PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Un enfoque bayesiano para estimar las temperaturas mínimas extremas a través de un modelo geoestadístico GEV

TESIS PARA OPTAR POR EL GRADO ACADÉMICO DE MAGISTRA EN ESTADÍSTICA

AUTORA

Anilda Maribel Guevara Alvarado

ASESORA

Dra. Zaida Jesús Quiroz Cornejo

Diciembre, 2022

Declaración jurada de autenticidad

Yo Zaida Jesús Quiroz Cornejo docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesora de la tesis titulada *Un enfoque bayesiano para estimar las temperaturas mínimas extremas a través de un modelo geoestadístico GEV*, de la autora Anilda Maribel Guevara Alvarado, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 16 %. Así lo consigna el reporte de similitud emitido por el software Turnitin el 13/08/2020.
- He revisado con detalle dicho reporte y confirmo que cada una de las coincidencias detectadas no constituyen plagio alguno.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lima, 01 de diciembre de 2022

<table>
<tr><td colspan="2">Apellidos y nombres de la asesora:
Quiroz Cornejo Zaida Jesús</td></tr>
<tr><td>DNI: 43704124</td><td rowspan="2">Firma:</td></tr>
<tr><td>ORCID: https://orcid.org/0000-0003-3821-0815</td></tr>
</table>

Dedicatoria

A mis queridos padres, por enseñarme que con perserverancia y estudio podemos lograr nuestros objetivos.
A mi jefe y amigo, Juan Carlos Torres, por apoyarme y motivarme a estudiar la maestría.

Agradecimentos

Mi especial agradecimiento para la profesora Zaida Quiroz, por su orientación, motivación, contribución y guia constante durante el desarrollo de la tesis.

A los profesores de la maestría, por su guía y enseñanza durante todos estos años estudio.

A mis amigos de la maestría, Anita y Francisco, por su compañerismo y apoyo.

Resumen

El desarrollo sostenible de un país puede verse limitado debido a cambios graduales del clima y eventos hidrometeorológicos extremos, que afectan de manera recurrente la infraestructura, medios de vida así como las inversiones. El Perú, es uno de los países más afectados por la variabilidad y cambio climático, por tanto la gestión del riesgo climático, entre ellas el estudio de temperaturas extremas, contribuye a reducir impactos socio-económicos y ambientales en las inversiones público-privadas. En este contexto, en esta tesis se propone aplicar un modelo bayesiano geoestadístico usando una distribución generalizada para valores extremos (GEV) para estimar y predecir las temperaturas minimas extremas en el Perú en el 2012. Así mismo, dado el alto costo computacional que ameritan los modelos bayesianos espaciales, se propone usar el enfoque de ecuaciones diferenciales parciales estocásticas (SPDE) y para la estimación de los parámetros se usa el método integrado de aproximación anidada de Laplace (INLA). El modelo propuesto permite estimar las temperaturas mínimas extremas en el Perú, con el propósito de mejorar la gestión de riesgo climático.

Palabras-clave: distribución generalizada para valores exremos (GEV), inferencia bayesiana, INLA, geoestadística, riesgo climático, temperaturas mínimas extremas.

Abstract

Sustainable development of a country could be affected by gradual changes in climate and extreme hydrometeorological hazards, which frequently damage infrastructure, livelihoods and investments. Peru is one of the most affected countries by climate variability and change, therefore implementing climate risk management policies which include the study of extreme temperatures, contributes to reducing socio-economic and environmental impacts on public and private investments. In this context, in this thesis it is proposed to implement a bayesian geostatistical model with a generalized extreme value distribution (GEV) to estimate and predict extreme minimum temperatures in Peru during 2012. Given the high computational complexity for implementing bayesian spatial models, we proposed use the stochastic partial differential equations (SPDE) and for the estimation of the parameters use the method, approach combined with the Integrated Nested Laplace Approximations (INLA). The proposed model allows estimating the minimum temperatures in Peru, with the purpose of improving climate risk management.

Keywords: bayesian inference, climate risk, extreme minimum temperatures, generalized extreme value distribution (GEV), geostatistics, INLA .

Índice general

Lista de Abreviaturas

$\mathbf{y}_{-i}$	: vector de observaciones, sin tomar en cuenta valor observado y_i .
GEV	: Distribución generalizada de valores extremos.
INLA	: Integrated nested laplace approximations.
MCMC	: Markov chain montecarlo.
SPDE	: Spatial stochastic differential equation.

Índice de figuras

Capítulo 1

Introducción

De acuerdo al estudio técnico de CENEPRED (2018) la temporada de las bajas temperaturas en el Perú, se caracteriza en las zonas alto andinas por la presencia de heladas, y algunas veces por la ocurrencia de nevadas y granizadas; mientras que en la selva, se tiene la presencia de friajes, que son caracterizadas por aire frío polar provenientes del sur, las temperaturas del aire disminuyen, hay mayor presencia de neblinas durante la noche y muy temprano. Debido a descensos críticos de la temperatura que sufre el Perú en sus distintas zonas geográficas (ver figura 1.1) y en ciertos periodos de tiempo, se propone: i) proponer un modelo estadístico que se adapte a las características de los datos, como dependencia espacial y temperaturas mínimas extremas; ii) proveer un análisis del comportamiento de la temperatura mínima en las diferentes regiones del Perú y evaluar qué variables pueden influir en ella y iii) estimar y predecir las temperaturas mínimas diarias extremas, con el fin de gestionar riesgos de desastres, impactos climatológicos, hidrológicos o agricultura.

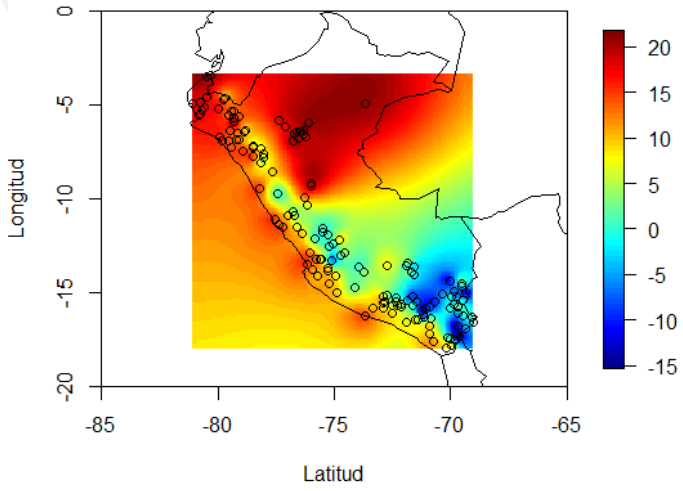

Figura 1.1: Temperaturas mínimas promedio en el Perú registradas en Agosto del 2012. Las estaciones metereológicas están representadas por círculos

1.1. Revisión de Literatura

En el estudio de Dyrrdal et al. (2014) se propuso un modelo geoestadístico, vinculando las distribuciones de valores extremos generalizados (GEV) con campos gaussianos latentes en un modelo jerárquico bayesiano que permita estimar la precipitación horaria extrema en Noruega. Los modelos lineales generalizados sobre los parámetros de la distribución GEV pueden incorporar información geográfica y meteorológica específica de la ubicación y por lo tanto, acomodar estos efectos en las precipitaciones extremas. Los parámetros de este modelo jerárquico bayesiano propuesto se estimaron a través de cadenas de Markov de Montecarlo (MCMC).

Por otro lado, en el trabajo de Rachmawati et al. (2018) proponen un modelo bayesiano espacio temporal y usan el método de aproximación de Laplace anidada integrada (INLA) dado el alto costo computacional de la estimación por MCMC. Así, la investigación modela el número de personas pobres ajustada por la distribución generalizada de valores extremos utilizando modelos espaciales y espacio temporales incorporando efectos aleatorios de tendencias temporales paramétricas y no paramétricas.

1.2. Objetivos

Debido a la distribución de las temperaturas mínimas medidas a través de las estaciones metereológicas del Servicio Nacional de Meteorología e Hidrología del Perú-SENAHMI, en este trabajo se propone modelar estas temperaturas usando una distribución generalizada de valores extremos (GEV) incorporando un efecto aleatorio espacial que tome en cuenta la dependencia en los datos (Figura 1.1). La estimación de los parámetros se realizará bajo el enfoque bayesiana mediante el método de aproximación de Laplace anidada integrada (INLA), dada su eficiencia computacional.

El objetivo principal de la tesis es aplicar un modelo geoestadístico con distribución generalizada de valores extremos (GEV) para estimar y predecir las temperaturas mínimas extremas en el Perú. De manera específica:

- Revisar la literatura acerca de los diferentes modelos espaciales propuestos para variables continuas de valores extremos.
- Aplicar el modelo espacial generalizado para valores extremos al conjunto de datos que contienen las temperaturas mínimas extremas medidas en las estaciones meteorológicas del Servicio Nacional de Meteorología e Hidrología del Perú-SENAHMI.

1.3. Organización del trabajo

La presente tesis se organiza de la siguiente manera. En el Capítulo 2, se presentan los conceptos previos y necesarios para el desarrollo del trabajo, como la definición de la distribución generalizada de valores extremos y sus reparametrizaciones, conceptos de geoestadística, el modelo Matérn para modelar la dependencia espacial y su aproximación por ecuaciones diferenciales parciales estocásticas, inferencia bayesiana usando INLA y criterios de selección de modelos. En el Capítulo 3, se presenta la estructura del modelo geoestadístico y la estimación de los parámetros bajo el enfoque bayesiano. El Capítulo 4 muestra los resultados del

estudio de simulación realizado. El Capítulo 5 muestra el análisis exploratorio de los datos de las temperaturas mínimas medidas en las estaciones y que fueron extraídas de SENAHMI y la aplicación del modelo. Finalmente, en el Capítulo 6, se discuten las conclusiones obtenidas del trabajo.

Capítulo 2

Marco teórico

En este capítulo se detallará la definición y propiedades de la distribución que será de uso para la estructura del modelo del capítulo posterior. Se introducirán además ciertos conceptos geoestadísticos y de inferencia bayesiana.

2.1. Teoría de valores extremos

La teoría de valores extremos (*Theory Extreme Values* - TEV) tiene como objetivo modelar eventos raros, que ocurren con una probabilidad muy pequeña, a fin de predecir dichos eventos o medir su riesgo. Es muy usado en diferentes disciplinas como en cambios climáticos, geología, finanzas, biomedicina, entre otras.

Así mismo, la TEV tiene como objetivo extrapolar información y se enfoca en el estudio de las colas de la distribución de los datos. En particular las propiedades de los valores extremos (como por ejemplo mínimos y máximos) son determinadas por las colas a la izquierda o derecha de la distribución subyacente (Kotz y Nadarajah (2000)). En la práctica, la TEV nos permite modelar distribuciones donde las colas tienen pocas observaciones y usarlas para estimar valores extremos que pueden ser menores que el valor mínimo o mayores que el valor máximo.

De acuerdo a Coles (2001), la teoría de valores extremos está enfocada en describir el comportamiento de la estadística máximo, denotada por $M_n = max\{X_1, ..., X_n\}$, donde $X_1, ..., X_n$ son variables aleatorias i.i.d con función de distribución acumulada F. La distribución del M_n puede obtenerse a partir de la distribución de las n variables aleatorias $X_1, ..., X_n$. Para ello se toma en cuenta que dichas v.a.s son independientes, entonces la función de distribución acumulada (f.d.a) de M_n es:

$$\begin{aligned} F_{M_n}(z) &= P\left[M_n \leq z\right] = P\left[X_1 \leq z, ..., X_n \leq z\right] \\ &= P\left[X_1 \leq z\right] \times ... \times P\left[X_n \leq z\right] \\ &= \left[F(z)\right]^n . \end{aligned} \tag{2.1}$$

La distribución calculada en (2.1) converge a cero cuando $n \to \infty$ para $z < z^*$ y a uno para $z > z^*$, donde $z^* = \sup\{z : F(z) < 1\}$. Por lo que, para obtener una distribución límite será necesario definir una sucesión de vectores aleatorios M_n^* en función de los M_n y en base a la teoría del límite central, la cual consiste en determinar sucesiones de constantes $\{b_n\}_{n\geq 1}$ y

$\{a_n\}_{n\geq 1}$, tales que

$$M_n^* = \frac{M_n - b_n}{a_n} \quad , \text{ con } a_n > 0.$$

Las posibles distribuciones límite que podrían surgir para M_n^* vendrán dadas por el siguiente teorema propuesto por Fisher y Tippett (1928) y demostrado por Gnedenko (1943)

Teorema 2.1.1 *Si existen sucesiones de constantes* $\{a_n > 0\}$ *y* $\{b_n\}$ *tales que*

$$P\left(\frac{M_n - b_n}{a_n} \geq z\right) \to G(z), \text{ cuando } n \to \infty,$$

entonces, la función de distribución acumulada de G *debe pertenecer a una de las siguientes familias:*

$$I: \quad G(z) = \exp\left\{-\exp\left[-\frac{z-b}{a}\right]\right\}, -\infty < z < \infty;$$

$$II: \quad G(z) = \begin{cases} 0, & z \leq b, \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b; \end{cases}$$

$$III: \quad G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\}, & z < b, \\ 1, & z \geq b, \end{cases}$$

con $a > 0$, $b \in \mathbb{R}$ *y* $\alpha > 0$.

De acuerdo a Coles (2001), el teorema 2.1.1 establece que el máximo de la muestra reescalado $(M_n - b_n)/a_n$ converge en distribución a una variable que tiene una distribución dentro de las familias etiquetadas como I, II y III. Colectivamente, estas tres distribuciones se denominan distribuciones de valor extremo, con los tipos I, II y III ampliamente conocidos como las familias Gumbel, Frechet y Weibull, respectivamente. Cada familia tiene un parámetro de localización y escala, b y a respectivamente; además, las familias Frechet y Weibull tienen un parámetro de forma α.

Así mismo, el teorema 2.1.1 implica que, cuando M_n puede ser estabilizada con una secuencia adecuada $\{a_n\}$ y $\{b_n\}$, la correspondiente variable normalizada M_n^* tiene una distribución que se restringe y corresponde a uno de los tres tipos de distribuciones de valores extremos. La característica más relevante de este resultado es que los tres tipos de distribuciones de valores extremos son los únicos límites posibles para la distribución de M_n^* sin importar la distribución F para la población. Es en este sentido, el teorema provee un valor extremo, de forma análoga al teorema de límite central.

2.2. Distribución generalizada de valores extremos (GEV) para el máximo

Von Mises (1936) y Jenkinson (1955) proponen una parametrización común de los tres tipos de distribuciones del Teorema 2.1.1, determinando una distribución generalizada de valores extremos (GEV *Generalized Extreme Values Distribution*) para la v.a Y y con f.d.a:

$$F(y,\xi) = \begin{cases} \exp\left\{-(1+\xi y)^{-1/\xi}\right\} & ;\xi \neq 0,\ 1+\xi y > 0 \\ \exp\left\{-\exp(-y)\right\} & ;\xi = 0,\ y \in \mathbb{R}, \end{cases}$$

donde ξ es un parámetro de forma. Incluyendo parámetros de localización μ y escala σ, su versión estandarizada es:

$$F(y,\mu,\sigma,\xi) = \begin{cases} \exp\left\{-\left(1+\xi\left(\frac{y-\mu}{\sigma}\right)\right)^{-1/\xi}\right\} & ;\xi \neq 0,\ 1+\xi\left(\frac{y-\mu}{\sigma}\right) > 0 \\ \exp\left\{-\exp\left(-\left(\frac{y-\mu}{\sigma}\right)\right)\right\} & ;\xi = 0,\ y \in \mathbb{R}. \end{cases} \tag{2.2}$$

Cabe recalcar que para $\xi = 0$, la aproximación $\xi \to 0$ permite mantener la continuidad de la f.d.a. Esto es

$$\begin{aligned} F(y,0) = \lim_{\xi\to 0} \exp\left\{-(1+\xi y)^{-1/\xi}\right\} &= \exp\left\{-\lim_{\xi\to 0}(1+\xi y)^{-1/\xi}\right\} \\ &= \exp\left\{-\exp(-y)\right\}. \end{aligned}$$

La función de densidad de probabilidad para Y con f.d.a (2.2), está dada por

$$f(y,\mu,\sigma,\xi) = \begin{cases} \frac{1}{\sigma}\left(1+\xi\left(\frac{y-\mu}{\sigma}\right)\right)^{-1-1/\xi} \exp\left\{-\left(1+\xi\left(\frac{y-\mu}{\sigma}\right)\right)^{-1/\xi}\right\} & ;\xi \neq 0 \\ \frac{1}{\sigma}\exp\left(-\left(\frac{y-\mu}{\sigma}\right)\right)\exp\left\{-\exp\left(-\left(\frac{y-\mu}{\sigma}\right)\right)\right\} & ;\xi = 0. \end{cases}$$

En la Figura 2.1 se visualiza la función de densidad de probabilidad (f.d.p) de la GEV descritas en el teorema 2.1.1. Para la simulación de los tres tipos distribuciones de valores extremos se utilizaron como parámetros $\mu = 0$, $\sigma = 1$ y $\xi = -0.8$. El valor esperado y la varianza de una variable aleatoria Y con distribución GEV estandarizada están dados por:

$$E[Y] = \begin{cases} \mu + (\Gamma(1-\xi)-1)\frac{\sigma}{\xi} & ,\xi \in (0,1), \\ \mu + \sigma e & ,\xi = 0, \\ \infty & ,\xi \in [1,\infty) \end{cases}$$

$$Var[Y] = \begin{cases} \left(\Gamma(1-2\xi) - \Gamma(1-\xi)^2\right)\frac{\sigma^2}{\xi^2} & ,\xi \in (0,1/2), \\ \sigma^2\frac{\pi^2}{6} & ,\xi = 0, \\ \infty & ,\xi \in [1/2,\infty) \end{cases}$$

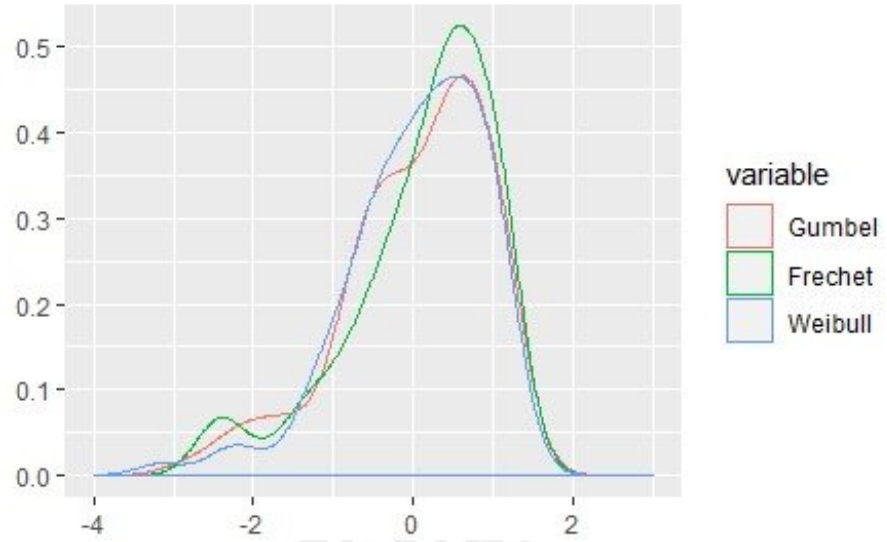

Figura 2.1: Funciones de densidad de v.a's con distribucion: Gumbel (linea roja), Fechet (linea verde) y Weibull (línea azul). Los parámetros utilizados para la simulación de las funciones de densidad son $\xi = -0.8$, $\mu = 0$ y $\sigma = 1$.

