagsUI.

[6] 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.

[7] M. PLUMMER, N. BEST, K. COWLES, K. VINES (2006). **'CODA'**: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6, 7–11.

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

[9] 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.

[10] N. UMLAUF, N. KLEIN, T. SIMON, A. ZELEIS (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.

Anexo

A.A. Análisis exploratorio para la media

A.A.1. Gráficos exploratorios para la evolución temporal a largo plazo

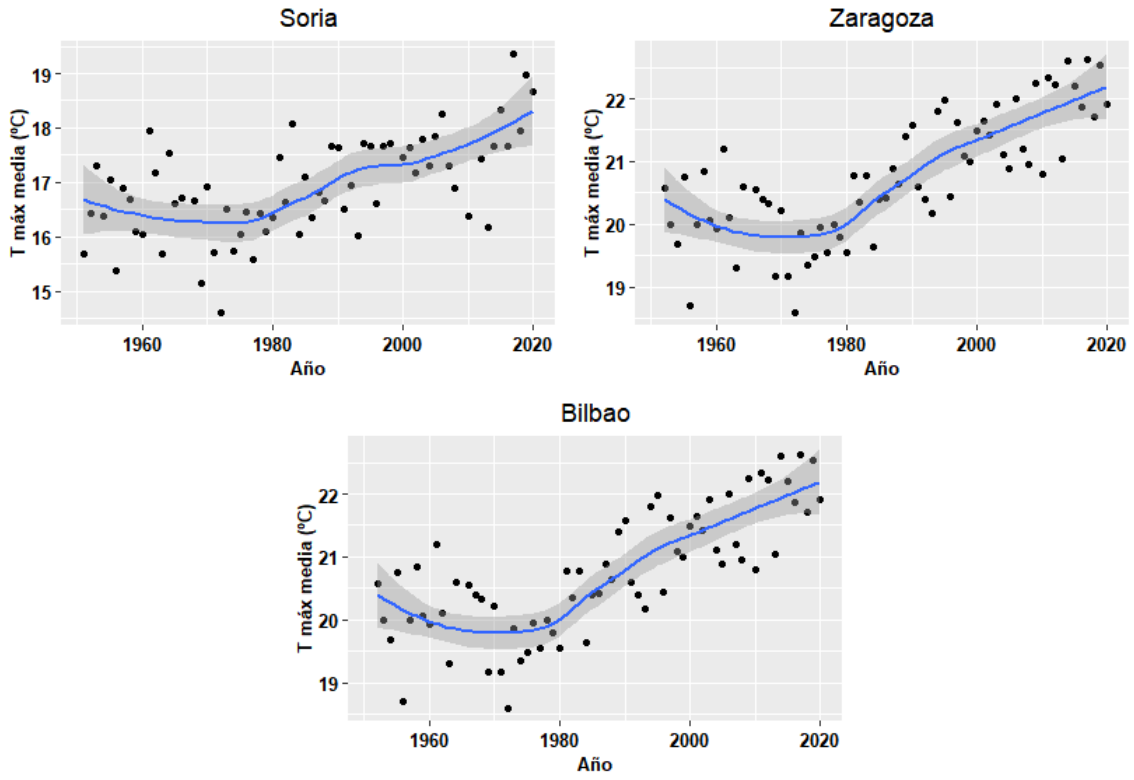

Figura A1: Tendencia lineal en las diferentes localidades.

A.A.2. Soria

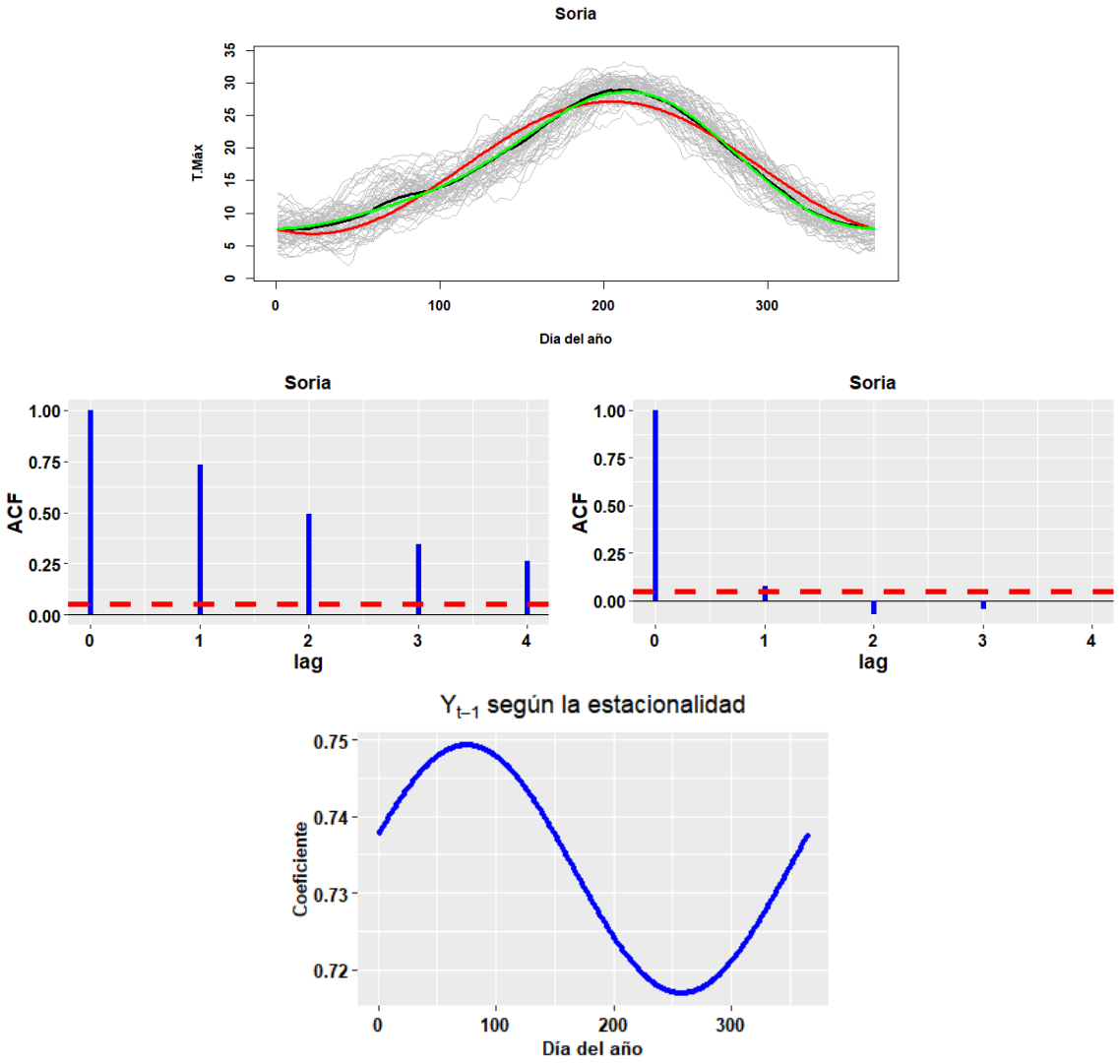

Figura A2: Análisis exploratorio para la media, Soria.

A.A.3. Zaragoza

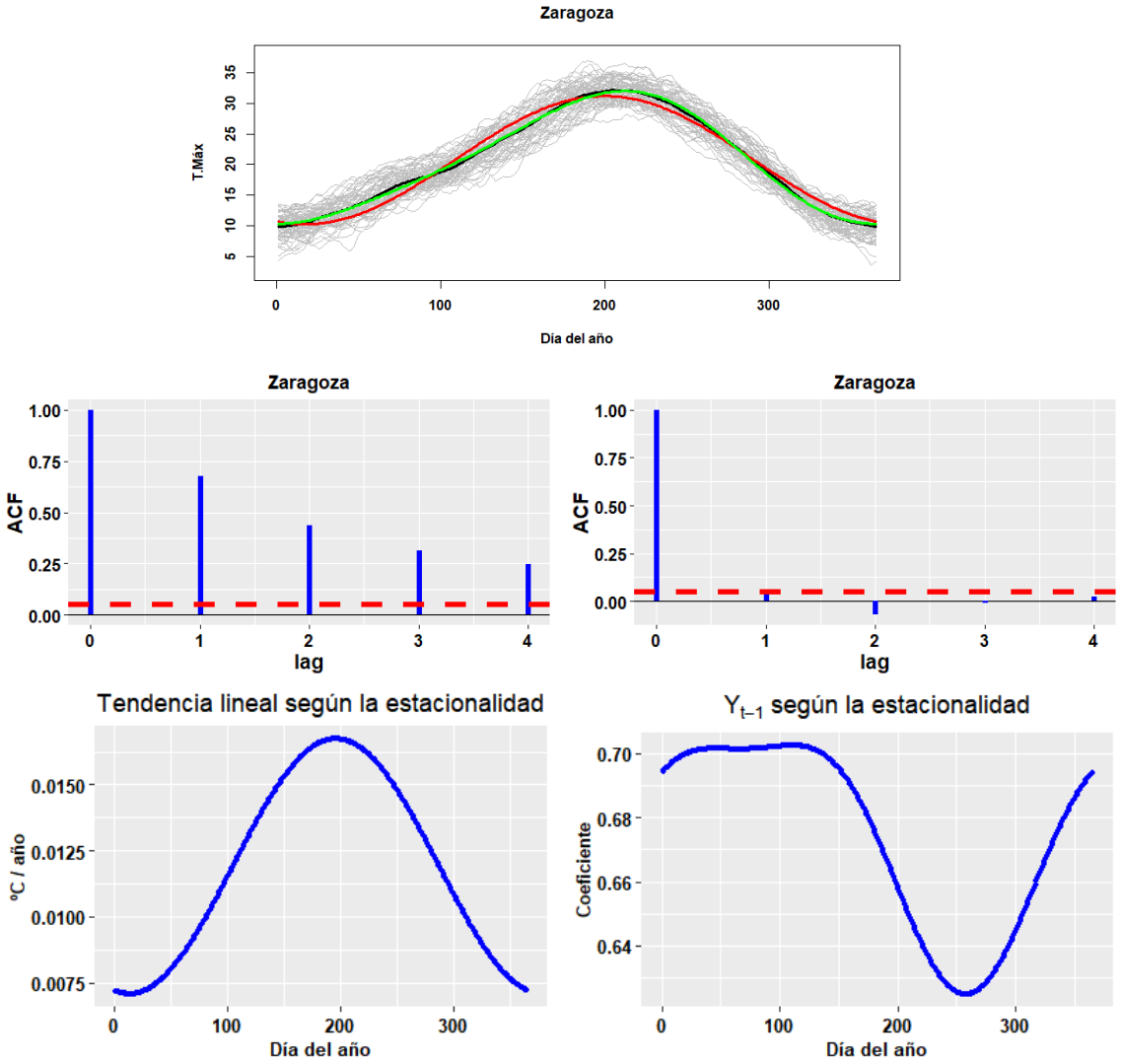

Figura A3: Análisis exploratorio para la media, Zaragoza.

A.A.4. Bilbao

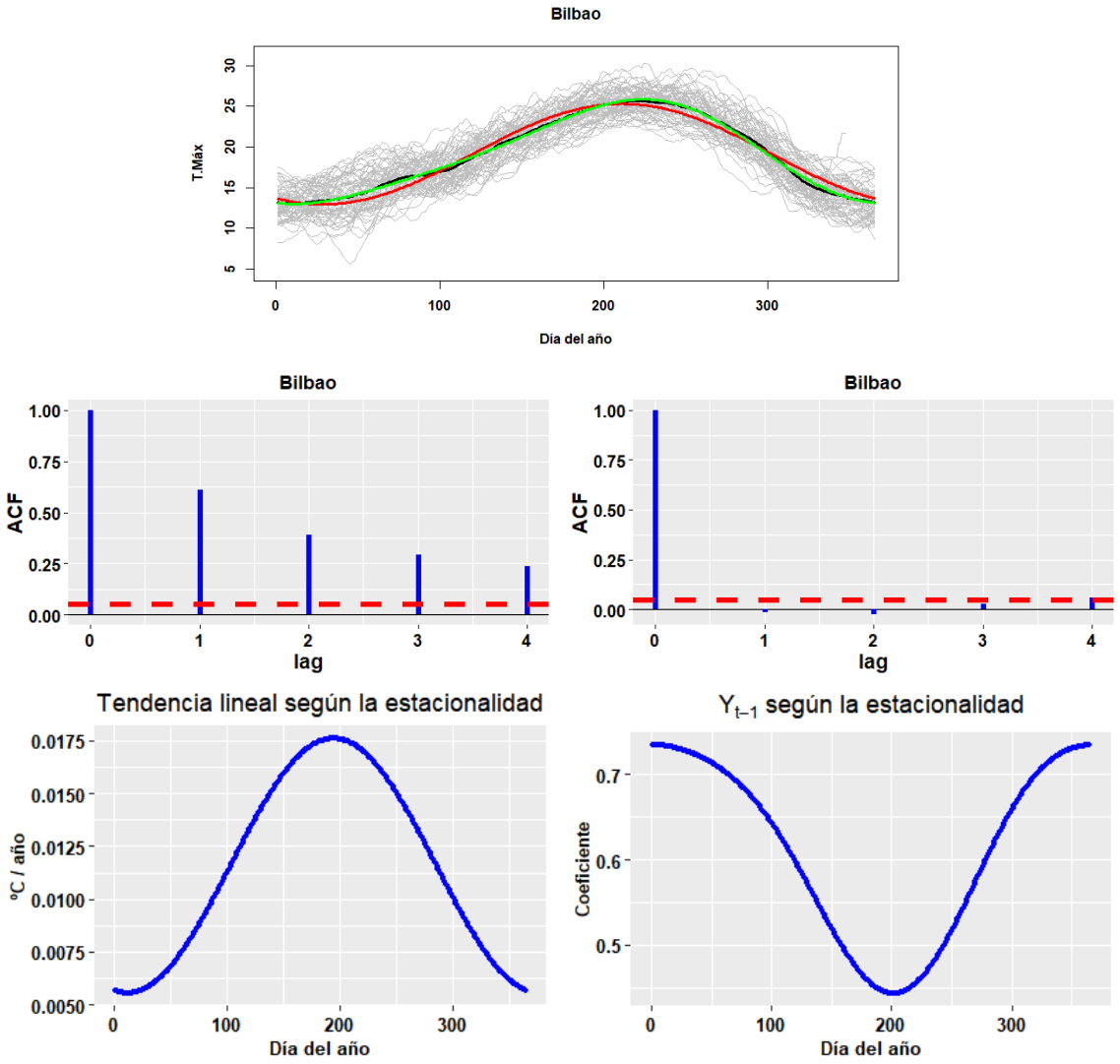

Figura A4: Análisis exploratorio para la media, Bilbao.