2.2.1. Reparametrización de la distribución GEV

La función de distribución acumulada de la GEV estandarizada (2.2) bajo la adición de un parámetro de escala fijo $s > 0$ y parámetros reestructurados de los vistos anteriormente, puede ser parametrizada bajo la siguiente función de distribución acumulada

$$F(y;\eta,\tau,\xi) \quad = \quad \begin{cases} \exp\left\{-\left(1+\xi\sqrt{\tau s}\,(y-\eta)\right)^{-1/\xi}\right\} & ;\xi \neq 0,\ 1+\xi\sqrt{\tau s}\,(y-\eta) > 0 \\ \exp\left\{-\exp\left(-\sqrt{\tau s}\,(y-\eta)\right)\right\} & ;\xi = 0,\ y \in \mathbb{R}, \end{cases} \tag{2.3}$$

donde se definen los nuevos parámetros:

- s: parámetro de escala fijo donde $s > 0$,
- τ: parámetro de precisión donde $\sqrt{\tau s} = \dfrac{1}{\sigma}$,
- η: predictor lineal, cuya representación matemática se basa bajo la reestructuración del parámetro μ de la forma siguiente

$$\eta \quad = \quad \mu.$$

La función de densidad de probabilidad para Y con la función de distribución acumulada (3.3), está dada por

$$f(y;\eta,\tau,\xi) \quad = \quad \begin{cases} \sqrt{\tau s}\left(1+\xi\sqrt{\tau s}\,(y-\eta)\right)^{-1-1/\xi}\exp\left\{-\left(1+\xi\sqrt{\tau s}\,(y-\eta)\right)^{-1/\xi}\right\} & ;\xi \neq 0 \\ \sqrt{\tau s}\exp\left(-\sqrt{\tau s}\,(y-\eta)\right)\exp\left\{-\exp\left(-\sqrt{\tau s}\,(y-\eta)\right)\right\} & ;\xi = 0. \end{cases} \tag{2.4}$$

Bajo esta reparametrización la distribución para valores extremos máximos, se denota

como

$$Y \sim GEV(\eta, \tau, \xi).$$

2.2.2. Distribución GEV para el mínimo

Si bien la distribución GEV se determina sobre valores extremos máximos, ésta puede ser reconstruida de tal forma que se pueda obtener una distribución generalizada para valores extremos mínimos que permita definir la siguiente estadística $\tilde{M}_n = min\{Y_1, \cdots, Y_n\}$ donde $Y_i = -X_i$, siendo X_i variables aleatorias de $M_n = max\{X_1, \cdots, X_n\}$, permitiendo que $\tilde{M}_n = -M_n$. El cálculo de la función de distribución acumulada para el proceso $\tilde{M}_n$ se puede determinar de la siguiente forma

$$\begin{aligned}
P(\tilde{M}_n \leq y) &= P(-M_n \leq y) \\
&= P(M_n \geq -y) \\
&= 1 - P(M_n \leq -y) \\
&\approx 1 - F(-y; \mu, \sigma, \xi) \\
&= \begin{cases} 1 - \exp\left\{-\left(1 + \xi\left(\frac{-y-\mu}{\sigma}\right)\right)^{-1/\xi}\right\} & ;\xi \neq 0,\ 1 + \xi\left(\frac{-y-\mu}{\sigma}\right) > 0 \\ 1 - \exp\left\{-\exp\left(-\left(\frac{-y-\mu}{\sigma}\right)\right)\right\} & ;\xi = 0,\ -y \in \mathbb{R} \end{cases} \\
&= \begin{cases} 1 - \exp\left\{-\left(1 - \xi\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right)^{-1/\xi}\right\} & ;\xi \neq 0,\ 1 - \xi\left(\frac{y-\tilde{\mu}}{\sigma}\right) > 0 \\ 1 - \exp\left\{-\exp\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right\} & ;\xi = 0,\ y \in \mathbb{R} \end{cases}
\end{aligned} \tag{2.5}$$

donde $\tilde{\mu} = -\mu$.

Dado que $X_i = -Y_i$ entonces:

$$\begin{aligned}
max(X_i, \cdots, X_n) &= max(-Y_i, \cdots, -Y_n) \\
&= min(Y_i, \cdots, Y_n).
\end{aligned}$$

Indicando que un modelo GEV bajo un proceso de valores máximos, definido en la ecuación (2.2), es equivalente a un modelo GEV de valores mínimos definido en la ecuación (2.5); siendo su función de densidad de probabilidad para valores mínimos de la siguiente forma:

$$f(y; \tilde{\mu}, \sigma, \xi) = \begin{cases} \frac{1}{\sigma}\left(1 - \xi\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right)^{-1-1/\xi} \exp\left\{-\left(-1 - \xi\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right)^{-1/\xi}\right\} & ;\xi \neq 0 \\ \frac{1}{\sigma}\exp\left(\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right)\exp\left\{-\exp\left(\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right)\right\} & ;\xi = 0 \end{cases}$$

En el anexo A se muestra cómo se usa el método de la transformada inversa para simular datos provenientes de la distribución GEV para valores extremos mínimos. Para más detalles de distribución generalizada para valores extremos mínimos ver Coles (2001).

2.2.3. Reparametrización de la distribución GEV para mínimos

Para la función de distribución acumulada (f.d.a) de la GEV para mínimos puede ser parametrizada de forma análoga a lo estructurado en (3.3) bajo la siguiente función de distribución acumulada

$$F(y;\tilde{\eta},\tau,\xi) \quad = \quad \begin{cases} 1-\exp\left\{-\left(1-\xi\sqrt{\tau s}(y-\tilde{\eta})\right)^{-1/\xi}\right\} & ;\xi\neq 0,\ 1-\xi\sqrt{\tau s}(y-\tilde{\eta})>0 \\ 1-\exp\left\{-\exp\left(\sqrt{\tau s}(y-\tilde{\eta})\right)\right\} & ;\xi=0,\ y\in\mathbb{R}, \end{cases} \tag{2.6}$$

donde se definen los nuevos parámetros

- s: parámetro de escala fijo donde $s>0$,
- τ: parámetro de precisión donde $\sqrt{\tau s}=\dfrac{1}{\sigma}$,
- $\tilde{\eta}$: predictor lineal, cuya representación matemática se basa bajo la reestructuración del parámetro μ de la forma siguiente

$$\tilde{\eta} \quad = \quad \tilde{\mu}=-\mu.$$

Por tanto, la función de densidad de probabilidad para Y con función de distribución acumulada (2.6), está dada por

$$f(y;\tilde{\eta},\tau,\xi) \quad = \quad \begin{cases} \sqrt{\tau s}\left(1-\xi\sqrt{\tau s}\left(y-\tilde{\eta}\right)\right)^{-1-1/\xi}\exp\left\{-\left(1-\xi\sqrt{\tau s}\left(y-\tilde{\eta}\right)\right)^{-1/\xi}\right\} & ;\xi\neq 0 \\ \sqrt{\tau s}\exp\left(\sqrt{\tau s}\left(y-\tilde{\eta}\right)\right)\exp\left\{-\exp\left(\sqrt{\tau s}\left(y-\tilde{\eta}\right)\right)\right\} & ;\xi=0. \end{cases}$$

Se denota la distribución, con esta función de distribución acumulada (f.d.a) como

$$Y \quad \sim \quad \text{GEV}(\tilde{\eta},\tau,\xi).$$

2.3. Dependencia espacial

Cuando nos referimos a dependencia espacial, enfocamos el estudio hacia un efecto de relaciones entre estaciones, entendiéndose por estas como coordenadas medidas bajo técnicas de georeferenciación. Como punto inicial, incluimos la definición de un campo espacial, denotado por $Y(s)$ donde s hace referencia a un vector de variables aleatorias $\{s_i\}_{i=1}^{n}$ de locales, con $s_i\in\mathcal{D}\subset\mathbb{R}^2$, e $Y(s_i)$ a una variable aleatoria medida en dicho local s_i.

En tanto, de acuerdo a Banerjee et al. (2003), un campo o proceso espacial $Y(s), s\in\mathcal{D}$, es llamado proceso espacial gaussiano si la distribución conjunta de cualquier realización $\mathbf{Y(s)}=\mathbf{Y}=(\mathbf{Y(s_1)},\ldots,\mathbf{Y(s_n)})^{\top}$ tiene una distribución normal multivariada con media $\mu=(\mu_{s_i},\cdots,\mu_{s_n})^{\top}$ y matriz de covarianza Σ compuesta por $Cov(Y(s_i),Y(s_j))=C_{ij}$, donde $C(d_{i,j})$ es la función de covarianza que solo depende de la distancia euclidiana entre los locales s_i y s_j siendo $d_{i,j}=\|s_j-s_i\|$,cuya función de densidad de probabilidad (f.p.d) es:

$$f_Y(y) \quad = \quad \frac{1}{(2\pi)^{n/2}\mid\Sigma\mid^{-1/2}}\exp\left\{-\frac{1}{2}(Y-\mu)^T\Sigma^{-1}(Y-\mu)\right\}.$$

Una importante definición sobre un campo espacial es la de ser estrictamente estacionario o débilmente estacionario. Un campo espacial $Y(s)$ es estrictamente estacionario, si, para todo $n \geq 1$, cualquier conjunto $\{s_i\}_{i=1}^{n}$ de locales y cualquier $d \in \mathbb{R}^2$, la distribución de $(Y(s_1), \ldots, Y(s_n))$ es la misma que la de $(Y(s_1 + d), \cdots, Y(s_n + d))$. Una condición menos estricta sobre un campo espacial está dado por una estacionaridad débil, o estacionaridad de segundo orden, la cual se define si la media $\mu(s)$ es constante sobre cada locación y la covarianza $Cov(Y(s_i), Y(s_i + d)) = Cov(d)$, para todo $d \in \mathbb{R}^2$. Nótese que un campo espacial estrictamente estacionario implica ser débilmente estacionario; sin embargo, la implicancia inversa solo es cierta para un proceso espacial gaussiano.

2.3.1. Variogramas

La definición de estacionaridad débil implica el hecho de que $E[Y(s + d)] = E[Y(s)]$; sin embargo, solo asumida esta igualdad se define una estacionaridad intrínseca, permitiendo lo siguiente

$$Var(Y(s_i + d) - Y(d)) \quad = \quad 2\gamma(d),$$

donde $2\gamma(d)$ se denomina un variograma y $\gamma(d)$ un semivariograma. La relación entre el semivariograma $\gamma(d)$ con la función de covarianza $C(.)$ es la siguiente

$$\begin{aligned}
\gamma(d) &= \left(\frac{1}{2}\right) Var(Y(s_i + d) - Y(s_i)) \\
&= \left(\frac{1}{2}\right) (Var(Y(s_i + d)) + Var(Y(s_i)) - 2Cov(Y(s_i + d), Y(s_i))) \\
&= \left(\frac{1}{2}\right) (C(0) + C(0) - 2C(d)) \\
\gamma(d) &= C(0) - C(d). \qquad (2.7)
\end{aligned}$$

El semivariograma presenta tres características principales las cuales permiten tener su implicancia bajo distintos modelos. Estas características son las siguientes:

- τ^2 (*pepita*): Valor de $\gamma(d)$ para el cual la distancia $d \to 0$, es decir cuando la distancia d entre dos locales tiende a ser nulo.
- $\tau^2 + \sigma^2$ (*meseta*): Valor de $\gamma(d)$ para el cual la distancia $d \to \infty$, es decir cuando la distancia d es amplia.
- $(1/\phi)$ (*Rango*): La distancia d a la cual $\gamma(d)$ se vuele constante.

2.3.2. Modelo Mátern

De acuerdo a Banerjee et al. (2003), el modelo Matérn es un tipo de modelo estacionario que permite solo su dependencia con la distancia d y define su semivariograma como la

función isotrópica:

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2 \left[1 + \frac{(\phi d)^\nu}{2^{(\nu-1)}\Gamma(\nu)} K_\nu(\phi d)\right] & , \ d > 0; \\ 0 & , d = 0, \end{cases} \tag{2.8}$$

donde ν es un parámetro de suavizamiento, K_ν es la función de Bessel modificada de orden ν. De (2.7) y (2.8), tenemos que la función de covarianza Matérn es definida por:

$$\begin{aligned} C(d) &= \lim_{u\to\infty} \gamma(u) - \gamma(d) \\ &= \tau^2 + \sigma^2 - \left[\tau^2 + \sigma^2 \left[1 + \frac{(\phi d)^\nu}{2^{(\nu-1)}\Gamma(\nu)} K_\nu(\phi d)\right]\right] \\ &= \sigma^2 \frac{(\phi d)^\nu}{2^{(\nu-1)}\Gamma(\nu)} K_\nu(\phi d), \end{aligned} \tag{2.9}$$

donde σ^2 la varianza marginal del campo gaussiano y $\rho(d) = \frac{(\phi d)^\nu}{2^{(\nu-1)}\Gamma(\nu)} K_\nu(\phi d)$ es la función de correlación Mátern.
Por otro lado, se define el *rango efectivo* como la distancia a la cual dicha correlación alcanza el valor de 0.1 aproximadamente, esto es

$$r_* = \frac{\sqrt{8\nu}}{\phi}. \tag{2.10}$$

En la Figura 2.2 se visualiza función de correlación Matérn, considerando como parámetros $\nu = 0.5, \rho(d) = 4.34$; $\nu = 1, \rho(d) = 3.11$ y $\nu = 2, \rho(d) = 2.25$. Se puede observar que a medida que la distancia entre dos locales, la autocorrelación decrece y se aproxima a 0.

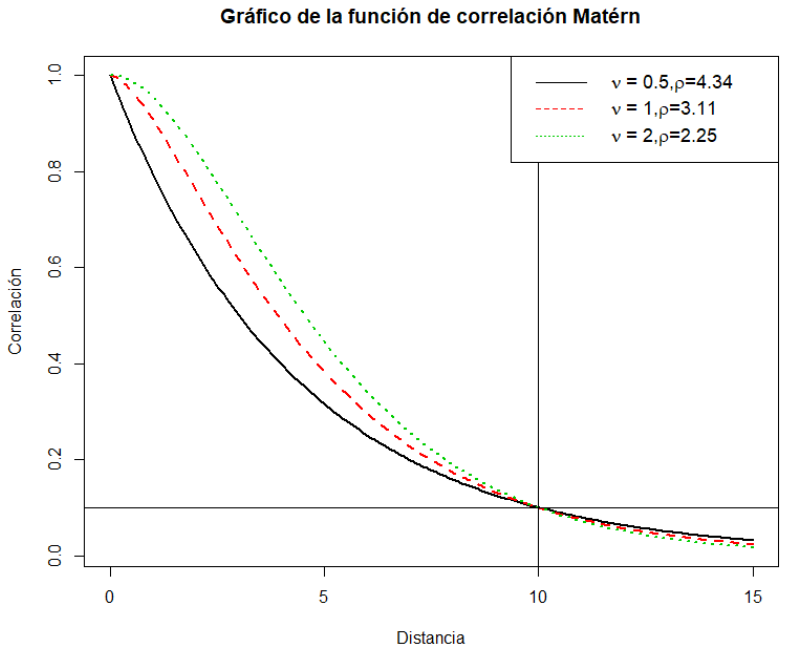

Figura 2.2: Función de correlación Matérn con diferentes valores de ν y $\rho(d)$.

2.3.3. Ecuaciones diferenciales parciales estocásticas (SPDE)

Las matrices de covarianza altamente densas en su estructura, requieren de un procesamiento computacional de estimación adecuado. Las ecuaciones diferenciales parciales estocásticas (SPDE, *spatial stochastic differential equation*) propuestas por Lindgren et al. (2011) permiten aproximar el campo gaussiano espacial (GF, *gaussian field*), de covarianza Matérn denotado como $f(s)$ en la presente tesis, como un campo aleatorio gaussiano discreto de Markov (GMRF, *Gaussian Markov Random Fields*). Para ello utiliza un resultado presentado en Whittle (1963) donde la solución de la siguiente ecuación diferencial parcial estocástica:

$$\underbrace{(\phi^2 - \Delta)^{(\nu+1)/2}}_{\substack{\text{operador} \\ \text{pseudo-diferencial}}}(\tau f(s)) \quad = \quad \mathcal{W}(s), \tag{2.11}$$

donde $f(s)$ es un campo gaussiano con función de covarianza Matérn. Además en (2.11), $s = (s_i, s_j) \in \mathbb{R}^2$ representa las coordenadas de un local, ν es un parámetro de suavizamiento, $\mathcal{W}(s)$ es un campo gaussiano espacial con ruido blanco y Δ es el operador Laplaciano, en coordenadas bidimensionales, definido como

$$\Delta \quad = \quad \frac{\partial^2}{\partial s_i^2} + \frac{\partial^2}{\partial s_j^2}.$$

La varianza marginal σ^2, en (2.9), queda definida por

$$\sigma^2 \quad = \quad \frac{\Gamma(\nu)}{\Gamma(\nu+1)(4\pi)\phi^{2\nu}\tau^2}\ . \tag{2.12}$$

Cabe señalar que si el parámetro de suavización es $\nu = 1$, las ecuaciones (2.10) y (2.12) quedarían definidas de la siguiente forma

$$r_* \quad = \quad \frac{\sqrt{8}}{\phi} \tag{2.13}$$

$$\sigma^2 \quad = \quad \frac{1}{(4\pi)\phi^{2\nu}\tau^2}\ . \tag{2.14}$$

Usando métodos de elementos finitos el campo gaussiano $f(s)$, en (2.11), se aproxima a $\tilde{f}$, a través de una representación con funciones base $\{\varphi_g\}$ definidas en una triangulación del dominio $\mathcal{D}$, de la forma siguiente

$$\tilde{f}(s) \quad = \quad \sum_{g=1}^{m} \varphi_g(s) f_g^*, \tag{2.15}$$

donde m es el número de vértices de la triangulación y $\mathbf{f}^* = (f_1^*, f_2^*, \cdots, f_m^*)^T$ se distribuye como un proceso gaussiano. Con el objetivo de obtener una estructura de Markov, las funciones base sobre un local s deben mantener un soporte local, esto quiere decir que cada $\varphi_g(s)$

es 1 en el vértice g y 0 en cualquier otro vértice. Luego,

$$f^* \quad \sim \quad N(0, Q^{-1}) \quad \text{con} \qquad Q = \tau^2 \left(\phi^4 C + 2\phi^2 G + G C^{-1} G\right), \tag{2.16}$$

donde Q se denomina matriz de precisión de f^* definida bajo una matriz diagonal C y una matriz llena de ceros G como sigue a continuación

$$C = [C_{ii}]_{m\times m} = \left[\int \varphi_i(s) ds\right]_{m\times m} \quad , \quad G = [G_{ij}]_{m\times m} = \left[\int \bigtriangledown \varphi_i(s) \bigtriangledown \varphi_j(s) ds\right]_{m\times m}. \tag{2.17}$$

2.4. Aproximación de laplace integrada y anidada (INLA *Integrated Nested Laplace Aproximation*)

El algoritmo INLA, propuesto por Rue et al. (2009), es un algoritmo determinístico para inferencia bayesiana, específicamente diseñado para modelos gaussianos latentes, y que comparados con MCMC (*Markov Chain Monte Carlo*) provee estimaciones en tiempos de cómputo más cortos.

2.4.1. Estructura del INLA en la clase de modelos gaussianos latentes

De acuerdo a Blangiardo y Cameletti (2015) el primer paso para definir un modelo gaussiano latente dentro del marco de referencia bayesiano es identificar una distribución para los datos observados $y = (y_1, \ldots, y_n)^T$. Un enfoque muy general consiste en especificar una distribución para y_i caracterizada por un parámetro μ_i (generalmente la media $E(y_i)$) definida como una función de un predictor aditivo estructurado η_i través de una función de enlace $g(\cdot)$, tal que $g(\mu_i) = \eta_i$. El predictor lineal aditivo η_i se define como sigue:

$$\eta_i = \beta_0 + \sum_{p=1}^{P} \beta_m x_{mi} + \sum_{l=1}^{L} f_l(z_{li}), \tag{2.18}$$

donde β_0 es un intercepto; $\beta_1, \ldots, \beta_P$ son los coeficientes de regresión que cuantifican el efecto lineal de las covariables $x_1, \cdots, x_P$; y $f_1(\cdot), \cdots, f_L$ son efectos aleatorios definidos para las covariables $z_1, \ldots, z_L$. Los términos $f_l(\cdot)$ pueden asumir diferentes formas, tales como efectos suaves y no lineales de covariables, tendencias temporales y efectos estacionales, efectos aleatorios para individuos, así como efectos aleatorios temporales. Por esta razón, la clase de modelos gaussianos latentes es muy flexible y puede adaptarse a una amplia gama de modelos que van desde modelos lineales generalizados y dinámicos a modelos espaciales y espacio-temporales.

Se recopila todos los componentes latentes (no observables) de interés para la inferencia en un conjunto de parámetros llamados $\mathbf{w}$ definidos como $\mathbf{w} = \{\beta_0, \beta_1, \cdots, \beta_p, f\}$. Además, se denota con $\theta = \{\theta_1, \cdots, \theta_K\}$ donde K es el vector de hiperparámetros. Al asumir la independencia condicional, la distribución de las n variables aleatorias (todas provenientes de la misma familia de distribución) viene dada por la función de verosimilitud

$$p(\mathbf{y} \mid \mathbf{w}, \theta) \quad = \quad \prod_{i=1}^{n} p(y_i \mid w_i, \theta). \tag{2.19}$$

Se asume una normal multivariada a priori para $\mathbf{w}$ con media 0 y matriz de precisión $Q(\theta)$, es decir, $\mathbf{w} \sim Normal(0, Q^{-1}(\theta))$ con función de densidad dada por

$$p(\mathbf{w} \mid \theta) = (2\pi)^{-n/2} \mid Q(\theta) \mid^{1/2} \exp\left(-\frac{1}{2}\mathbf{w}^T Q^{-1}(\theta)\mathbf{w}\right) \tag{2.20}$$

donde $\mid \cdot \mid$ denota el determinante, y $Q(\theta)$ es una matriz de precisión llena de ceros. De acuerdo a Rue y Held (2005) esta especificación se conoce como campo aleatorio de Markov gaussiano (GMRF *Gaussian Markov Random Fields*). La distribución a posteriori conjunta de $\mathbf{w}$ y θ viene dada por la productoria de la verosimilitud definida en la ecuación (2.19) de la función de densidad del GMRF dado en la ecuación (2.20) y la distribución a priori de los hiperparámetros $p(\theta)$. Esto es,

$$\begin{aligned}
p(\mathbf{w}, \theta \mid y) &\propto p(\theta) \times p(\mathbf{w} \mid \theta) \times p(y \mid \mathbf{w}, \theta) \\
&\propto p(\theta) \times p(\mathbf{w} \mid \theta) \times \prod_{i=1}^{n} p(y_i \mid w_i, \theta) \\
&\propto p(\theta) \times \mid Q(\theta) \mid^{1/2} \exp\left(-\frac{1}{2}\mathbf{w}^T Q(\theta)\mathbf{w}\right) \times \prod_{i=1}^{n} \exp(\log(p(y_i \mid w_i, \theta))) \\
&\propto p(\theta) \times \mid Q(\theta) \mid^{1/2} \exp\left(-\frac{1}{2}\mathbf{w}^T Q(\theta)\mathbf{w} + \sum_{i=1}^{n} \log(p(y_i \mid w_i, \theta))\right).
\end{aligned} \tag{2.21}$$

2.4.2. Inferencia bayesiana con INLA

La estimación de los parámetros será realizado bajo un enfoque bayesiano, haciendo uso del método de aproximación de Laplace integrada y anidada (INLA. *Integrated Nested Laplace Aproximation*), propuesta por Rue et al. (2009), restringida para modelos gaussianos latentes. Este método realiza un cálculo numérico directo de las densidades marginales posteriores para una gran sub-clase gaussiana latente (LGM) de modelos jerárquicos bayesianos, evitando el alto costo computacional en las simulaciones realizas por el método MCMC. La implementación abarca modelos de la siguiente forma

$$\begin{aligned}
\theta &\sim p(\theta) \\
\mathbf{w} \mid \theta &\sim N(0, Q(\theta)^{-1}) \\
y_i \mid \mathbf{w}, \theta &\sim p(y_i \mid \mathbf{w}, \theta)
\end{aligned}$$

donde θ son los hiperparámetros, $\mathbf{w}$ es el campo gaussiano latente, e y es el vector de datos observados.

De acuerdo a Blangiardo y Cameletti (2015), el objetivo del INLA es hallar las distribuciones de las marginales a posteriori para cada elemento del vector de parámetros

$$\begin{aligned}
p(w_i \mid \mathbf{y}) &= \int p(w_i, \theta \mid \mathbf{y}) d\theta \\
&= \int p(w_i \mid \theta, \mathbf{y}) p(\theta \mid \mathbf{y}) d\theta,
\end{aligned} \tag{2.22}$$

y para cada elemento de vector de hiperparámetros

$$p(\theta_k \mid y) \quad = \quad \int p(\theta \mid y) d\theta_{-k}.$$

Así necesitamos llevamos a cabo los siguientes pasos:

1. Calcular la condicional de $p(\theta \mid \mathbf{y})$ con lo que también puede obtener todas las marginales $p(\theta_k \mid \mathbf{y})$.

2. Calcular $p(w_i \mid \theta, \mathbf{y})$ el cual se necesita para calcular las marginales a posteriori $p(w_i \mid \mathbf{y})$.

El enfoque INLA explota los supuestos del modelo para producir una aproximación numérica de las posterioris de interés basados en el método de aproximación de Laplace, introducida por Tierney y Kadane (1986).

El primer paso consiste en calcular una aproximación de la posteriori conjunta de los hiperparámetros como

$$\begin{aligned} p(\theta \mid \mathbf{y}) &= \frac{p(\mathbf{w}, \theta \mid \mathbf{y})}{p(\mathbf{w} \mid \theta, \mathbf{y})} \\ &= \frac{p(\mathbf{y} \mid \mathbf{w}, \theta) p(\mathbf{w}, \theta)}{p(\mathbf{y})} \frac{1}{p(\mathbf{w} \mid \theta, \mathbf{y})} \\ &= \frac{p(\mathbf{y} \mid \mathbf{w}, \theta) p(\mathbf{w}, \theta) p(\theta)}{p(\mathbf{y})} \frac{1}{p(\mathbf{w} \mid \theta, \mathbf{y})} \\ &\propto \frac{p(y \mid \mathbf{w}, \theta) p(\mathbf{w}, \theta) p(\theta)}{p(\mathbf{w} \mid \theta, \mathbf{y})} \\ &\approx \frac{p(\mathbf{y} \mid \mathbf{w}, \theta) p(\mathbf{w}, \theta) p(\theta)}{\tilde{p}(\mathbf{w} \mid \theta, \mathbf{y})} \Big|_{\mathbf{w}=\mathbf{w}^*(\theta)} =: \tilde{p}(\theta \mid \mathbf{y}), \end{aligned} \tag{2.23}$$

donde $\tilde{p}(\mathbf{w} \mid \theta, \mathbf{y})$ es la aproximación gaussiana de $p(\mathbf{w} \mid \theta, y)$ y $\mathbf{w}^*(\theta)$ es la moda para un θ en particular. Dado que $p(\mathbf{w} \mid \theta, \mathbf{y})$ se aproxima a una distribución gaussiana usualmente hace el análisis menos complejo.

El segundo paso es ligeramente más complejo, porque en general existen más elementos en $\mathbf{w}$ que en θ, siendo este cálculo más costoso. Una primera posibilidad es aproximar la distribución condicional a posteriori $p(w_i \mid \theta, \mathbf{y})$ directamente de las marginales de $\tilde{p}(\mathbf{w} \mid \theta, \mathbf{y})$, es decir usando una distribución Normal, donde la descomposición Cholesky es usada para la matriz de precisión. Aunque esto es muy rápido, la aproximación generalmente no es tan buena. La segunda posibilidad es reescribir el vector de parámetros como $\mathbf{w} = (w_i, \mathbf{w_{-i}})$ y usar el método de aproximación Laplace para obtener

$$\begin{aligned} p(w_i \mid \theta, \mathbf{y}) &= \frac{p(w_i, \mathbf{w}_{-i} \mid \theta, \mathbf{y})}{p(\mathbf{w_{-i}} \mid w_i, \theta, \mathbf{y})} \\ &= \frac{p(\mathbf{w}, \theta \mid \mathbf{y})}{p(\mathbf{w} \mid \mathbf{y})} \frac{1}{p(\mathbf{w}_{-i} \mid w_i, \theta, \mathbf{y})} \\ &\propto \frac{p(\mathbf{w}, \theta \mid \mathbf{y})}{p(\mathbf{w}_{-i} \mid w_i, \theta, \mathbf{y})} \end{aligned}$$

$$p(w_i \mid \theta, \mathbf{y}) \approx \left.\frac{p(\mathbf{w}, \theta \mid \mathbf{y})}{\tilde{p}(\mathbf{w}_{-i} \mid w_i, \theta, \mathbf{y})}\right|_{\mathbf{w}_{-i}=\mathbf{w}^*_{-i}(w_i,\theta)} =: \tilde{p}(w_i \mid \theta, \mathbf{y}),$$

donde $\tilde{p}(\mathbf{w}_{-i} \mid w_i, \theta, \mathbf{y})$ es la aproximación gaussiana $p(\mathbf{w}_{-i} \mid w_i, \theta, \mathbf{y})$ y $\mathbf{w}^*_{-i}(w_i, \theta)$ es la moda. Debido a que las variables aleatorias $\mathbf{w}_{-i} \mid w_i, \theta, \mathbf{y}$ se asemejan a una normal, la aproximación dada en la ecuación (2.24) típicamente trabaja muy bien. Esta estrategia, sin embargo, puede ser muy costosa computacionalmente dado que términos como $\tilde{p}(\mathbf{w}_{-i} \mid w_i, \theta, \mathbf{y})$ deben ser recomputarizados para cada valor de $\mathbf{w}$ y θ.

La tercera posibilidad es la *Aproximación Laplace simplificada*, la cual está basada en la expansión de la serie de Taylor de la aproximación de Laplace $\tilde{p}(w_i \mid \theta, \mathbf{y})$ en la ecuación (2.24). Esto es usualmente corregido al incluir un término de mistura, es decir una spline, para incrementar el ajuste a la distribución requerida. La precisión de esta distribución es suficiente en muchos casos aplicados y dado que el tiempo necesitado para el cálculo es considerablemente menor.

Una vez que obtenemos $\tilde{p}(w_i \mid \theta, \mathbf{y})$ y $\tilde{p}(\theta \mid \mathbf{y})$, las distribuciónes marginales a posteriori $p(w_i \mid \mathbf{y})$, introducidas en la ecuación (3.8) son aproximadas por,

$$\tilde{p}(w_i \mid \mathbf{y}) \approx \int \tilde{p}(w_i \mid \theta, \mathbf{y})\tilde{p}(\theta \mid \mathbf{y})d\theta, \tag{2.24}$$

donde la integral puede ser resuelta numéricamente a través de una suma ponderada finita:

$$\tilde{p}(w_i \mid \mathbf{y}) \approx \sum_j \tilde{p}(w_i \mid \theta^j, \mathbf{y})\tilde{p}(\theta^j \mid \mathbf{y})\Delta_j, \tag{2.25}$$

para algunos puntos de integración relevantes θ^j con el conjunto correspondiente de pesos Δ_j.

2.4.3. Penalised Complexity Prior (PC a Priori)

De acuerdo a Fuglstad et al. (2017), la inclusión de un GRF (*Gaussian Random Field*) en un modelo puede llevar a un sobreajuste de, por ejemplo, la estimación de tendencias espaciales espurias o tendencias temporales espurias. Simpson et al. (2017) sugiere manejar el problema del sobre ajuste observando los componentes del modelo.