A.B. Gráficos para exploratorio de la varianza

A.B.1. Soria

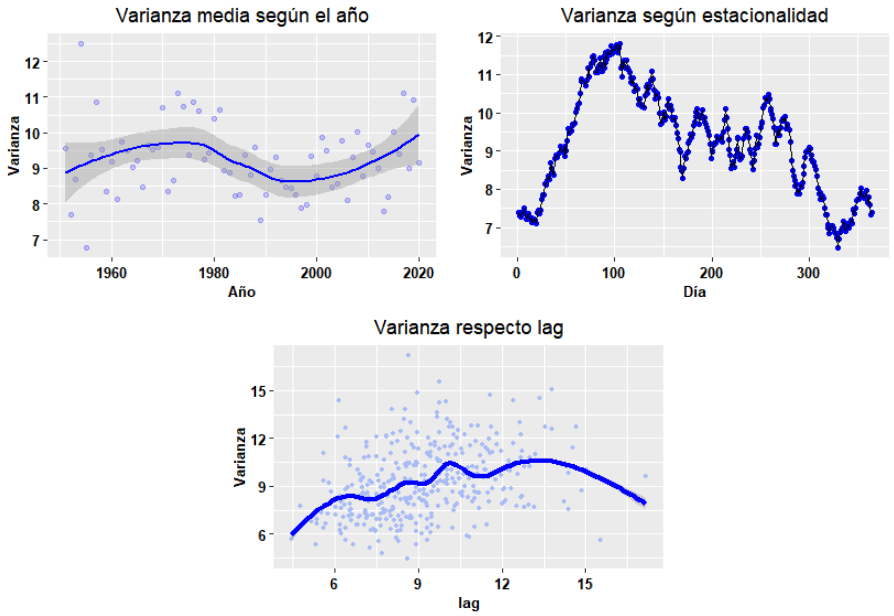

Figura A5: Gráficos para exploratorio de la varianza, Soria.

A.B.2. Zaragoza

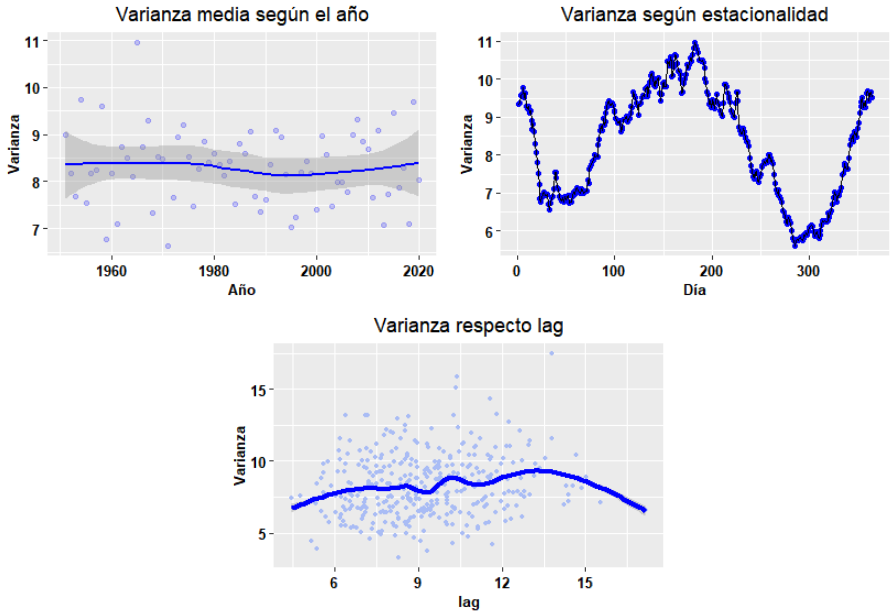

Figura A6: Gráficos para exploratorio de la varianza, Zaragoza.

A.B.3. Bilbao

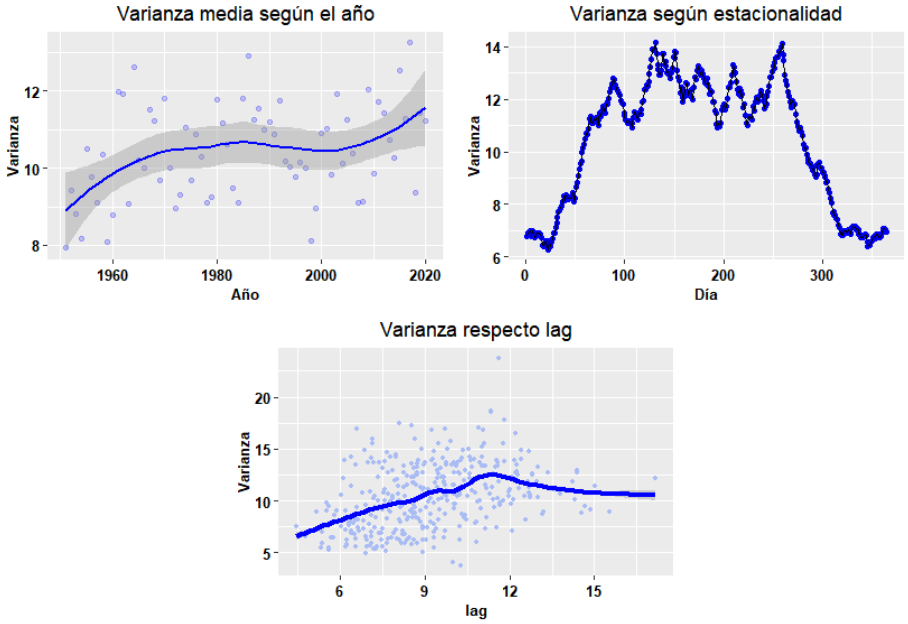

Figura A7: Gráficos para exploratorio de la varianza, Bilbao.