El primer paso de su enfoque es derivar una distancia desde el modelo base hasta su extensión flexible utilizando la divergencia Kullback-Leibler (KLD). El propósito de la distancia es proporcionar una mejor parametrización del componente del modelo donde el tamaño del cambio en el parámetro corresponde al tamaño del cambio en la diferencia entre el componente del modelo y su modelo base. En Fuglstad et al. (2017) describen el modelo base para el GRF mediante la medida gaussiana P_0 y el modelo flexible mediante la medida gaussiana P, y luego definiendo la distancia mediante $dist(P\|P_0) = \sqrt{2KL(P\|P_0)}$, donde $KL(P_0\|P)$ es la KLD de P_0 a P y se definide como lo siguiente:

Definición 2.4.1 *(Divergencia de Kullback-Leibler) Sean P_0 y P dos medidas sobre un conjunto χ, donde P es absolutamente continuo con respecto a P_0, luego la divergencia de*

Kullback-Leibler de P_0 a P se define como

$$KL(P|P_0) \quad = \quad \int_{\chi} \log \frac{dP}{dP_0} dP,$$

donde $\frac{dP}{dP_0}$ *es la derivada de Randon-Nikodym de* P *con respecto a* P_0.

La KLD cuantifica la información perdida al usar el modelo base para aproximarse a un modelo flexible (Simpson et al. (2017)). El segundo paso de la construcción de la priori es definir la priori sobre la distancia derivada utilizando tres principios: *razon Occam*, la penalización de velocidad constante y la escala definida por el usuario. *razon Occam* significa que lo anterior penaliza más y más fuertemente cuanto más se aleje del modelo base y se puede lograr utilizando una penalización de tasa constante, tal que para la distancia, t, se satisface

$$\frac{\pi(t+\delta)}{\pi(t)} = r^{\delta} \quad , \quad t, \delta > 0,$$

para una tasa de caída constante $0 < r < 1$. La única distribución continua con esta propiedad es la distribución exponencial $\pi(t) = \lambda \exp(-\lambda t)$, para $t > 0$, donde el cambio relativo en la a priori cuando la distancia aumenta no depende de la distancia actual t. La distancia en sí no suele ser interpretada directamente por el usuario y debe transformarse a un tamaño interpretable $Q(t)$. La información previa se puede incluir, por ejemplo, a través de las probabilidades de cola $P(Q(d) > U) = \alpha$ o $P(Q(d) < L) = \alpha$, donde U o L es un límite superior o inferior, respectivamente y α es la probabilidad de cola superior o inferior de la distribución anterior. A través de esta construcción,la PC a priori combina la geometría del espacio de parámetros con la creencia previa sobre un tamaño interpretable.

La PC Prior conjunta de la función de covarianza Matérn utilizada en la presente tesis, se encuentra basada en el siguiente teorema propuesto en Fuglstad et al. (2017):

Teorema 2.4.1 *(PC Prior para la Mátern* (σ, ϕ)*) Sea* u *un GRF definido en* $\mathbb{R}^d$*, donde* $d \leq 3$*, con una función de covarianza de Matérn con los parámetros* σ*,* ϕ *y* ν*. Entonces, la PC prior conjunta correspondiente a un modelo base con rango infinito y varianza cero es*

$$\pi(\sigma, \phi) = \frac{d}{2} \tilde{\lambda_1} \tilde{\lambda_2} \phi^{-d/2-1} \exp(-\tilde{\lambda_1} \phi^{-d/2} - \tilde{\lambda_2} \sigma) \quad , \quad \sigma > 0, \rho > 0,$$

donde $P(\phi < \phi_0) = \alpha_1$*,* $P(\sigma > \sigma_0) = \alpha_2$ *y:*

$$\tilde{\lambda_1} = -\log(\alpha_1)\phi_0^{d/2} \quad y \quad \tilde{\lambda_2} = -\frac{\log(\alpha_2)}{\sigma_0},$$

donde ϕ_0 *y* σ_0 *son definidas como límite inferior del parámetro phi y límite superior del parámetro* σ*, respectivamente.*

La a priori es fácil y rápida de calcular independientemente del número de observaciones y $d = 2$ proporciona el caso espacial bidimensional.

2.5. Evaluación del modelo bayesiano

2.5.1. Comparación de modelos

Para estudiar la bondad de ajuste de los modelos, usaremos el criterio de información de deviance (DIC), el logaritmo de la probabilidad pseudo marginal (LPML), y la estimación del error medio cuadrático (MSE).

El DIC es un criterio de información popular diseñado para modelos jerárquicos cuya principal aplicación es la selección del mejor modelo bayesiano (Spiegelhalter et al. (2002)).

El DIC se calcula en INLA como: $DIC = E^{\mathbf{w},\theta}\left[D(\theta,\mathbf{w})\right] + p_D$, donde la devianza está dada por:

$$D\left(\theta,\mathbf{w}\right) \quad = \quad -2\sum_{i\in G}^{n}\log\left\{p(y_i \mid \mathbf{w},\theta)\right\}, \tag{2.26}$$

la media a posteriori de la devianza, es calculado por INLA por

$$E^{\mathbf{w},\theta}\left[D(\theta,\mathbf{w})\right] \quad = \quad \int_{\theta,\mathbf{w}} D(\theta,\mathbf{w})\tilde{p}(\theta \mid \mathbf{y})p(\mathbf{w} \mid \theta,\mathbf{y})d\theta d\mathbf{w}, \tag{2.27}$$

y el número efectivo de parámetros es aproximado por

$$p_D \approx N_{\mathbf{w}} - traza\left\{Q(\theta^{me})Q^{*}(\theta^{me})^{-1}\right\}, \tag{2.28}$$

donde $N_{\mathbf{w}}$ es la dimensión de $\mathbf{w}$, θ^{me} denota la media posteriori, Q denota la matriz de precisión de la mediana posteriori y $(Q*)^{-1}$ denota la matriz de covarianza a posteriori de la aproximación gaussiana $\tilde{p}(\theta \mid \mathbf{y})$ en la mediana a posteriori (Rue et al. (2009)).

Otro criterio de comparación de modelos bayesianos es el de la ordenada predictiva condicional (CPO, *conditional predictive ordinate*) (Geisser y Eddy (1979)), definida por

$$\mathrm{CPO}_i = p(y_i \mid \mathbf{y}_{-i}), \tag{2.29}$$

donde $\mathbf{y}_{-i}$ representa a $\mathbf{y}$ sin el componente y_i. La estimación Monte Carlo (De et al. (1997); Held et al. (2010)), computacionalmente estructurada en INLA, es definida como la media armónica de la distribución condicional $p(y_{ij} \mid w_k, \theta_k)$

$$\widehat{\mathrm{CPO}}_i \quad = \quad \left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{p(y_i \mid w_k,\theta_k)}\right]^{-1}, \tag{2.30}$$

evaluada en muestras $w_1,\cdots,w_K$ de $p(w_i \mid \mathbf{y})$ para $i \in G$.

Esta buena medida de ajuste para cada observación puede ser resumida hacia todos los datos observados, como el logaritmo de la probabilidad pseudomarginal definida como

$$\mathrm{LPML} \quad = \quad \sum_{i\in G}^{n}\log p(y_i \mid \mathbf{y}_{-i}) \quad \approx \quad \sum_{i\in G}^{n}\log \widehat{CPO}_i, \tag{2.31}$$

un valor más alto de LPML corresponde a un mejor modelo. Otra medida de ajuste de un

modelo es el de la raíz de la estimación del error cuadrático medio (RMSE) determinada como

$$\text{RMSE} \quad = \quad \sqrt{\frac{1}{n}\sum_{i \in G}^{n} d_i^2}, \;\; \text{con } d_i = y_i - E[\mathbf{Y}_i \mid \mathbf{w}, \theta]. \tag{2.32}$$

Capítulo 3

Modelos para valores extremos mínimos

3.1. Modelo GEV para valores extremos mínimos

Sea $\mathbf{Y}^\star = (Y_1^\star, \ldots, Y_n^\star)^T$ el vector de variables aleatorias $Y_i^\star$ que representan la temperatura mínima en la estación $s_i \in \mathcal{D} \subset \mathbb{R}^2$, donde $i = 1, \ldots, n$. Se asume que $Y_i^\star \sim GEV(\tilde{\eta_i}, \tau, \xi)$, es decir sigue una distribución generalizada de valores extremos (GEV) mínimos, cuya función de densidad está dada por:

$$f_{Y_i^\star}(y_i^\star; \tilde{\eta}, \tau, \xi) = \begin{cases} \sqrt{\tau s}\left(1 - \xi\sqrt{\tau s}\left(y_i^\star - \tilde{\eta_i}\right)\right)^{-1-1/\xi} \exp\left\{-\left(1 - \xi\sqrt{\tau s}\left(y_i^\star - \tilde{\eta_i}\right)\right)^{-1/\xi}\right\} & ;\xi \neq 0 \\ \sqrt{\tau s} \quad \exp\left(\sqrt{\tau s}\left(y_i^\star - \tilde{\eta_i}\right)\right) \exp\left\{-\exp\left(\sqrt{\tau s}\left(y_i^\star - \tilde{\eta_i}\right)\right)\right\} & ;\xi = 0, \end{cases}$$

donde $s > 0$ es un parámetro de escala fijo, τ es el parámetro de precisión, ξ el parámetro de forma y $\tilde{\eta_i}$ es el predictor lineal asociado a la variable $Y_i^\star$. Mientras que su función de distribución acumulada (f.d.p) está dada por:

$$F_{Y_i^\star}(y_i^\star; \tilde{\eta_i}, \tau, \xi) = \begin{cases} 1 - \exp\left\{-\left(1 - \xi\sqrt{\tau s}(y_i^\star - \tilde{\eta_i})\right)^{-1/\xi}\right\} & ;\xi \neq 0,\ 1 - \xi\sqrt{\tau s}(y_i^\star - \tilde{\eta_i}) > 0 \\ 1 - \exp\left\{-\exp\left(\sqrt{\tau s}(y_i^\star - \tilde{\eta_i})\right)\right\} & ;\xi = 0,\ y_i^\star \in \mathbb{R}. \end{cases} \tag{3.1}$$

Como se explicó en el capítulo 2.2.2 de forma equivalente se puede asumir que la v.a. $Y_i = -Y_i^\star$ representa el máximo de las temperaturas $-Y_i^\star$. Luego se puede asumir que Y_i sigue una distribución generalizada de valores extremos máximos, $Y_i \sim \text{GEV}(\eta_i, \tau, \xi),\ \forall\ i = 1, \ldots, n$, cuya función de distribución acumulada (f.d.p) es dada por:

$$f_{Y_i}(y_i; \eta_i, \tau, \xi) = \begin{cases} \sqrt{\tau s}\left(1 + \xi\sqrt{\tau s}\left(y_i - \eta_i\right)\right)^{-1-1/\xi} \exp\left\{-\left(1 + \xi\sqrt{\tau s}\left(y_i - \eta_i\right)\right)^{-1/\xi}\right\} & ;\xi \neq 0 \\ \sqrt{\tau s} \quad \exp\left(-\sqrt{\tau s}\left(y_i - \eta_i\right)\right) \exp\left\{-\exp\left(-\sqrt{\tau s}\left(y_i - \eta_i\right)\right)\right\} & ;\xi = 0, \end{cases} \tag{3.2}$$

donde η_i es el predictor lineal asociado a la variable Y_i. En particular se usa esta transformación de la v.a $Y_i^\star$ por Y_i porque nos permite estimar los parámetros a partir de la fdp de una distribución GEV para máximos, que es la usualmente implementada en los paquetes estadísticos como el INLA.

La función de distribución acumulada de Y_i está definida por:

$$F(y_i; \eta_i, \tau, \xi) = \begin{cases} \exp\left\{-\left(1 + \xi\sqrt{\tau s}\left(y_i - \eta_i\right)\right)^{-1/\xi}\right\} & ;\xi \neq 0,\ 1 + \xi\sqrt{\tau s}\left(y_i - \eta_i\right) > 0 \\ \exp\left\{-\exp\left(-\sqrt{\tau s}\left(y_i - \eta_i\right)\right)\right\} & ;\xi = 0,\ y_i \in \mathbb{R}, \end{cases} \tag{3.3}$$

El predictor lineal η_i es definido de la siguiente forma:

$$\eta_i \quad = \quad \mathbf{X_i^T}\beta,$$

donde el vector β es definido por los coeficientes de la regresión, $[\mathbf{X}_i]_{n\times p}$ representa la matriz de covariables.

De acuerdo a la sección 2.4.1, se define el vector del campo gaussiano latente como $\mathbf{w} = (\beta)$, y el vector de hiper-parámetros $\theta = (\tau, \xi)$.

Recopilados los componentes latentes de interés para la inferencia y asumiendo independencia condicional entre las $\mathbf{Y_i}$, dados $\mathbf{w}$ y θ, las Y_i's siguen una distribución GEV, luego la función de verosimilitud viene dada por:

$$p(\mathbf{y} \mid \theta) = \prod_{i=1}^{n} f_{Y_i}(y_i \mid \eta_i, \xi, \tau),$$

la cual es definida en la ecuación 3.2 dependiendo de $\xi = 0$ o $\xi > 0$ y $\eta_i = \mathbf{X_i^T}\beta$.

3.1.1. Inferencia bayesiana bajo el enfoque INLA

Asumiendo, que sobre n datos observados $\mathbf{y}$ provienen de una distribución generalizada de valores extremos máximos (GEV) con parámetros η, $\theta_1 = \tau$ y $\theta_2 = \xi$. Así mismo, siendo η un predictor lineal asociado a un vector de coeficientes β y p covariables, se tendría lo siguiente

$$\begin{aligned} \theta &\sim p(\theta_k) \quad , \text{ para } k = 1, 2. \\ \eta_i &= X_i^T\beta \\ \mathbf{y} \mid \beta, \theta &\sim p(\mathbf{y} \mid \eta, \theta) \quad , \text{ donde } \theta = (\theta_1, \theta_2). \end{aligned}$$

Bajo el criterio de estimación INLA, el modelo gaussiano latente jerárquico tiene la siguiente forma

$$\begin{aligned} \theta &\sim p(\theta_k) \quad , \text{ para } k = 1, 2, 3, 4. \\ \eta_i &= X_i^T\beta \\ \mathbf{y} \mid \mathbf{w}, \theta &\sim p(\mathbf{y} \mid \eta, \theta) \quad , \ \theta = (\theta_1, \theta_2). \end{aligned}$$

donde η_i es el predictor lineal.

Entonces, el modelo completo para la inferencia bayesiana está definido de la siguiente forma

$$\begin{aligned} Y_i \mid \mathbf{w}, \theta &\sim GEV\left(\eta_i, \tau, \xi\right) \\ p(\beta_0) &\propto 1 \\ \beta_l &\sim N\left(0, 0.00001\right); l = 1, \ldots, p-1 \\ \log(\tau) &\sim LogGamma(1, 0.000001) \\ \xi &\sim N(0, 25), \end{aligned} \tag{3.4}$$

donde INLA define por defecto una priori impropia para β_0 y distribuciones gaussiana a prioris no informativa para el vector β_l. Para la definición de las a prioris de los hiperparámetros

τ y σ, se consideraron a prioris no informativas.

Luego, haciendo uso de la ecuación (2.21), la función de densidad a posteriori conjunta del campo gaussiano latente $\mathbf{w} = \{\beta\}$ y los hiperparámetros θ, toma la forma siguiente

$$\begin{aligned} p(\mathbf{w}, \theta \mid \mathbf{y}) &\propto p(\theta)p(\mathbf{w} = \{\beta\} \mid \theta)p(\mathbf{y} \mid \mathbf{w}, \theta) \\ &\propto p(\theta)p(\beta) \mid \theta^*)p(\mathbf{y} \mid \mathbf{w}, \theta) \end{aligned}$$

$$\begin{aligned} p(\mathbf{w}, \theta \mid \mathbf{y}) &\propto p(\theta)p(\beta) \exp\left\{\sum_{i=1}^{n} \log(f_{Y_i}(y_i \mid \eta_i, \theta))\right\}. \\ &\propto p(\tau)p(\xi)p(\beta_0)\prod_{l=1}^{p} p(\beta_l) \exp\left\{\sum_{i=1}^{n} \log(f_{Y_i}(y_i \mid \eta_i, \theta))\right\}. \end{aligned}$$

A partir de esta conjunta a posteriori, se procede a estimar los parámetros utilizando la aproximación INLA.

3.2. Modelo geoestadístico GEV para valores extremos mínimos

Sea $\mathbf{Y}^\star = (Y_1^\star, \ldots, Y_n^\star)^T$ el vector de variables aleatorias $Y_i^\star$ que representan la temperatura mínima en la estación $s_i \in \mathcal{D} \subset \mathbb{R}^2$, donde $i = 1, \ldots, n$. Se asume que $Y_i^\star$ sigue una distribución generalizada de valores extremos mínimos. Como se explicó en el capítulo 2.2.2 de forma equivalente se puede asumir que la v.a. $Y_i = -Y_i^\star$ representa el máximo de las temperaturas $-Y_i^\star$. Luego se puede asumir que Y_i sigue una distribución generalizada de valores extremos máximos, $Y_i \sim \text{GEV}(\eta_i, \tau, \xi),\ \forall\ i = 1, \ldots, n$, cuya función de distribución acumulada (f.d.p) es dada por:

$$f_{Y_i}(y_i; \eta_i, \tau, \xi) = \begin{cases} \sqrt{\tau s}\,(1 + \xi\sqrt{\tau s}\,(y_i - \eta_i))^{-1-1/\xi} \exp\left\{-(1 + \xi\sqrt{\tau s}\,(y_i - \eta_i))^{-1/\xi}\right\} & ;\xi \neq 0 \\ \sqrt{\tau s}\exp\left(-\sqrt{\tau s}\,(y_i - \eta_i)\right)\exp\left\{-\exp\left(-\sqrt{\tau s}\,(y_i - \eta_i)\right)\right\} & ;\xi = 0, \end{cases} \tag{3.5}$$

donde $s > 0$ es un parámetro de escala fijo, τ es el parámetro de precisión, ξ el parámetro de forma y η_i es el predictor lineal asociado a la variable Y_i. En particular se usa esta transformación de la v.a $Y_i^\star$ por Y_i porque nos permite estimar los parámetros a partir de la fdp de una distribución GEV para máximos, que es la usualmente implementada en los paquetes estadísticos como el INLA.