A.C. Resultados de los modelos exploratorios para la media y la varianza

μ	Soria	Bilbao	Zaragoza	σ	Soria	Zaragoza	Bilbao
$Year_t$	0.0073	0.0120	0.0120	$Year_t$			0.70
$Year_t : S1_t$		-0.0011	-0.0012	$Year_t : S1_t$			-2.0 e-4
$Year_t : C1_t$		-0.0011	-0.0012	$Year_t : C1_t$			1.9 e-3
$Year_t : S2_t$				$Year_t : S2_t$			-1.6 e-3
$Year_t : C2_t$				$Year_t : C2_t$			5.7 e-5
$Y_{t-1} : S1_t$	0.0160	0.0370	0.0380	$Y_{t-1} : S1_t$	0.018	2.7 e-4	-3.3e-3
$Y_{t-1} : C1_t$	0.0043	0.0098	0.1400	$Y_{t-1} : C1_t$	0.015	0.039	0.048
$Y_{t-1} : S2_t$		- 0.0061	- 0.0170	$Y_{t-1} : S2_t$	-8.3 e-3	5.0 e-3	-3.0 e-4
$Y_{t-1} : C2_t$		0.0096	- 0.0140	$Y_{t-1} : C2_t$	-2.4 e-3	0.015	-6.7e-3

Cuadro A1: Coeficientes con 2 cifras significativas de los parámetros más relevantes asociados a los Modelos (4.3) y (4.4) para la media y los Modelos (4.5) y (4.6) para la varianza.

A.D. Resultados de los modelos bayesianos-'bamlss'

A.D.1. Resumen numérico de la distribución a posteriori de los parámetros del modelo final

Modelo bayesiano final Soria							
μ	Mean	2.5%	97.5%	σ	Mean	2.5%	97.5%
$Year_t$	0.007	0.006	0.009				
$S1_t$	-1.179	-1.391	-0.971	$S1_t$	-0.0367	-0.0848	0.0163
$C1_t$	-2.664	-2.875	-2.455	$C1_t$	-0.0830	-0.1368	-0.0326
$S2_t$	0.567	0.489	0.649	$S2_t$	0.0729	0.0361	0.1053
$C2_t$	0.048	-0.031	0.128	$C2_t$	-0.0113	-0.0460	0.0253
Y_{t-1}	0.730	0.722	0.738	Y_{t-1}	0.0069	0.0049	0.0089
$Y_{t-1} : S1_t$	0.014	0.002	0.026	$Y_{t-1} : S1_t$	0.0061	0.0030	0.0088
$Y_{t-1} : C1_t$	0.002	-0.009	0.013	$Y_{t-1} : C1_t$	0.0038	0.0012	0.0067
				$Y_{t-1} : S2_t$	-0.0023	-0.0041	-0.0005
				$Y_{t-1} : C2_t$	-0.0022	-0.0040	-0.0004

Cuadro A2: Resumen numérico de la distribución a posteriori de los parámetros del modelo bayesiano final para μ y σ incluyendo media e intervalo de credibilidad.(Soria)

Modelo bayesiano final Zaragoza							
μ	Mean	2.5%	97.5%	σ	Mean	2.5%	97.5%
$Year_t$	0.012	0.010	0.014				
$Year_t : S1_t$	-0.001	-0.003	0.002				
$Year_t : C1_t$	-0.004	-0.007	-0.002				
$S1_t$	-0.037	-4.964	4.508	$S1_t$	0.1143	0.0506	0.1813
$C1_t$	5.264	0.133	10.393	$C1_t$	0.3617	0.3005	0.4227
$S2_t$	0.760	0.596	0.929	$S2_t$	-0.0895	-0.1302	-0.0487
$C2_t$	-0.337	-0.508	-0.158	$C2_t$	0.1263	0.0847	0.1707
Y_{t-1}	0.676	0.668	0.685	Y_{t-1}	0.0093	0.0072	0.0114
$Y_{t-1} : S1_t$	0.033	0.020	0.045	$Y_{t-1} : S1_t$	-0.0001	-0.0032	0.0030
$Y_{t-1} : C1_t$	0.006	-0.006	0.018	$Y_{t-1} : C1_t$	-0.0181	-0.0210	-0.0153
$Y_{t-1} : S2_t$	-0.004	-0.012	0.004	$Y_{t-1} : S2_t$	0.0018	0.0001	0.0037
$Y_{t-1} : C2_t$	0.010	0.003	0.018	$Y_{t-1} : C2_t$	-0.0073	-0.0091	-0.0056

Cuadro A3: Resumen numérico de la distribución a posteriori de los parámetros del modelo bayesiano final para μ y σ incluyendo media e intervalo de credibilidad.(Zaragoza)

Modelo bayesiano final Bilbao							
μ	Mean	2.5%	97.5%	σ	Mean	2.5%	97.5%
$Year_t$	0.010	0.008	0.012				
$Year_t : S1_t$	-0.001	-0.001	-0.001	$Year_t : S2_t$	0.0000	0.0000	0.0000
$Year_t : C1_t$	-0.002	-0.002	-0.002	$Year_t : C2_t$	-0.0001	-0.0001	-0.0001
$S2_t$	0.570	0.504	0.640	$S1_t$	0.1298	0.0716	0.1857
$C2_t$	0.026	-0.041	0.096	$C1_t$	0.5473	0.4860	0.6078
Y_{t-1}	0.647	0.637	0.657	Y_{t-1}	0.0339	0.0317	0.0362
$Y_{t-1} : S1_t$	0.038	0.025	0.051	$Y_{t-1} : S1_t$	0.0003	-0.0024	0.0032
$Y_{t-1} : C1_t$	0.096	0.082	0.111	$Y_{t-1} : C1_t$	-0.0263	-0.0292	-0.0232

Cuadro A4: Resumen numérico de la distribución a posteriori de los parámetros del modelo bayesiano final para μ y σ incluyendo media e intervalo de credibilidad.(Bilbao)

A.D.2. Gráficos de interacciones de parámetros

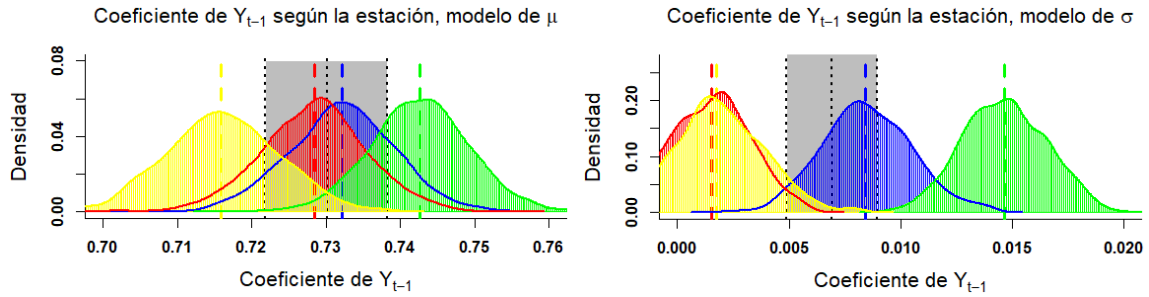

Figura A8: Efectos de Y_{t-1} según la estación del año en el modelo de Soria. A la izquierda pertenecientes al modelo para la media y a la derecha para la varianza. En gris se representa el intervalo de credibilidad y la media a posteriori de Y_{t-1}. En verde se representa el coeficiente de Y_{t-1} el 20 Abril, en amarillo el 23 Septiembre, en azul el 29 Diciembre y en rojo el 29 de Junio.