El predictor lineal η_i es definido de la siguiente forma:

$$\eta_i = \mathbf{X_i^T}\beta + f_i, \tag{3.6}$$

donde el vector β es definido por los coeficientes de la regresión, $[\mathbf{X}_i]_{n\times p}$ representa la matriz de covariables y f_i es el i-ésimo efecto aleatorio espacial en el local s_i, tal que $\mathbf{f}(\mathbf{s}) = (f(_1), \ldots, f(_n))^T \sim N(\mathbf{0}, \mathbf{\Sigma})$ con la función de covarianza Matérn con elementos, $\Sigma_{ij} = \sigma^2\rho(d_{ij})$, donde d_{ij} es la distancia euclidiana entre dos estaciones s_i y s_j, σ^2 es la varianza marginal del campo gaussiano y $\rho(d_{ij}) = \dfrac{(\phi d_{ij})^\nu}{2^{(\nu-1)}\Gamma(\nu)}K_\nu(\phi d_{ij})$ es la función de correlación Matérn.

En esta sección definiremos la estructura del modelo geoestadístico gaussiano latente para datos geoestadísticos. De acuerdo a Blangiardo et al. (2013), cuando se trata con datos georeferenciados, un método computacionalmente efectivo es usar el enfoque de SPDE, propuesto por Lindgren et al. (2011). Este consiste en representar un proceso espacial continuo, en el caso de la presente tesis, un campo gaussiano (GF) con función de covarianza Matérn, como un campo gaussiano markoviano (GMRF) discretamente indexado, mediante una representación de funciones base, definida en una triangulación de dominio $\mathcal{D}$, de la siguiente forma

$$\tilde{f}(s) \quad = \quad \sum_{g=1}^{m} \varphi_g(s) f_g^*, \tag{3.7}$$

donde m es el número de vértices de la triangulación, $\varphi_g(s)$ es el conjunto de funciones base y f_g^* son pesos con distribución gaussiana. Lindgren et al. (2011) demostraron que el vector de funciones base $\mathbf{f}^* = (f_1^*, f_2^*, \ldots, f_m^*)^T$ es un GMRF con matriz de precisión $\mathbf{Q}_{\mathbf{f}^*}$ llena de ceros que depende de una función de covarianza Matérn con parámetros ϕ y σ.

Dada la representación del campo gaussiano (GF) de la ecuación (3.7), el predictor lineal de la ecuación (3.6) puede ser reescrito como

$$\begin{aligned} \tilde{\eta}_i &= \beta_0 + \sum_{p=1}^{P} \beta_p x_{ip} + \sum_{g=1}^{m} \varphi_g(s) f_g^* \\ &= \beta_0 + \sum_{p=1}^{P} \beta_p x_{ip} + \sum_{g=1}^{m} A_{ig} f_g^*, \end{aligned}$$

donde A_{ig} es el elemento genérico de la matriz dispersa $\mathbf{A}$ que mapea el GMRF f^* desde los m vertices de la triangulación hasta las n estaciones observadas. La dimensión resultante de la matriz $\mathbf{A}$ esta dada por el número de estaciones observadas y por el número de nodos de la matriz $(n \times m)$. Así, para la implementación del modelo geoestadístico gaussiano latente en R-INLA, se reemplaza la ecuación (3.6) para los n predictores lineales $\eta = (\eta_1, \ldots, \eta_n)^T$ por la siguiente:

$$\tilde{\eta} \quad = \quad \mathbf{1}\beta_0 + \mathbf{X}^{\mathbf{T}}\beta + \mathbf{A}f^*,$$

donde $\mathbf{1}$ es un vector de unos y $\mathbf{X}$ es la matriz de covariables $P \times n$.

De acuerdo a la sección 2.4.1, se define el vector del campo gaussiano latente como $\mathbf{w} = (\tilde{\mathbf{f}}, \beta)$ donde $\tilde{\mathbf{f}} = \mathbf{A}\mathbf{f}^* = (\tilde{f}_1, \ldots, \tilde{f}_n)^T \sim \text{PG}(0, \tilde{\Sigma})$, y el vector de hiper-parámetros $\theta = (\bar{\theta} = (\tau, \xi), \theta^* = (\phi, \sigma))$.

Recopilados los componentes latentes de interés para la inferencia y asumiendo independencia condicional entre las $\mathbf{Y_i}$, dados $\mathbf{w}$ y $\bar{\theta}$, las Y_i's siguen una distribución GEV, luego la función de verosimilitud viene dada por:

$$p(\mathbf{y} \mid \mathbf{w}, \bar{\theta}) = \prod_{i=1}^{n} f_{Y_i}(y_i \mid \tilde{\eta}_i, \xi, \tau),$$

la cual es definida en la ecuación 3.5 dependiendo de $\xi = 0$ o $\xi > 0$ y $\tilde{\eta}_i = \mathbf{X_i^T}\beta + \tilde{f}_i$.

3.2.1. Inferencia bayesiana bajo el enfoque INLA

Asumiendo, que sobre n datos observados $\mathbf{y}$ provienen de una distribución generalizada de valores extremos máximos (GEV) con parámetros η, $\theta_1 = \tau$ y $\theta_2 = \xi$. Así mismo, siendo η un predictor lineal asociado a un vector de coeficientes β, p covariables y efectos espaciales $\tilde{f}$ distribuidos bajo un proceso gaussiano SPDE (Lindgren y Rue, 2015) con parámetros $\theta_3 = \phi$ y $\theta_4 = \sigma$, se tendría lo siguiente

$$\begin{aligned}
\theta &\sim p(\theta_k) && , \text{ para } k = 1, 2, 3, 4. \\
\tilde{\mathbf{f}} \mid \theta^* &\sim \text{PG}\left(0, \tilde{\mathbf{\Sigma}}\right) && , \text{ donde } \tilde{f} = \{\tilde{f}_i\}_{i=1}^n, \quad \theta^* = (\theta_3, \theta_4) \\
\tilde{\eta}_i &= X_i^T\beta + \tilde{f}_i && \\
\mathbf{y} \mid \tilde{f}, \beta, \bar{\theta} &\sim p(\mathbf{y} \mid \tilde{\eta}, \bar{\theta}) && , \text{ donde } \bar{\theta} = (\theta_1, \theta_2).
\end{aligned}$$

Bajo el criterio de estimación INLA, el modelo gaussiano latente jerárquico tiene la siguiente forma

$$\begin{aligned}
\theta &\sim p(\theta_k) && , \text{ para } k = 1, 2, 3, 4. \\
f^* \mid \theta^* &\sim N\left(0, Q_{f^*}^{-1}(\theta^*)\right) && , \text{ donde } dim(f^*) = m, \theta^* = (\theta_3, \theta_4) \\
\tilde{\eta}_i &= X_i^T\beta + A_i f^* && \\
\mathbf{y} \mid \mathbf{w}, \bar{\theta} &\sim p(\mathbf{y} \mid \tilde{\eta}, \bar{\theta}) && , \text{ donde } \bar{\theta} = (\theta_1, \theta_2).
\end{aligned}$$

donde θ está compuesta por los hiperparámetros $\bar{\theta}$ y θ^*, para f^* se asume una normal multivariada a priori con medias cero y matriz de precision $Q(\theta^*)$, para el vector de coeficientes β se asume una distribución gaussiana a priori con parámetros conocidos, $\tilde{\eta}_i$ es el predictor lineal y $\mathbf{w}$ es el campo gaussiano latente.

Entonces, el modelo completo para la inferencia bayesiana está definido de la siguiente forma

$$\begin{aligned}
Y_i \mid \mathbf{w}, \bar{\theta} &\sim GEV\left(\tilde{\eta}_i, \tau, \xi\right) \\
f^* \mid \theta^* &\sim N\left(0, Q_{f^*}(\phi, \sigma)^{-1}\right) \\
p(\beta_0) &\propto 1 \\
\beta_l &\sim N\left(0, 0.00001\right); l = 1, \ldots, p-1 \\
\log(\tau) &\sim LogGamma(1, 0.000001) \\
\xi &\sim N(0, 25) \\
p(\sigma, \phi) &= \tilde{\lambda_1}\tilde{\lambda_2}\phi^{-2}\exp(-\tilde{\lambda_1}\phi^{-1} - \tilde{\lambda_2}\sigma),
\end{aligned}$$

donde $\tilde{\lambda_1} = -\log(0.5)0.7$ y $\tilde{\lambda_2} = -\log(0.5)$, además INLA define por defecto una priori impropia para β_0 y distribuciones gaussiana a prioris no informativa para el vector β_p. Para la definición de las a prioris de los hiperparámetros τ y σ, se consideraron a prioris no informativas y para los hiperparámetros ϕ y σ serán estimados por la PC prior conjunta definido en el teorema 2.4.1, de tal forma que no sea informativa por lo que se asignan probabilidades de 0.5 en cada intervalo de confianza de ρ y σ sobre ciertas cotas de valores $\phi_0 = 0.7$ y $\sigma_0 = 1$, respectivamente, es decir, $P(\phi < 0.7) = 0.5$, $P(\sigma > 1) = 0.5$. Según

Fuglstad et al. (2017), estudios indican que buenas coberturas son obtenidas cuando σ_0 es 2.5 a 40 veces el verdadero valor de σ y ϕ_0 es 1/10 a 1/2.5 veces el verdadero valor de ϕ.

Luego, haciendo uso de la ecuación (2.21), la función de densidad a posteriori conjunta del campo gaussiano latente $\mathbf{w} = \{\tilde{f}, \beta\}$ y los hiperparámetros θ, toma la forma siguiente

$$\begin{aligned}
p(\mathbf{w}, \theta \mid \mathbf{y}) &\propto p(\theta)p(\mathbf{w} = \{\tilde{f}, \beta\} \mid \theta)p(\mathbf{y} \mid \mathbf{w}, \theta) \\
&\propto p(\theta^*)p(\bar{\theta})p(\beta)p(\tilde{f} \mid \theta^*)p(\mathbf{y} \mid \mathbf{w}, \bar{\theta}) \\
&\propto p(\theta^*)p(\bar{\theta})p(\beta)\,|Q_{f^*}(\theta^*)|^{1/2} \exp\left\{-\frac{1}{2}f^{*T}Q_{f^*}(\theta^*)\tilde{f}^* + \sum_{i=1}^{n}\log(f_{Y_i}(y_i \mid \tilde{\eta}_i, \bar{\theta}))\right\}. \\
&\propto p(\sigma, \phi)p(\tau)p(\xi)p(\beta_0)\prod_{l=1}^{p}p(\beta_l)\,|Q_{f^*}(\theta^*)|^{1/2} \\
&\quad \exp\left\{-\frac{1}{2}f^{*T}Q_{f^*}(\theta^*)\tilde{f}^* + \sum_{i=1}^{n}\log(f_{Y_i}(y_i \mid \tilde{\eta}_i, \bar{\theta}))\right\}.
\end{aligned}$$

A partir de esta conjunta a posteriori, se procede a estimar los parámetros utilizando la aproximación INLA, es decir se procede a hallar las distribuciones de las marginales a posteriori para cada elemento del vector de parámetros

$$\begin{aligned}
p(w_i \mid \mathbf{y}) &= \int p(w_i, \theta \mid \mathbf{y})d\theta \\
&= \int p(w_i \mid \theta, \mathbf{y})p(\theta \mid \mathbf{y})d\theta,
\end{aligned} \tag{3.8}$$

y para cada elemento de vector de hiperparámetros

$$p(\theta_k \mid y) = \int p(\theta \mid y)d\theta_{-k}. \tag{3.9}$$

Así necesitamos llevamos a cabo los siguientes pasos:

1. Aproximar la condicional de $p(\theta \mid \mathbf{y})$ con lo que también puede aproximar todas las marginales $p(\theta_k \mid \mathbf{y})$.

2. Aproximar $p(w_i \mid \theta, \mathbf{y})$ el cual se necesita para aproximar las distribuciones marginales a posteriori $p(w_i \mid \mathbf{y})$.

Capítulo 4

Estudio de simulación

4.1. Simulación del modelo GEV para valores extremos mínimos

En esta sección se simulan datos provenientes del modelo definido en la sección 3.1. Sea el vector aleatorio $\mathbf{Y}^\star = (Y_1^\star, \ldots, Y_n^\star)^T$, donde $Y_i^\star$ siguen una distribución GEV$(\tilde{\eta}, \tau, \xi)$ para valores extremos mínimos, en particular para $\xi \neq 0$, tal que la fdp está dada por:

$$f_{Y_i}^\star(y_i^\star; \tilde{\eta}_i, \tau, \xi) = \sqrt{\tau s}\left(1 - \xi\sqrt{\tau s}\,(y_i^\star - \tilde{\eta}_i)\right)^{-1-1/\xi} \exp\left\{-\left(1 - \xi\sqrt{\tau s}\,(y_i^\star - \tilde{\eta}_i)\right)^{-1/\xi}\right\},$$

donde $s > 0$ es un parámetro de escala fijo, τ es el parámetro de precisión, ξ el parámetro de forma y $\tilde{\eta}_i$ es el predictor lineal asociado a la variable Y_i. Para simular los datos se usa la fda del GEV para mínimos,

$$F_{Y_i^\star}(y_i^\star; \tilde{\eta}_i, \tau, \xi) = 1 - \exp\left\{-\left(1 - \xi\sqrt{\tau s}(y_i^\star - \tilde{\eta}_i)\right)^{-1/\xi}\right\} \quad ; \xi \neq 0,\ 1 - \xi\sqrt{\tau s}(y_i^\star - \tilde{\eta}_i) > 0.$$

En el anexo B se muestra cómo se simulan datos provenientes de la distribución GEV para valores extremos mínimos. Como se explicó en el capítulo 2.2.2 de forma equivalente se puede asumir que la v.a. $Y_i = -Y_i^\star$ representa el máximo de las $-Y_i^\star$. Luego se puede asumir que Y_i sigue una distribución generalizada de valores extremos máximos, $Y_i \sim \text{GEV}(\eta_i, \tau, \xi),\ \forall\ i = 1, \ldots, n$. En particular se usa esta transformación de la v.a $Y_i^\star$ por Y_i porque nos permite estimar los parámetros a partir de la fdp de una distribución GEV para máximos, que es lo que se propone en esta tesis, para más detalles ver la sección 3.1.

El cuadro 4.1 muestras las estimaciones a posteriori, intervalos de credibilidad y la desviación estándar de cada parámetro de la simulación de una distribución generailizada para valores mínimos. Podemos observar que las medias a posteriori se acercan a los valores originales de los parámetros, el parámetro original se encuentra dentro del intervalo de credibilidd al 95 % y los intervalos de credibilidad no tienen al valor 0.

En resumen, con esta simulación se verifica que los datos provenientes de una distribución GEV para mínimos, puede ser modelada por una distribución GEV para máximos, siempre que se tome el valor negativo de la variable original. Sin pérdida de generalidad, a continuación procedemos a evaluar el ajuste del modelo geoestadístico para variables aleatorias con distribución GEV para máximos, dado que la estimación para el modelo geoestadístico para variables aleatorias con distribución GEV para mínimos es un caso particular del modelo geoestadístico para variables aleatorias con GEV para máximos.

Cuadro 4.1: Media a posteriori, desviación estandar a posteriori e intervalo de credibilidad (al 95 %), para valores extremos mínimos.

Parámetro	**Original**	**Media**	**Desviación estándar**	**Intervalos de credibilidad**		
				2.5 %	**50 %**	**97.5 %**
β_0	17	17.0121	0.0235	16.9658	17.0122	17.0582
β_1	-3	-3.0112	0.021	-3.0524	-3.0112	-2.9701
τ	2	2.1479	0.1042	1.9509	2.1452	2.3605
ξ	-0.3	-0.2825	0.0181	-0.3173	-0.2829	-0.2463

4.2. Simulación del modelo geoestadístico GEV para valores extremos máximos usando SPDE

En esta sección se simulan 200 coordenadas en el cuadrado unitario D $=[0,1]\times[0,1]\in\Re^2$ (Figura 4.1), las cuales se usan para la simulación de los datos provenientes del modelo geoestadístico de valores extremos máximos definido en el capítulo 3. Luego se estiman los parámetros mediante inferencia bayesiana usando $\mathbf{R-INLA}$. Se consideran diferentes escenarios para evaluar la efectividad de la estimación de los parámetros del modelo en estudio.