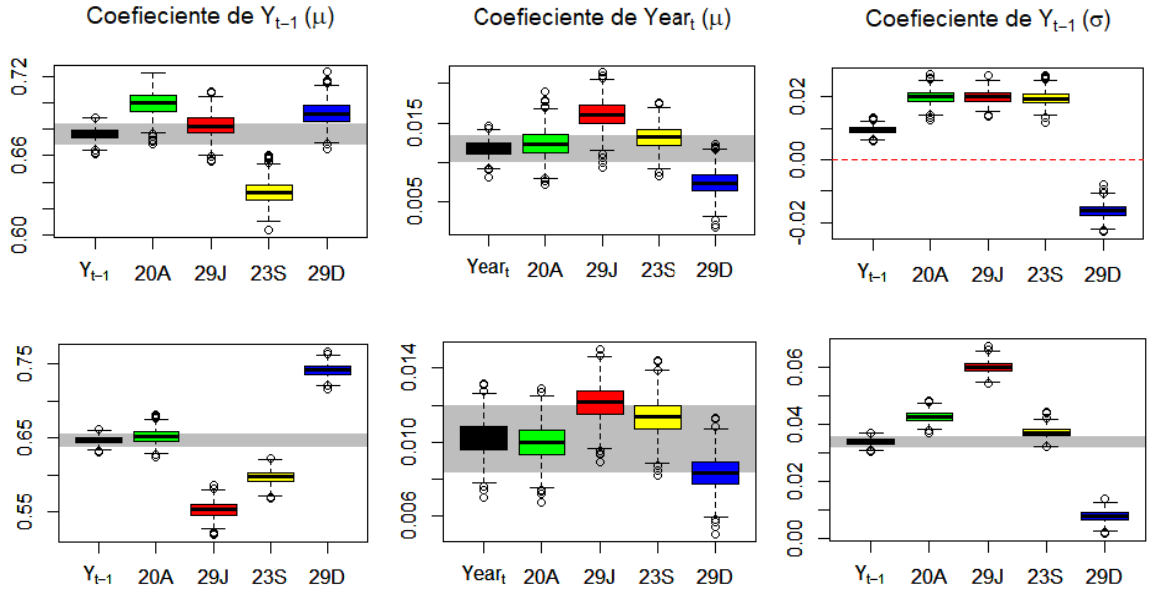

Figura A9: Box-plot de las distribuciones a posteriori de los parámetros con interacciones en Zaragoza (primera fila) y Bibao (segunda fila). En negro se representa el efecto sin interacción, en gris su intervalo de credibilidad, en verde el valor del efecto para el 20 Abril, en rojo para el 29 de Junio, en amarillo para el 23 Septiembre y en azul para el 29 Diciembre.

A continuación se mostrarán los datos tabulados de la distribución a posteriori asociada a los términos con interacciones en el MBF para cada observatorio.

Distribución a posteriori de los términos con interacciones, MBF Soria.							
Submodelo para μ	Media	2.5%	97.5 %	Submodelo para σ	Media	2.5%	97.5 %
Y_{t-1}	0.730	0.722	0.738	Y_{t-1}	0.007	0.005	0.009
20A	0.743	0.730	0.757	20A	0.002	-0.002	0.005
29J	0.729	0.714	0.743	29J	0.008	0.005	0.013
23S	0.716	0.701	0.731	23S	0.015	0.011	0.018
29D	0.732	0.718	0.746	29D	0.002	-0.002	0.006

Cuadro A5: Resumen de la distribución a posteriori de los términos con interacción, MBF de Soria. Primero consideramos el término sin interaccionar, luego en los días seleccionados 20A, 29J, 23S y 29D.

Distribución a posteriori de los términos con interacciones, MBF Zaragoza.							
Submodelo para μ	Media	2.5%	97.5 %	Submodelo para σ	Media	2.5%	97.5 %
Y_{t-1}	0.676	0.668	0.685	Y_{t-1}	0.009	0.007	0.011
20A	0.700	0.682	0.717	20A	0.020	0.016	0.024
29J	0.683	0.664	0.699	29J	0.020	0.016	0.024
23S	0.632	0.615	0.649	23S	0.019	0.015	0.024
29D	0.692	0.676	0.708	29D	-0.016	-0.020	0.012
$Year_t$	0.012	0.010	0.014				
20A	0.012	0.009	0.016				
29J	0.016	0.013	0.019				
23S	0.013	0.010	0.016				
29D	0.007	0.004	0.011				

Cuadro A6: Resumen de la distribución a posteriori de los términos con interacción, MBF de Zaragoza. Primero consideramos el término sin interaccionar, luego en los días seleccionados 20A, 29J, 23S y 29D.

Distribución a posteriori de los términos con interacciones, MBF Bilbao.							
Submodelo para μ	Media	2.5 %	97.5 %	Submodelo para σ	Media	2.5 %	97.5 %
Y_{t-1}	0.647	0.637	0.657	Y_{t-1}	0.034	0.032	0.036
20A	0.652	0.634	0.669	20A	0.043	0.039	0.046
29J	0.553	0.532	0.573	29J	0.060	0.056	0.064
23S	0.597	0.578	0.614	23S	0.037	0.033	0.041
29D	0.742	0.726	0.757	29D	0.008	0.004	0.011
$Year_t$	0.010	0.008	0.012				
20A	0.010	0.008	0.012				
29J	0.012	0.010	0.014				
23S	0.011	0.010	0.013				
29D	0.008	0.006	0.010				

Cuadro A7: Resumen de la distribución a posteriori de los términos con interacción, MBF de Bilbao. Primero consideramos el término sin interaccionar, luego en los días seleccionados 20A, 29J, 23S y 29D.