En particular se asume que el efecto espacial originalmente sigue un proceso gaussiano

$$f(s)\sim PG(0,\Sigma),$$

donde Σ es la función de covarianza Mátern, que será aproximada numéricamente, dado su costo computacional, por ecuaciones diferenciales parciales estocásticas (SPDE) con hiperparámetros de rango ϕ y desviación estándar σ definido en la sección 2.3.3. Los valores que utilizaremos para los hiperparámetros del campo espacial serán:

- Rango de correlación para el campo espacial $(r_*)=0.7$, note que $r_*=\dfrac{\sqrt{8\nu}}{\phi}$, donde $\nu=2$.
- Desviación estándar para el campo espacial (σ)= 2.

Luego el proceso espacial gaussiano simulado usando SPDE es definido como

$$\tilde{f}(s)=A_i*f^*,$$

donde f^* es el campo espacial gaussiano que contiene la matriz de vértices de la triangulación, cuya solución, bajo SPDE, es definida en la ecuación (2.16), y A_i es una matriz de pesos determinada por $[\varphi_g(s_i)]_{g=1}^m$ (2.15), para todo $i=1,\ldots,n=200$ estaciones. Por tanto, A_i contiene la proyección de las 200 locaciones simuladas y los 461 vértices de la triangulación. El proceso espacial gaussiano simulado $(\tilde{f}_i)$, considerando los parámetros anteriores, se observa en la Figura 4.1.

Para la simulación de los datos asumimos que estos provienen de una distribución generalizada de valores extremos máximos, $Y_i\sim GEV(\eta_i,\tau,\xi)$, cuya fdp fue definida en la ecuación (3.5). El predictor lineal es definido como:

$$\eta_i \;=\; \beta_0+\beta_1X_i+\tilde{f}_i,$$

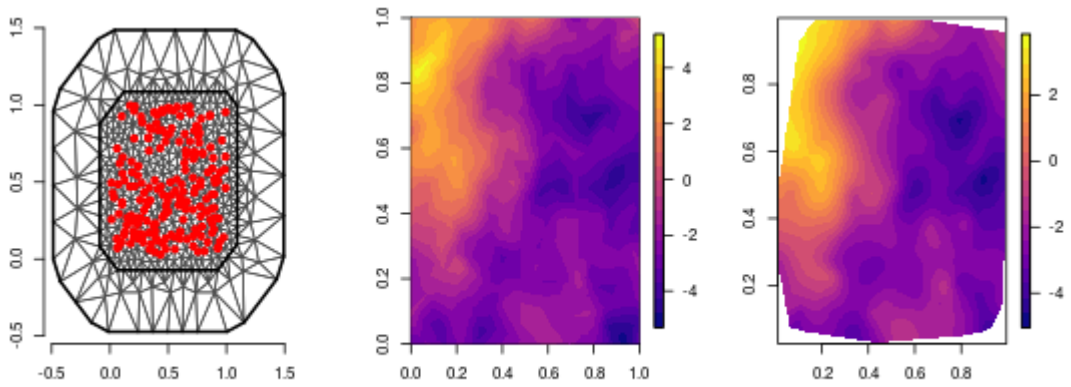

Figura 4.1: Izquierda: Construcción de la malla compuesta por 461 triangulaciones, los puntos rojos son las 200 locaciones simuladas. Intermedio: Campo espacial simulado (f^*). Derecha: Campo espacial gaussiano $\tilde{f}(s)$ simulado.

para $i = 1, \cdots, n$, β_0 es el intercepto, β_1 es el coeficiente de regresión, X_i es una variable explicativa definida en el local i, y los $\tilde{f}_i$ son los efectos espaciales simulados y estructurado en la sección anterior.
Se simula datos de la covariable tal que $X_i \overset{iid}{\sim}$ Normal$(0, 1)$. Los valores considerados para la simulación del predictor lineal son: $\beta_0 = -3$, $\beta_1 = 6$, $\tau = 2$, parámetro de precisión de la distribución GEV y $\xi = -0.3$, parámetro de forma de la distribución GEV.

En la Figura 4.2 se observa los valores puntuales simulados del efecto espacial (derecha) y el histograma de los valores Y_i simulados (izquierda), el cual es ligeramente asimétrico a la izquierda, debido al parámetro ξ negativo.

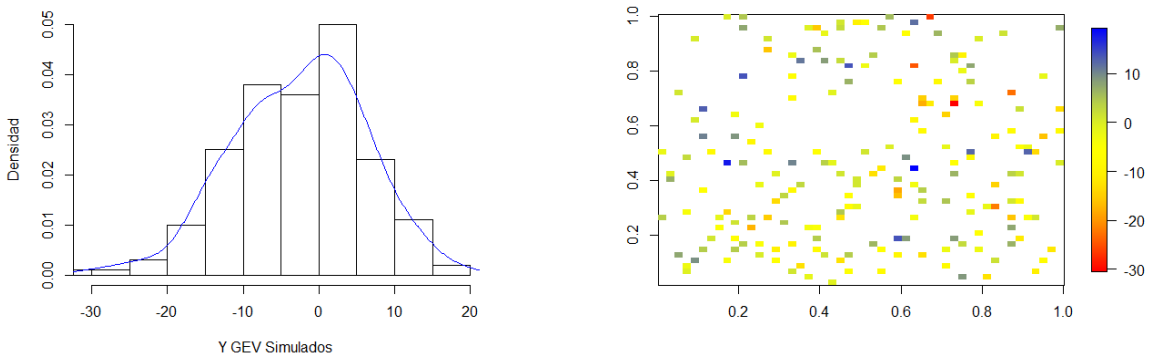

Figura 4.2: Izquierda: Histograma de Y_i simulados con efecto espacial. Derecha: Datos simulados con efecto espacial sobre las locaciones simuladas

4.2.1. Estimación de los parámetros

Para estimar los parámetros se consideró la distribución a posteriori conjunta definida en la ecuación (5.2), asumiendo un modelo bayesiano jerárquico con distribuciones a priori definidas en la Sección 3.2.1. El cuadro 4.2 muestra las estimaciones a posteriori, intervalos de credibilidad y la desviación estándar a posteriori de cada parámetro. Podemos observar

que las medias a posteriori se acercan a los valores originales de los parámetros y que los intervalos de credibilidad contienen los valores reales de los parámetros, siendo evidencia de una estimación adecuada de los parámetros del modelo ajustado. Asimismo, estos resultados se comprueban en la Figura 4.3 se muestran las densidades marginales a posteriori de los coeficientes de regresión (β_0, β_1) y de los hiperparámetros (τ, ξ, r_*, σ) obtenidos bajo el modelo bayesiano jerárquico geoestadístico con distribución generalizada de valores extremos máximos.

Cuadro 4.2: Media a posteriori, desviación estandar a posteriori e intervalo de credibilidad (al 95 %).

Parámetro	**Original**	**Media**	**Desviación estándar**	**Intervalos de credibilidad**		
				2.5 %	**50 %**	**97.5 %**
β_0	-3	-4.6169	2.0372	-9.041	-4.6124	-0.4098
β_1	6	5.9804	0.5092	4.9766	5.9815	6.9778
τ	2	2.4700	0.3200	1.8700	2.4600	3.1100
ξ	-0.3	-0.32	0.057	-0.4356	-0.3185	-0.2119
r_*	0.7	0.7201	0.6435	0.1606	0.5296	2.4109
σ	2	2.6761	0.8256	1.3455	2.5868	4.5556

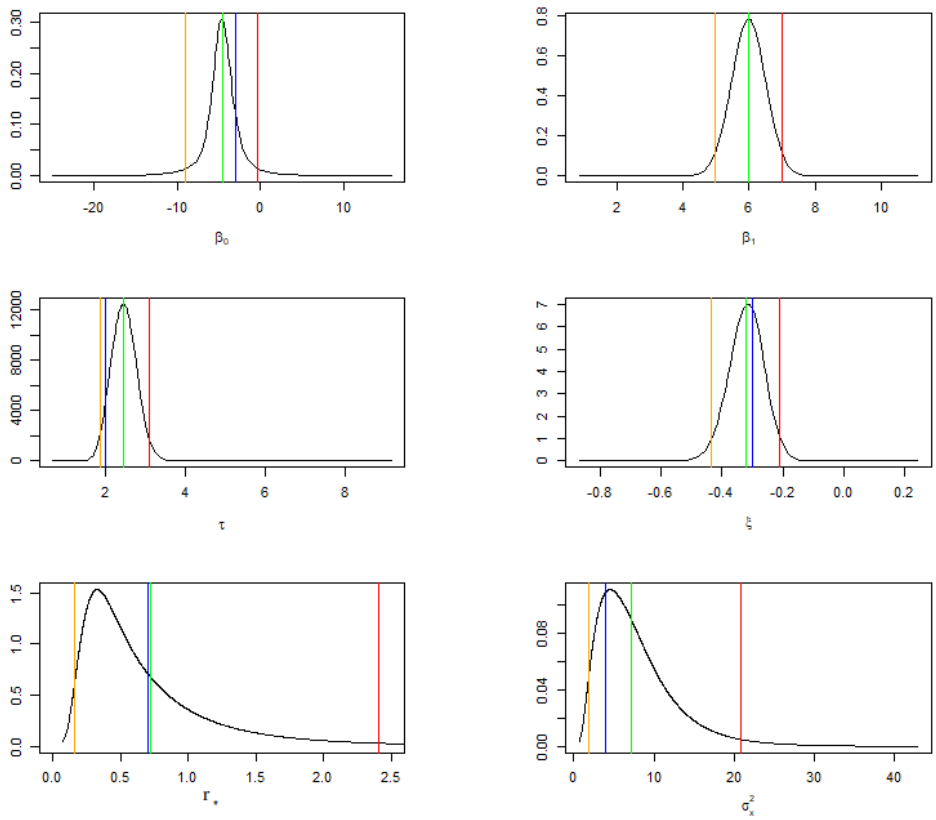

Figura 4.3: Gráficos de las funciones de densidad marginales a posteriori de los hiperparámetros del modelo geoestadístico GEV. La linea azul representa el valor original del parámetro, la línea verde la media estimada, la línea naranja el límite inferior del intervalo de crédibilidad de la estimación del parámetro y la línea roja el límite superior del intervalo de crédibilidad de la estimación del parámetro.

Otro resultado importante, es analizar la estimación de los efectos espaciales a través de la media a posteriori, la desviación estándar a posteriori, el límite inferior y superior de los intervalos de credibilidad al 95 % del campo espacial proyectado estimado $\tilde{f}(s)$ en toda el área de estudio. En la Figura 4.4 se muestran los resultados obtenidos a partir de los hiperparámetros (ϕ, σ) estimados. Según estos resultados se puede observar que la estimación de la media a posteriori del campo espacial $\tilde{f}(s)$ es muy similar al campo espacial proyectado original de la Figura 4.1.

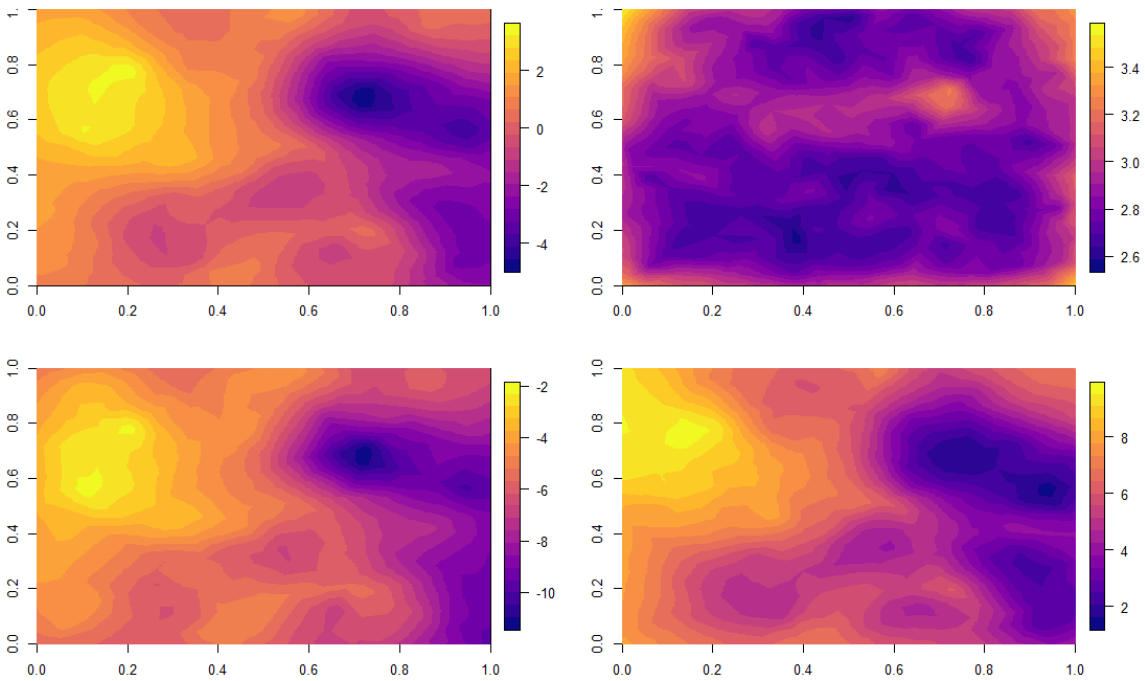

Figura 4.4: Superior Izquierda: Media a posteriori del campo espacial proyectado en las 200 locaciones simuladas. Superior Derecha: Desviación estándar a posteiori del campo espacial proyectado en las 200 locaciones simuladas. Inferior Izquierda: Límite inferior del intervalo de credibilidad al 95 % del campo espacial proyectado en las 200 locaciones simuladas. Inferior Derecha : Límite superior del intervalo de credibilidad al 95 % del campo espacial proyectado en las 200 locaciones simuladas

Finalmente en la Figura 4.5 se muestran la media estimada y sus intervalos de credibilidad al 95 % de los valores simulados bajo una distribución GEV con efecto espacial.

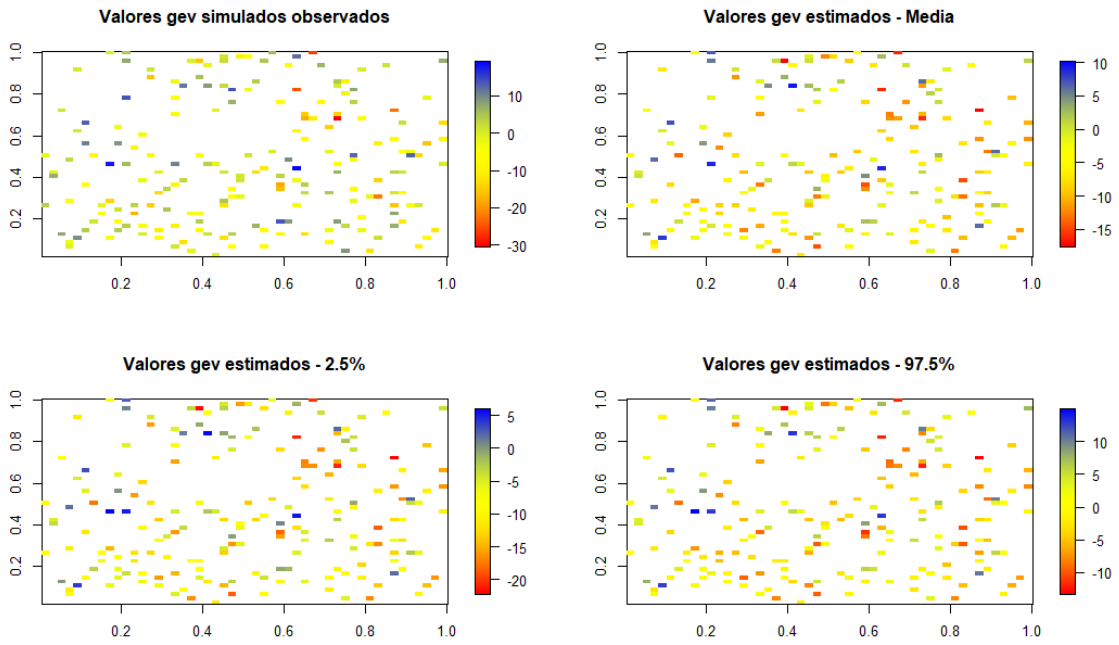

Figura 4.5: Superior Izquierda: Valores GEV simulados, Superior Derecha: Media estimada, Inferior Izquierda: Límite inferior del intervalo de credibilidad al 95 % de los valores GEV simulados, Inferior Derecha : Límite superior del intervalo de credibilidad al 95 % de los valores GEV simulados

Como la distribución para modelar los datos es la GEV para máximos, este modelo en particular debe estimar mejor los máximos valores extremos, lo cual es evidente analizando estos resultados.

4.2.2. Simulación e inferencia bayesiana del modelo geoestadístico GEV para máximos usando distintas réplicas muestrales

En esta sección se detallan los resultados obtenidos al simular una cantidad de muestra n replicándola m veces, obteniéndose una matriz de proyección A de dimensión $(n \times m, k \times m)$ que será usada para la inferencia, donde k son los vértices proyectados de las matriz de precisión Q.

Como se logra observar para una muestra fija de $n = 200$, si el número de réplicas es mayor, se obtienen mejores estimaciones y menores desviaciones estándar en cada de una de los parámetros. Por otro lado, al incrementar el número de estaciones n (de 200 a 400) y el número de vértices k (de 622 a 972) y dejando fijo el número de réplicas ($m = 5$ y $m = 20$) se logra una ligera mejora en las estimaciones de los hiperparámetros.