A.D.3. Gráfico de diferencia de temperaturas

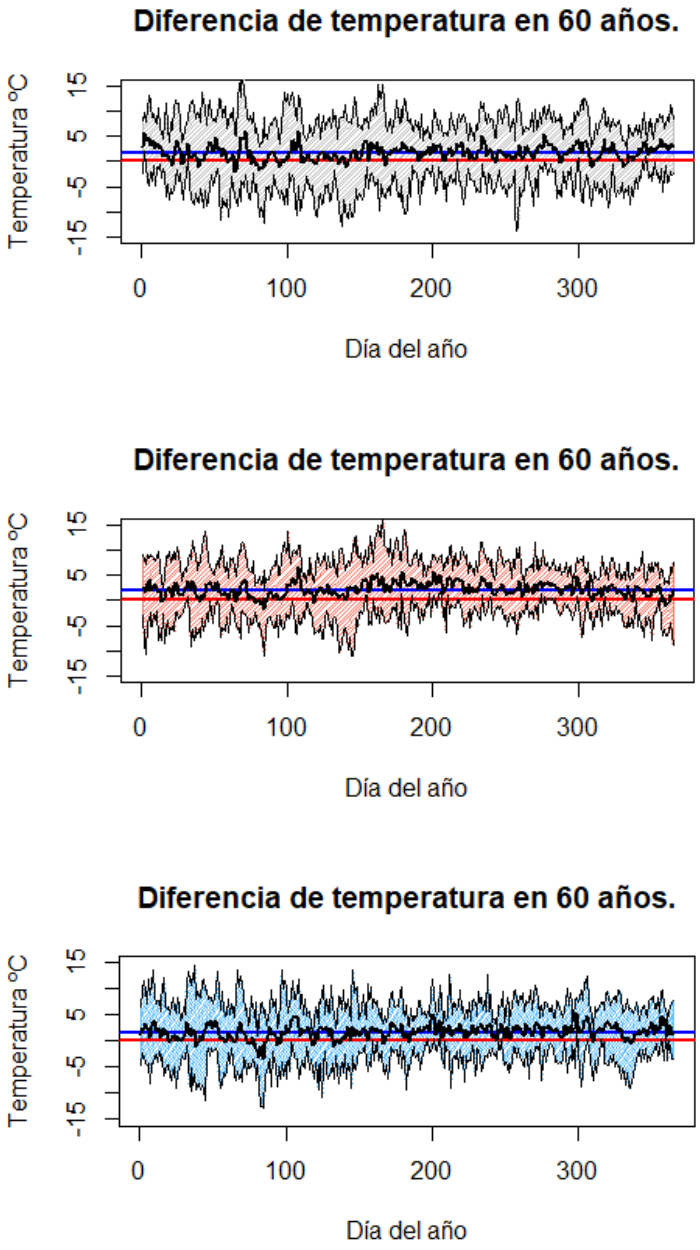

Figura A10: Intervalo de credibilidad (en gris Soria, en rojo claro Zaragoza y en azul claro Bilbao) con media a posteriori (en negro) para mostrar la diferencia de T.máx el mismo día del año 60 años después. En rojo la línea horizontal resalta el 0 y en azul oscuro el incremento medio de la T.máx. en todo el año.

A.E. Convergencia de los modelos bayesianos-'bamlss'

Se muestran los 'trace plot' de cada parámetro del modelo bayesiano MBF en cada localidad:

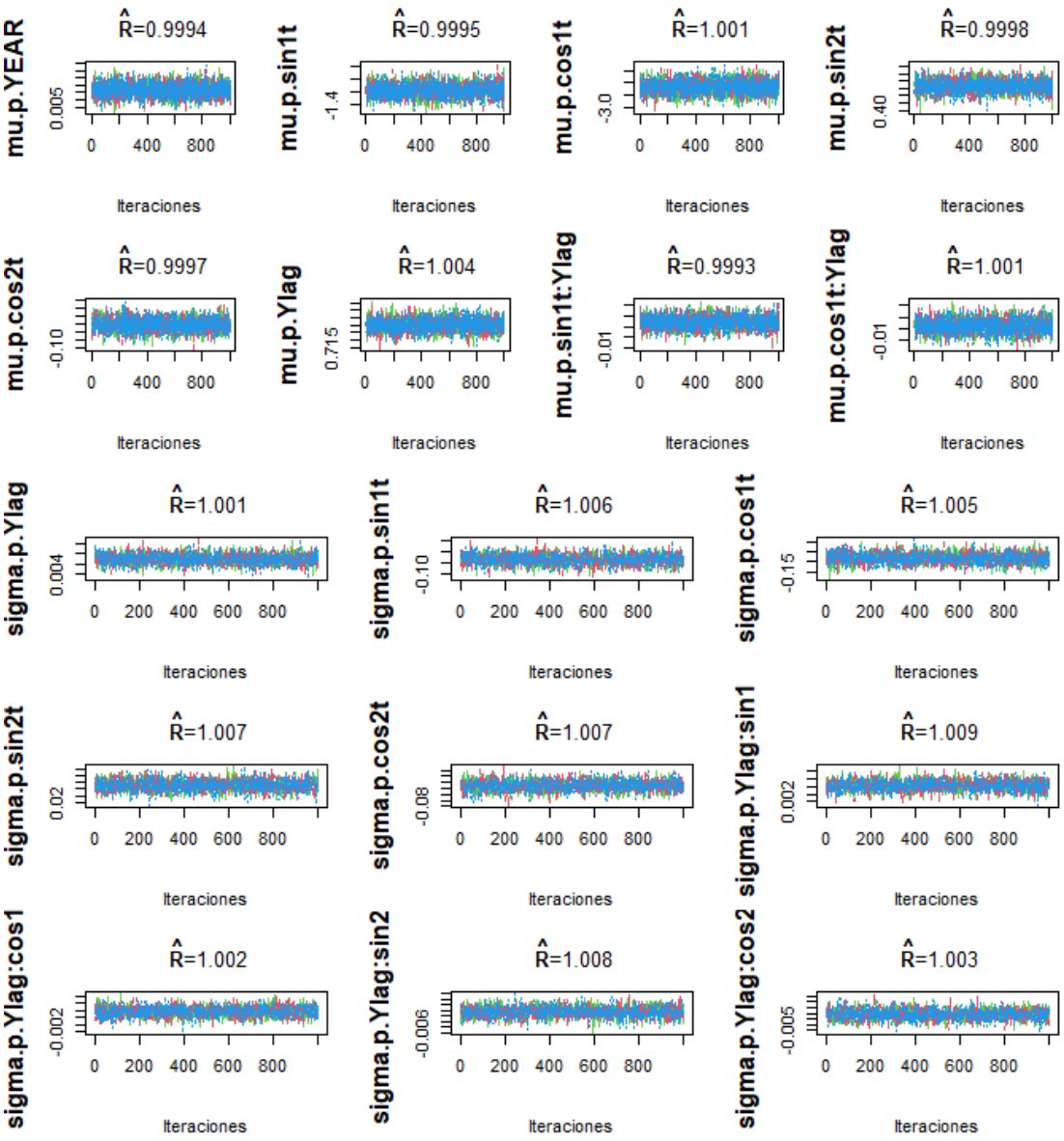

Figura A11: Trace plot asociado a cada parámetro del modelo bayesiano MBF de Soria con su correspondiente valor del diagnóstico de convergencia de Gelman-Rubin $\hat{R}$.

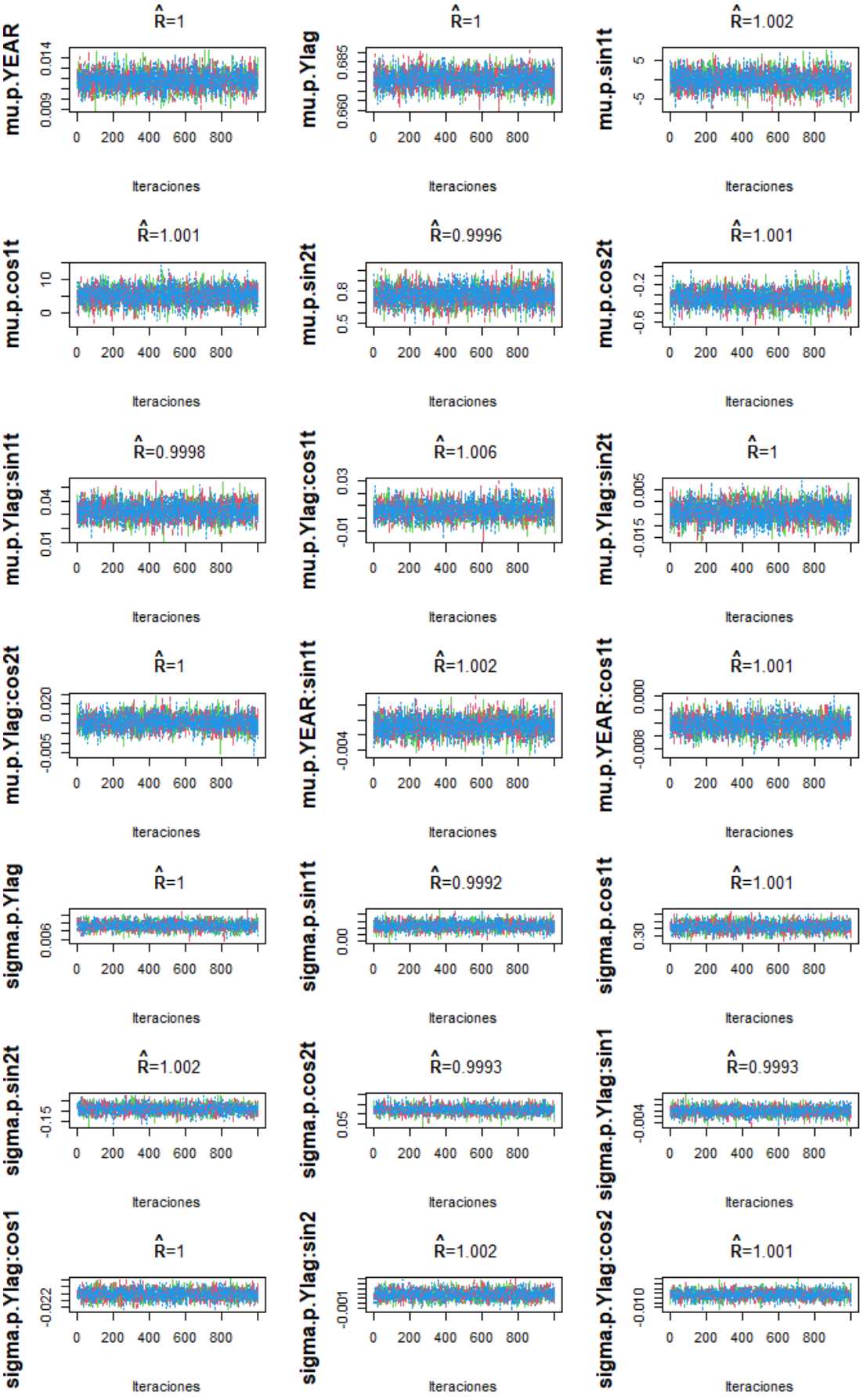

Figura A12: Trace plot asociado a cada parámetro del modelo bayesiano MBF de Zaragoza con su correspondiente valor del diagnóstico de convergencia de Gelman-Rubin $\hat{R}$.

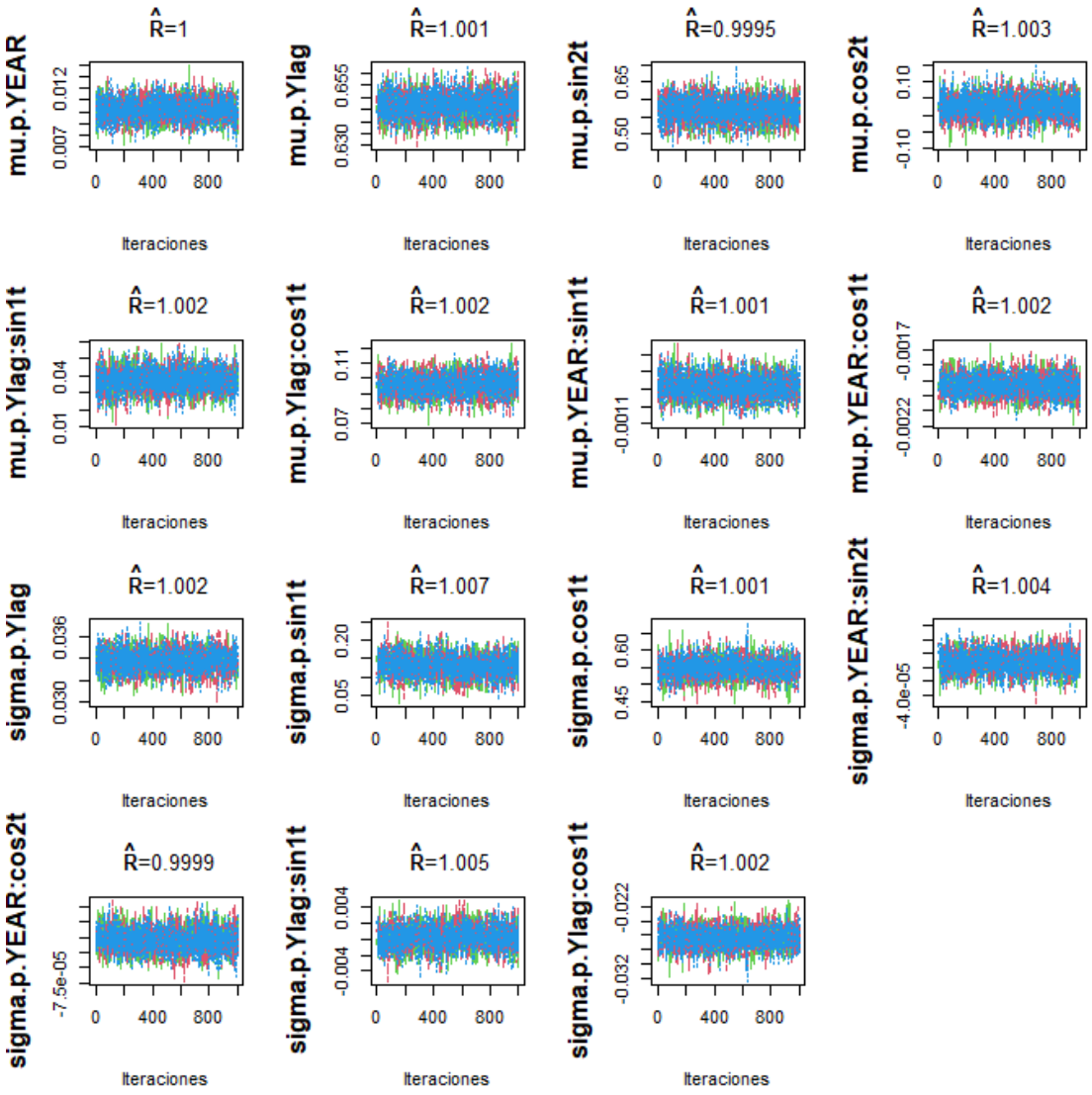

Figura A13: Trace plot asociado a cada parámetro del modelo bayesiano MBF de Bilbao con su correspondiente valor del diagnóstico de convergencia de Gelman-Rubin $\hat{R}$.