Cuadro 4.3: Media a posteriori, desviación estándar a posteriori, intervalos de credibilidad (al 95 %) sobre una muestra de tamaño n=200 y k=622 vértices.

Número de réplicas	**Parámetro**	**Original**	**Media**	**Desviación estándar**	**Intervalos de credibili**		
					2.5 %	**50 %**	**97**
m=5	β_0	-3	-4.0151	0.6985	-5.4282	-4.0161	-2
	β_1	6	5.6574	0.2106	5.2445	5.6572	6.
	τ	2	2.1300	0.1100	1.9200	2.1300	2.
	ξ	-0.3	-0.2947	0.0191	-0.3319	-0.2949	-0
	ϕ	0.7	0.67	0.2726	0.2777	0.6245	1.
	σ	2	2.1602	0.3684	1.5172	2.1332	2.
m=20	β_0	-3	-3.0293	0.2933	-3.6118	-3.0282	-2
	β_1	6	5.8087	0.1072	5.5982	5.8087	6.
	τ	2	1.9900	0.0500	1.8800	1.9800	2.
	ξ	-0.3	-0.312	0.0105	-0.3324	-0.3121	-0
	r_*	0.7	0.5516	0.113	0.3685	0.538	0.
	σ	2	2.2754	0.2066	1.8952	2.2664	2.

Cuadro 4.4: Media a posteriori, desviación estándar a posteriori, intervalos de credibilidad (al 95 %) sobre una muestra de tamaño n=400 y k=972 vértices.

Número de réplicas	**Parámetro**	**Original**	**Media**	**Desviación estándar**	**Intervalos de credibili**		
					2.5 %	**50 %**	**97**
m=5	β_0	-3	-4.1620	0.6310	-5.4388	-4.1562	-2
	β_1	6	5.5817	0.1473	5.2926	5.5817	5.
	τ	2	2.0400	0.0700	1.9000	2.0400	2.
	ξ	-0.3	-0.3119	0.0138	-0.3387	-0.3120	-0
	ϕ	0.7	0.6228	0.1672	0.3551	0.6026	1.
	σ	2	2.3508	0.3136	1.7997	2.3282	3.
m=20	β_0	-3	-3.3204	0.3154	-3.9420	-3.3206	-2
	β_1	6	5.9206	0.0761	5.7713	5.9206	6.
	τ	2	1.9700	0.0300	1.9100	1.9700	2.
	ξ	-0.3	-0.3119	0.0004	-0.3127	-0.3120	-0
	r_*	0.7	0.7029	0.0510	0.6113	0.6993	0.
	σ	2	2.1769	0.1474	1.8793	2.1817	2.

Capítulo 5

Aplicación

En este capítulo se aplica el modelo propuesto sobre un conjunto de datos que está compuesto por las temperaturas mínimas en el Perú.

5.1. Descripción de los datos

La base de datos ha sido extraída del Servicio Nacional de Meteorología e Hidrología del Perú- SENAHMI (https://www.senamhi.gob.pe/?&p=estaciones), en donde a través de estaciones meteorológicas han sido medidas las temperaturas mínimas en todo el Perú. En particular, se seleccionó para esta tesis los datos correspondientes al periodo de Agosto 2012, encontrándose dentro del periodo de friajes en la región Selva del Perú, cuyos departamentos más afectados por este fenómeno son Madre de Dios, Puno, Ucayali, Huánuco, San Martín y Loreto.

Este conjunto de datos está compuesta por los campos descritos en el Cuadro 5.1

Cuadro 5.1: Descripción de los campos de la base de datos.

Campo	**Descripción del campo**
Temperatura mínima	Temperatura mínima diaria de cada estación medida en grados centígrados (C).
Temperatura máxima	Temperatura máxima diaria de cada estación medida en grados centígrados (C).
Latitud	Ángulo entre el plano ecuatorial y el segmento que une el punto terrestre al centro de la tierra ().
Longitud	Ángulo entre el semiplano que pasa por el meridiano de Greenwich con el semiplano del eje de la tierra que pasa por el punto terrestre ().
Altitud	Distancia en metros sobre el nivel del mar ($m.s.n.m$).
Precipitación	Medida en milímetros (mm) de agua caídos por unidad de superficie (m^2).

Los datos están compuestos por 151 estaciones donde se midieron las temperaturas mínimas, siendo esta la variable de interés. Las potenciales covariables con las cuales se intentará explicar la variable de respuesta son la altitud y la precipitación.

5.2. Análisis exploratorio

En la Figura 5.1 se observa el histograma de los datos, donde podemos observar que los datos Y_i tienen un asimetría a la izquierda; sin embargo como ya se explicó en el capítulo 3 la presente tesis utilizará la variable de respuesta $Y_i^\star$ la cual se asume que sigue una distribución generalizada de valores exremos mínimos. Así mismo, la Figura 5.1 muestra a la derecha los datos (Y_i) de las temperaturas mínimas observadas, donde los círculos corresponden a las estaciones.

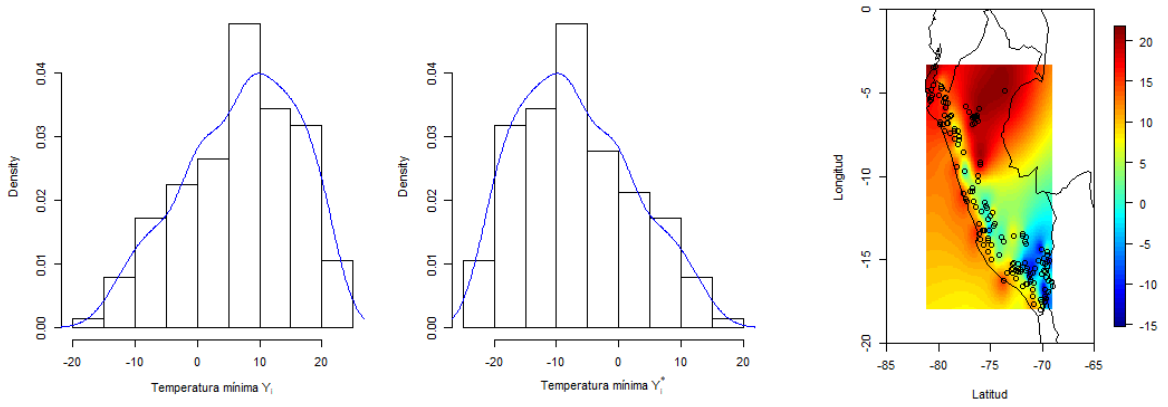

Figura 5.1: Izquieda:Histograma de distribución de las temperaturas mínimas del Perú (Y_i), medidas en Agosto 2012.La línea azul representa la fdp estimada.Intermedio: Histograma de distribución de las temperaturas mínimas del Perú $(Y_i^\star)$ medidas en Agosto 2012.La línea azul representa la fdp estimada. Derecha: Mapa interpolado de las temperaturas mínimas en el Perú observadas en Agosto 2012. Los círculos corresponden a las estaciones metereológicas.

Para evaluar las potenciales convariables, se analizan las correlaciones de la temperatura mínima $(Y_i^\star)$, con cada una de las variables de la base de datos con el objetivo de identificar las posibles covariables que formen parte de los modelos que serán propuestos. En la Figura 5.2 en la parte superior se muestran los gráficos de dispersión de las variables altitud y precipitación, observándose que en la primera existe una relación lineal positiva con la variable de respuesta. Este resultado tiene sentido en el Perú, dado que las regiones con mayor altitud tienen temperaturas mínimas extremas. Por otro lado para el caso de la variable precipitación se observa una ligera relación negativa. Para encontrar una mejor asociación lineal, se transformó la variable precipitación a un polinomio de grado 2 y además se usó una transformación logarítmica, en la cual se observa una mejor asociación lineal con la variable temperaturas mínimas. En la Figura 5.3 se muestran las correlaciones de las covariables con la variable de respuesta $(Y_i^\star)$ y además la correlación entre ellas. Según estos resultados, se continua el modelamiento tomando como covariables la altitud, la precipitación y su transformación logarítmica.

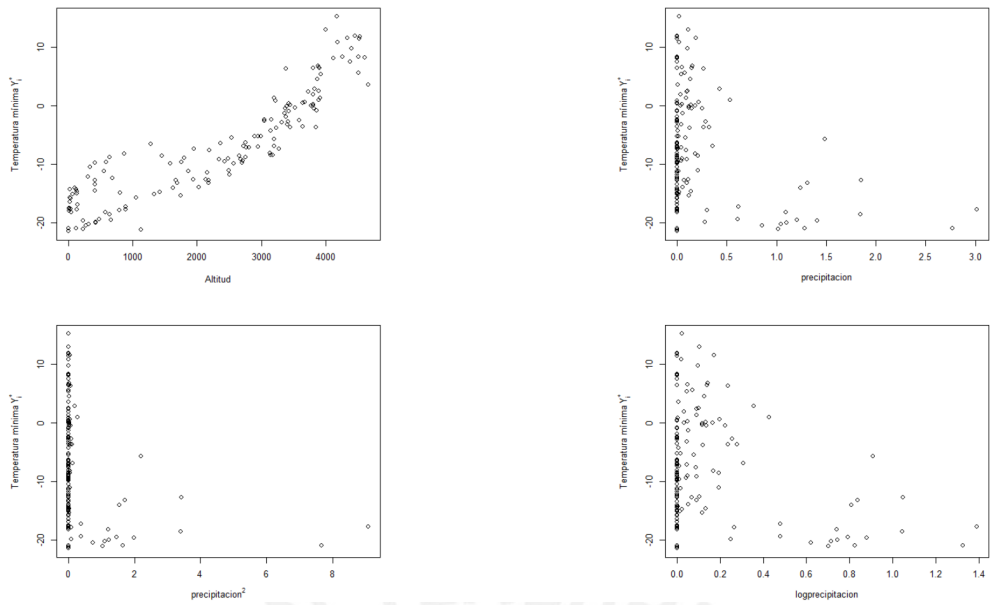

Figura 5.2: Izquierda superior: Diagrama de dispersión de las altitudes vs. las temperatura mínimas ($Y_i^\star$). Derecha superior: Diagrama de dispersión de las precipitaciones vs. las temperatura mínimas ($Y_i^\star$). Izquierda inferior: Diagrama de dispersión de las precipitaciones al cuadrado vs. las temperatura mínimas ($Y_i^\star$). Derecha inferior: Diagrama de dispersión del logaritmo de las precipitaciones vs. las temperatura mínimas ($Y_i^\star$).

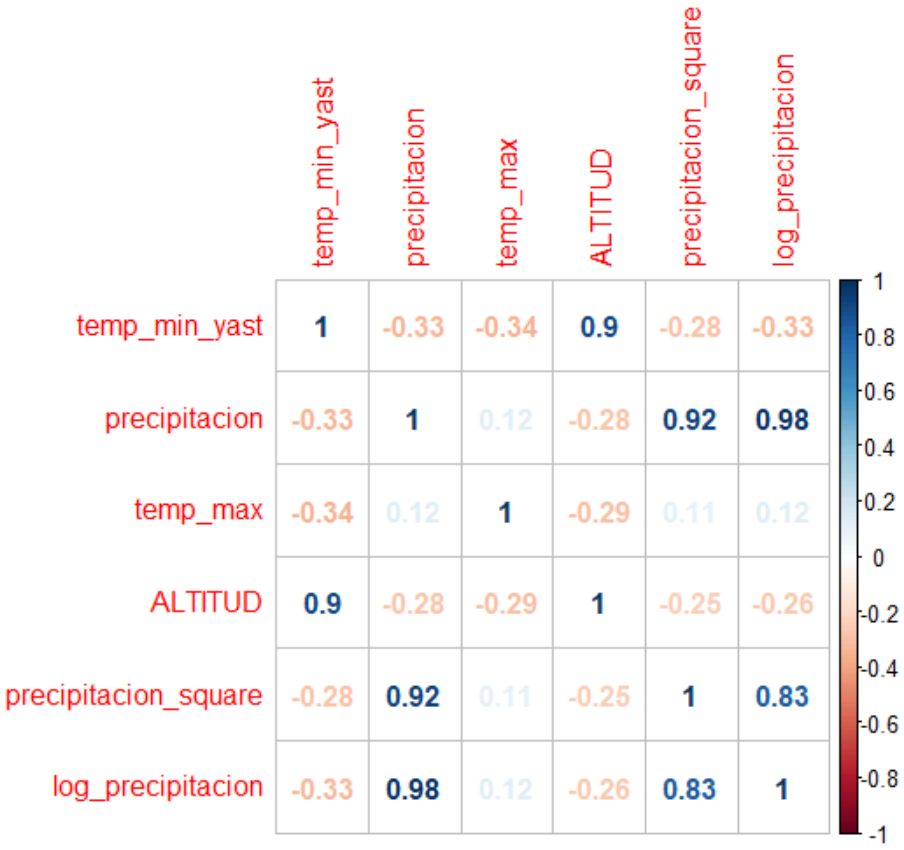

Figura 5.3: Gráfico de correlaciones entre las covariables con las temperaturas mínimas.

Parte del análisis exploratorio también consiste en determinar la existencia de autocorrelación espacial en las temperaturas mínimas en estaciones vecinas. En la Figura 5.4 se observa el semivariograma empírico y teórico (Matérn) de los datos de las temperaturas mínimas. En términos generales, se tiene que cuando la distancia entre las estaciones es pequeña la semivarianza es de menor magnitud, es decir habria menor autocorrelación espacial. Y además observamos que cuando la distancia entre las estaciones es mayor, la semivarianza se incrementa hasta llegar a un umbral el cual es conocido como meseta. En el semivariograma se observa una meseta igual a $(\tau^2 + \sigma^2) = 43.1$. Otro componente importante del semivariograma es el rango, que representa la distancia a partir de la cual las temperaturas mínimas no dependen espacialmente entre las estaciones. Cuanto más pequeño es el rango, existe menor autocorrelación espacial. En este caso se observa un rango de aproximadamente 1 grado; por otro lado, podemos observar que el modelo Matérn teórico se ajusta a los datos, el cual es usado en la estimación del campo gaussiano latente.

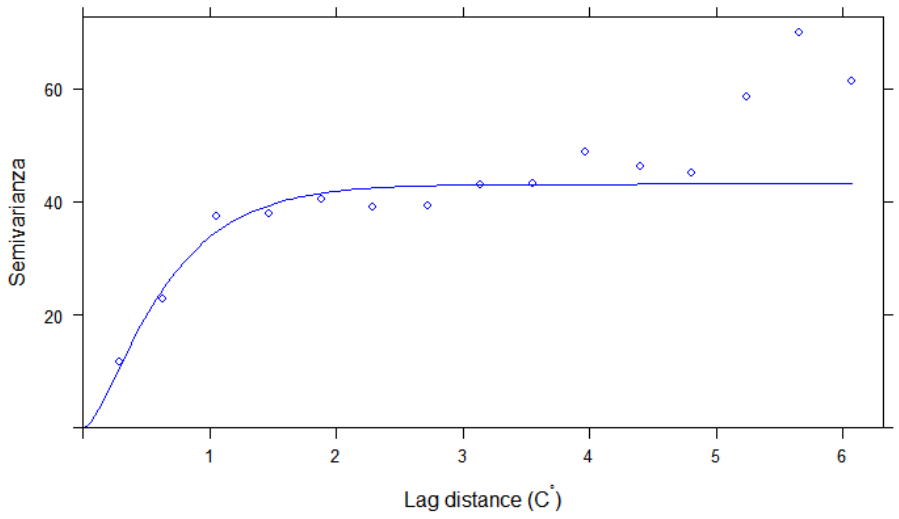

Figura 5.4: Semivariograma de las temperaturas mínimas del Perú medidas en Agosto 2012 bajo un modelo Matérn.

5.3. Construcción del efecto espacial ($\tilde{f}(s)$)

Una vez obtenida la matriz de datos que formarán parte del modelamiento, se crea la estructura de las estaciones, representada por el par (Latitud y Longitud), sobre cierta cantidad de vértices establecidos con el objetivo de generar posteriormente una matriz de proyección A que permita realizar la inferencia del modelo.