A continuación se mostrará una tabla conteniendo el tamaño de muestra efectivo para cada parámetro del MBF correspondiente a cada observatorio. Dicha medida debería ser cercana a 1001 para cada parámetro, el número de sorteos a posteriori pedidos.

Tamaño de muestra efectivo de cada parámetro							
Submodelo μ	Soria	Zaragoza	Bilbao	Submodelo σ	Soria	Zaragoza	Bilbao
$S1_t$	1001	1001		$S1_t$	664	1209	690
$C1_t$	1001	1001		$C1_t$	910	1001	763
$S2_t$	1001	1001	1001	$S2_t$	833	1001	715
$C2_t$	1001	1001	1001	$C2_t$	884	728	841
Y_{t-1}	1001	1001	1001	Y_{t-1}	701	1442	787
$Y_{t-1} : S1_t$	1001	1132	1001	$Y_{t-1} : S1_t$	866	1291	698
$Y_{t-1} : C1_t$	1001	1147	1001	$Y_{t-1} : C1_t$	806	1001	829
$Y_{t-1} : S2_t$		1001		$Y_{t-1} : S2_t$	764	1001	
$Y_{t-1} : C2_t$		1001		$Y_{t-1} : C2_t$	1001	913	
$Year_t$	863	1001	1001	$Year_t$			
$Year_t : S1_t$		1001	1001	$Year_t : S2_t$			715
$Year_t : C1_t$		1001	1001	$Year_t : C2_t$			841

Cuadro A8: Tamaño de muestra efectivo para los parámetros del MBF en Soria, Zaragoza y Bilbao.

A.F. Código de R para algunas funciones

El código de R desarollado para este TFG se puede encontrar en el enlace de GitHub: `https://github.com/JavierTorcal/TFG-Bayesian-models-for-climate-time-series`.

A.F.1. Función para depurar el conjunto de datos

Este archivo recoge la función utilizada para preparar los conjuntos de datos de acuerdo a nuestras necesidades. Se aplica sobre la serie de T.máx. diaria de la ciudad deseada y requiere la instalación de la librería **'lubridate'** para el manejo de las fechas.

```
fun_depurar_datos<-function(datos){

require(lubridate)
#Eliminar columnas que no queremos.
datos$SOUID<-NULL
datos$STAID<-NULL
datos$Q_TX<-NULL

#Formato de fechas mediante Lubridate
datos$DATE<-ymd(datos$DATE)
datos$YEAR<-year(datos$DATE)
datos$MONTH<-month(datos$DATE)
datos$DAY<-day(datos$DATE)


#Eliminar las observaciones -9999 y dividir entre 10:
datos[datos$TX==-9999,]$TX<-NA
datos$TX<-datos$TX/10

#Eliminar 29 Feb
diabisiesto<-is.element(datos$MONTH,2)&is.element(datos$DAY,29)
datos<-datos[!diabisiesto,]

#Años 1951-2020
datos<-datos[(datos$YEAR>=1951) & (datos$YEAR<=2020),]

#Incluir los armónicos en el conjunto de datos
n <- length(datos$TX)
t <- 1:n
datos$sin1t <- sin(2 * pi * t / 365)
datos$cos1t <- cos(2 * pi * t / 365)
datos$sin2t <- sin(4 * pi * t / 365)
datos$cos2t <- cos(4 * pi * t / 365)

#Añadir el día previo (Ylag)
datos$Ylag<-lag(c(NA,datos$TX[-length(datos$TX)]),k=1)

#Poner el número adecuado a los días
rownames(datos)<-c(1:25550)
datos$YEAR_DAY<-rep(1:365,70)
```

```
#Devuelve el conjunto de datos listo para su uso.
return(datos)}
```

A.F.2. Función usada para inferir resultados respecto al calentamiento

Esta función nos ha permitido obtener los resultados sobre la explotación del modelo MBF para el estudio del calentamiento global. Se aplica sobre el conjunto de datos y el modelo bayesiano final de la ciudad deseada.

```
fun_estudio.dif.temp<-function(datos,modelo,color.grafico){
#Conseguir simulaciones de temperatura máxima en década1
decada1<-datos[366:3650,]#1952-1960
p1 <- predict(modelo,newdata=decada1, model = "mu", What="samples",
              na.rm=T, FUN = function(x) { x })
matriz1<-as.matrix(p1[,1:1000])

#Conseguir simulaciones de temperatura máxima en década7
decada7<-datos[22266:25550,]#2012-2020
p7 <- predict(modelo,newdata=decada7, model = "mu", What="samples",
              FUN = function(x) { x })
matriz7<-as.matrix(p7[,1:1000])

matrizResta<-matriz7-matriz1
medias<-rowMeans(matrizResta)

#Obtener distrib posteriori de diferencia media temp diaria entre dos décadas
aux.mean<-aggregate(matrizResta,by=list(rep(c(1:365),9)),mean)
aux.mean.mean<-apply(aux.mean,1,mean)

#Conseguir el INtervalo de Credibilidad.
aux.matrix<-matrix(0,nrow=365,ncol=9000)

for(i in 1:365){
        aux.matrix[i,]<- c(matrizResta[seq(i,i+8*365,365),])}

c.025<-apply(aux.matrix,1,quantile,0.025)
c.975<-apply(aux.matrix,1,quantile,0.975)

#Graficar la media a posteriori e intervalo de credibilidad a posteriori
Resultado<-NULL
Resultado$YEAR_DAY<-c(1:365)
Resultado$T.media<-aux.mean.mean
Resultado$c2.5<-c.025
Resultado$c97.5<-c.975

Resultado<-as.data.frame(Resultado)
plot(T.media~YEAR_DAY,data=Resultado,type="n",ylim=c(-15,15),lwd=2,
     xlab="Día del año",ylab="Temperatura ºC",
     main="Diferencia de temperatura en 60 años.")
polygon(c(Resultado$YEAR_DAY, rev(Resultado$YEAR_DAY)), c(Resultado$c2.5,
rev(Resultado$c97.5)),col = color.grafico , density = 50, angle = 45)
```

```
abline(h=0,col="red",lwd=2)
abline(h=mean(medias),col="blue",lwd=2)
lines(loess(T.media~YEAR_DAY,data=Resultado),type="l",ylim=c(-10,10),lwd=2)
lines(c2.5~YEAR_DAY,data=Resultado,type="l",ylim=c(-10,10),lwd=1)
lines(c97.5~YEAR_DAY,data=Resultado,type="l",ylim=c(-10,10),lwd=1)

#Incremento en media de la temperatura máxima diaria
print(paste("Incremento medio de",mean(medias), "grados"))

#Probabilidad a posteriori de que la temp. máx.
#en un día sea mayor que el mismo día hace 60 años.
Resultado2<-aux.matrix
print(paste("Prob a posteriori de  mayor T.máx.:",mean(Resultado2>0)))}
```