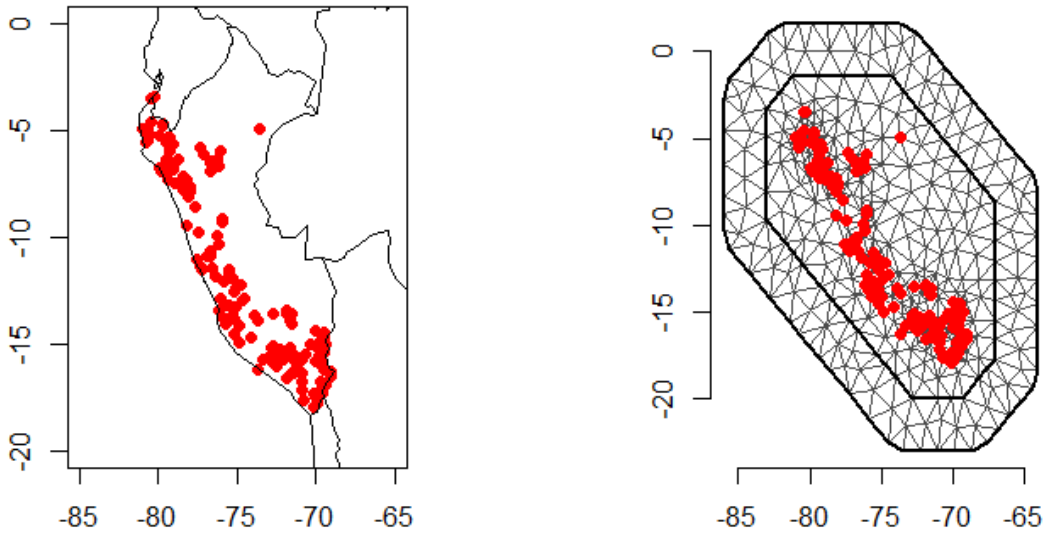

Figura 5.5: Izquierda: Mapa de las coordenadas de las estaciones meteorológicas en el Perú. Derecha: Triangulación de las estaciones meteorológicas del Perú compuesta por 506 vértices de $n = 151$ estaciones

5.4. Modelamiento de los datos y resultados

Para el modelamiento de los datos, sea $Y_i^\star$ v.a. que representa a las temperaturas mínimas, se define $Y_i = -Y_i^\star$ $i = 1, \cdots, n$ en $n = 151$ estaciones. Se asume $Y_i \sim GEV(\eta_i^{(j)}, \tau, \xi)$, donde $s > 0$ es un parámetro de escala fijo, τ es el parámetro de precisión, ξ el parámetro de forma y η_i es el predictor lineal asociado a la variable Y_i. Se proponen tres modelos sin efecto espacial y tres modelos con efecto espacial. El cuadro 5.2 muestra el predictor lineal de cada uno de los modelos propuestos, donde β_0 es el intercepto, $\beta_{altitud}$ es el coeficiente de regresión para la covariable altitud, $\beta_{precipitacion}$ es el coeficiente de regresión para la covariable precipitación, $\beta_{logprecipitacion}$ es el coeficiente de regresión para la covariable precipitación con transformación logarítmica y $\tilde{f}_i$ son los efectos espaciales.

Cuadro 5.2: Predictor lineal de los modelos propuestos.

Modelo	Predictor lineal
GEV-NS1	$\eta_i^{(1)} = \beta_0 + \beta_{altitud} * \text{Altitud}_i$
GEV-S1	$\eta_i^{(2)} = \beta_0 + \beta_{altitud} * \text{Altitud}_i + \tilde{f}_i$
GEV-NS2	$\eta_i^{(3)} = \beta_0 + \beta_{altitud} * \text{Altitud}_i + \beta_{logprecipitacion} * \text{LogPrecipitacion}$
GEV-S2	$\eta_i^{(4)} = \beta_0 + \beta_{altitud} * \text{Altitud}_i + \beta_{logprecipitacion} * \text{LogPrecipitacion} + \tilde{f}_i$
GEV-NS3	$\eta_i^{(5)} = \beta_0 + \beta_{altitud} * \text{Altitud}_i + \beta_{precipitacion} * \text{Precipitacion}$
GEV-S3	$\eta_i^{(6)} = \beta_0 + \beta_{altitud} * \text{Altitud}_i + \beta_{precipitacion} * \text{Precipitacion} + \tilde{f}_i$

El modelo jerárquico para los modelos no espaciales GEVNS1, GEVNS2 y GEVNS3, está definido de la siguiente forma:

$$\begin{array}{rcl} Y_i \mid \mathbf{w}, \theta & \sim & GEV\left(\eta_i, \tau, \xi\right) \\ p(\beta_0) & \propto & 1 \\ \beta_l & \sim & N\left(0, 0.00001\right); l = 1, \ldots, p-1 \\ \log(\tau) & \sim & LogGamma(1, 0.000001) \\ \xi & \sim & N(0, 0.001), \end{array} \tag{5.1}$$

donde η_i difiere para cada modelo según su definición en el cuadro 5.2.

Entonces, el modelo jerárquico para los modelos espacialesGEV-S1, GEV-S2 y GEVS3 está definido de la siguiente forma:

$$\begin{array}{rcl} Y_i \mid \mathbf{w}, \bar{\theta} & \sim & GEV\left(\tilde{\eta}_i, \tau, \xi\right) \\ f^* \mid \theta^* & \sim & N\left(0, Q_{f^*}(\phi, \sigma)^{-1}\right) \\ p(\beta_0) & \propto & 1 \\ \beta_l & \sim & N\left(0, 0.00001\right); l = 1, \ldots, p-1 \\ \log(\tau) & \sim & LogGamma(1, 0.000001) \\ \xi & \sim & N(0, 0.001) \\ p(\sigma, \phi) & = & \tilde{\lambda_1}\tilde{\lambda_2}\phi^{-2}\exp(-\tilde{\lambda_1}\phi^{-1} - \tilde{\lambda_2}\sigma), \end{array}$$

donde $\tilde{\lambda_1} = -\log(0.5)0.7$ y $\tilde{\lambda_2} = -\log(0.5)$, η_i difiere para cada modelo según su definición en el cuadro 5.2 y $\tilde{f} = Af^*$ es una aproximación de un proceso gaussiano con función de covarianza Matérn que depende de los parámetros ϕ y σ.

Sea el campo gaussiano latente $\mathbf{w} = \{\beta\}$ para los modelos no espaciales y $\mathbf{w} = \{\tilde{f}, \beta\}$ para los modelos espaciales y los hiperparámetros $\theta = \{\tau, \xi\}$ para los modelos no espaciales y $\theta = \{\tau, \xi, \phi, \sigma\}$ para los modelos espaciales, entonces la función de densidad a posteriori conjunta toma las formas siguientes:
i) Para modelos no espaciales:

$$p(\mathbf{w}, \theta \mid \mathbf{y}) \propto p(\tau)p(\xi)p(\beta_0)\prod_{l=1}^{p} p(\beta_l)\exp\left\{\sum_{i=1}^{n}\log(f_{Y_i}(y_i \mid \eta_i, \theta))\right\}.$$

ii) Para modelos espaciales:

$$\begin{aligned} p(\mathbf{w}, \theta \mid \mathbf{y}) \propto\ & p(\sigma, \phi)p(\tau)p(\xi)p(\beta_0)\prod_{l=1}^{p} p(\beta_l)\left|Q_{f^*}(\theta^*)\right|^{1/2} \\ & \exp\left\{-\frac{1}{2}f^{*T}Q_{f^*}(\theta)\tilde{f}^* + \sum_{i=1}^{n}\log(f_{Y_i}(y_i \mid \eta_i, \theta))\right\}. \end{aligned}$$

A partir de esta conjunta a posteriori, se procede a estimar los parámetros utilizando la aproximación INLA,

Los cuadros 5.3, 5.4, 5.5, 5.6, 5.7 y 5.8 muestran los resultados obtenidos, la estimación de las medias a posteriori, la desviación estándar a posteriori y los intervalos de credibilidad al

95 % de los parámetros de cada modelo propuesto. Según estos resultados, se puede observar que los modelos con efecto aleatorio $(\eta^{(2)}, \eta^{(4)}, \eta^{(6)})$ tienen menor desviación estandar en la estimación de sus parámetros comparado con los modelos sin efecto aleatorio $(\eta^{(1)}, \eta^{(3)}, \eta^{(5)})$, especialmente en los hiperparámetros (τ, ξ).

Cuadro 5.3: Media,desviación estándar e intervalos de credibilidad al 95 % de los parámetros del modelo propuesto GEV-NS1

Predictor lineal	**Parámetro**	**Media**	**Desviación estándar**	**Intervalos de credibilidad**		
				2.5 %	**50 %**	**97.5 %**
$\eta^{(1)}$	β_0	-20.9407	0.5911	-22.1244	-20.9335	-19.7985
	$\beta_{altitud}$	0.0054	0.0002	0.0049	0.0054	0.0058
	τ	0.0916	0.0125	0.0681	0.0914	0.1170
	ξ	-0.0799	0.0621	-0.2024	-0.0798	0.0423

Cuadro 5.4: Media,desviación estándar e intervalos de credibilidad al 95 % de los parámetros del modelo propuesto GEV-S1

Predictor lineal	**Parámetro**	**Media**	**Desviación estándar**	**Intervalos de credibilidad**		
				2.5 %	**50 %**	**97.5 %**
$\eta^{(2)}$	β_0	-17.6464	0.9193	-19.4081	-17.6618	-15.7849
	$\beta_{altitud}$	0.0043	0.0003	0.0036	0.0043	0.0049
	τ	38.1342	0.0576	27.1919	38.0460	50.2217
	ξ	-0.1539	0.0043	-0.1632	-0.1535	-0.1463
	r_*	1.6241	0.2154	1.2648	1.5997	2.1105
	σ	3.1497	0.4094	2.3684	3.1469	3.9863

Cuadro 5.5: Media,desviación estándar e intervalos de credibilidad al 95 % de los parámetros del modelo propuesto GEV-NS2

Predictor lineal	**Parámetro**	**Media**	**Desviación estándar**	**Intervalos de credibilidad**		
				2.5 %	**50 %**	**97.5 %**
$\eta^{(3)}$	β_0	-20.2559	0.6948	-21.6455	-20.2479	-18.9122
	$\beta_{altitud}$	0.0053	0.0002	0.0048	0.0053	0.0057
	$\beta_{logprecipitacion}$	-2.0121	1.1890	-4.4348	-1.9795	0.2287
	τ	9.1142	0.0126	6.7266	9.1005	11.6347
	ξ	-0.1135	0.0554	-0.2256	-0.1123	-0.0073

Cuadro 5.6: Media,desviación estándar e intervalos de credibilidad al 95 % de los parámetros del modelo propuesto GEV-S2

Predictor lineal	**Parámetro**	**Media**	**Desviación estándar**	**Intervalos de credibilidad**		
				2.5 %	**50 %**	**97.5 %**
$\eta^{(4)}$	β_0	-17.4634	0.9870	-19.3736	-17.4768	-15.4743
	$\beta_{altitud}$	0.0043	0.0003	0.0036	0.0043	0.0049
	$\beta_{logprecipitacion}$	-0.4479	1.2060	-2.7992	-0.4535	1.9298
	τ	39.4341	0.0169	36.5757	39.2534	43.1867
	ξ	-0.1872	0.0012	-0.1894	-0.1873	-0.1849
	r_*	1.4222	0.2876	0.8859	1.4184	2.0003
	σ	3.4614	0.4541	2.7418	3.3982	4.5092

Cuadro 5.7: Media,desviación estándar e intervalos de credibilidad al 95 % de los parámetros del modelo propuesto GEV-NS3

Predictor lineal	**Parámetro**	**Media**	**Desviación estándar**	**Intervalos de credibilidad**		
				2.5 %	**50 %**	**97.5 %**
$\eta^{(5)}$	β_0	-20.3418	0.6846	-21.7107	-20.3341	-19.0174
	$\beta_{altitud}$	0.0053	0.0002	0.0048	0.0053	0.0057
	$\beta_{precipitacion}$	-1.0679	0.6894	-2.4879	-1.0437	0.2144
	τ	9.0904	0.0126	6.6919	9.0804	11.6203
	ξ	-0.1077	0.0569	-0.2225	-0.1064	0.0010

Cuadro 5.8: Media,desviación estándar e intervalos de credibilidad al 95 % de los parámetros del modelo propuesto GEV-S3

Predictor lineal	**Parámetro**	**Media**	**Desviación estándar**	**Intervalos de credibilidad**		
				2.5 %	**50 %**	**97.5 %**
$\eta^{(6)}$	β_0	-17.4006	0.8986	-19.1636	-17.4014	-15.6356
	$\beta_{altitud}$	0.0042	0.0003	0.0037	0.0042	0.0048
	$\beta_{precipitacion}$	-0.4115	0.5849	-1.5666	-0.4092	0.7296
	τ	38.5614	0.0034	37.8776	38.5729	39.1925
	ξ	-0.1977	0.0003	-0.1982	-0.1978	-0.1972
	r_*	1.4845	0.0358	1.4146	1.4844	1.5551
	σ	3.1706	0.0937	2.9672	3.1799	3.3289

Lo mencionado en el párrafo anterior, puede ser corroborado en el cuadro 5.9, en el cual se observa el error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE) de la estimación de cada modelo propuesto lo siguiente, siendo los modelos con efecto aleatorio los que obtienen mejores indicadores y el modelo GEV-S1 el que presenta mejor bondad de ajuste, así mismo todos los parámetros estimados de este modelo son significativos pues no tienen el valor 0 dentro de su intervalo de credidibilidad 95 %.

Cuadro 5.9: Criterio de selección de los modelos propuestos: MSE, RMSE, Marginal Likelihood (ML), DIC, LCPO (Log-CPO) de los diferentes modelos ajustados para la estimación de temperaturas mínimas.

Modelo propuesto	**MSE**	**RMSE**	**ML**	**DIC**	**LCPO**
GEV-NS1	17.95051	4.23680	-452.00	835.71	852.97
GEV-S1	1.55594	1.24737	-433.00	681.40	1334.41
GEV-NS2	17.11627	4.13718	-454.00	834.20	2300.81
GEV-S2	1.47433	1.21422	-439.00	717.11	1598.39
GEV-NS3	17.22873	4.15075	-455.00	834.63	1909.83
GEV-S3	1.89143	1.37529	-446.00	725.53	1455.63

Por otro lado, en la Figura 5.6, los modelos que contienen el efecto espacial (lado derecho) ajustan mejor con respecto a los modelos sin efecto espacial (lado izquierdo), a pesar de que en ambos tipos de modelos se utiliza la distribución generalizada de valores extremos. Cabe resaltar que la cola que representa a las temperaturas mínimas es estimada mejor, debido esencialmente al modelo GEV ajustado. Además, según estos resultados nuevamente el modelo GEV-S2 estima mejor las temperaturas mínimas.

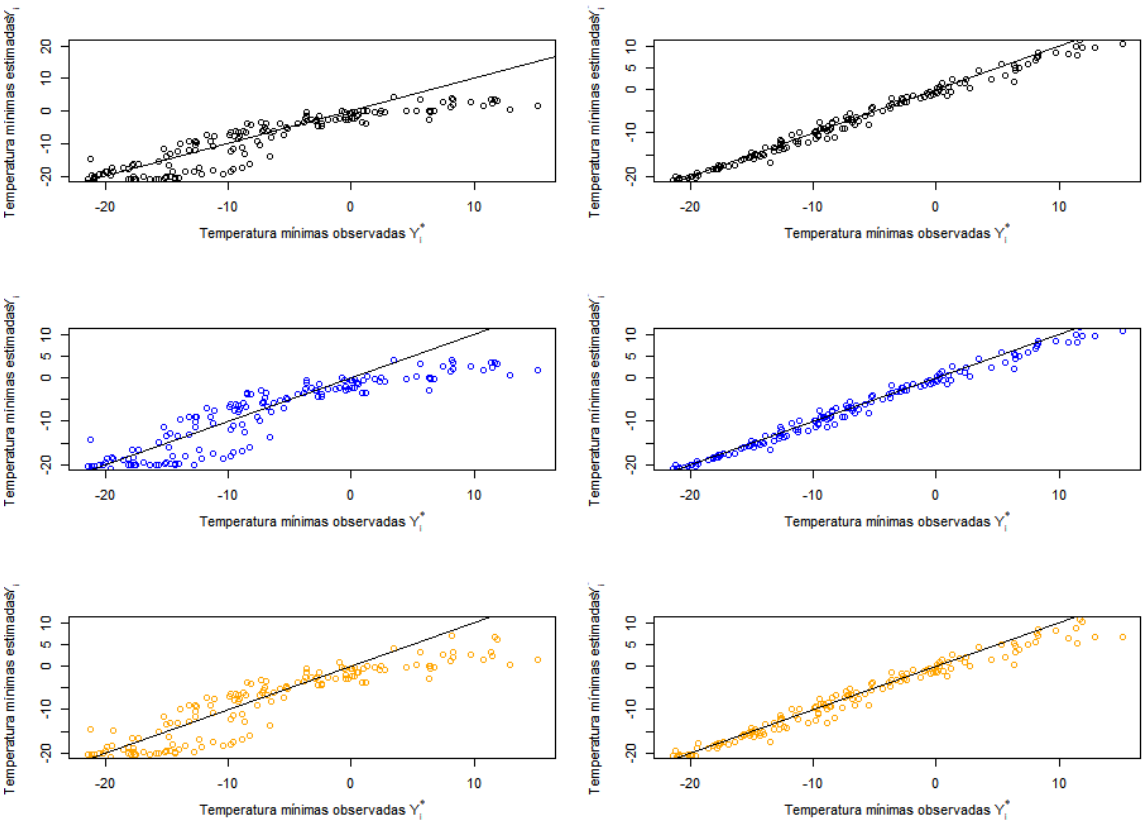

Figura 5.6: Izquierda: Ajustes de modelos GEV sin efecto espacial ($\eta^{(1)}, \eta^{(3)}, \eta^{(5)}$). Derecha: Ajustes de modelos GEV con efecto espacial ($\eta^{(2)}, \eta^{(4)}, \eta^{(6)}$).

Dado que los modelos espaciales tienen mejores estimaciones, menor error y mayor ajuste

a los datos, en la Figura 5.7 se grafica el histograma de las temperaturas mínimas observadas y la densidad de los modelos espaciales. En términos generales, se observa que estos modelos espaciales tienen un buen ajuste de los temperaturas mínimas.

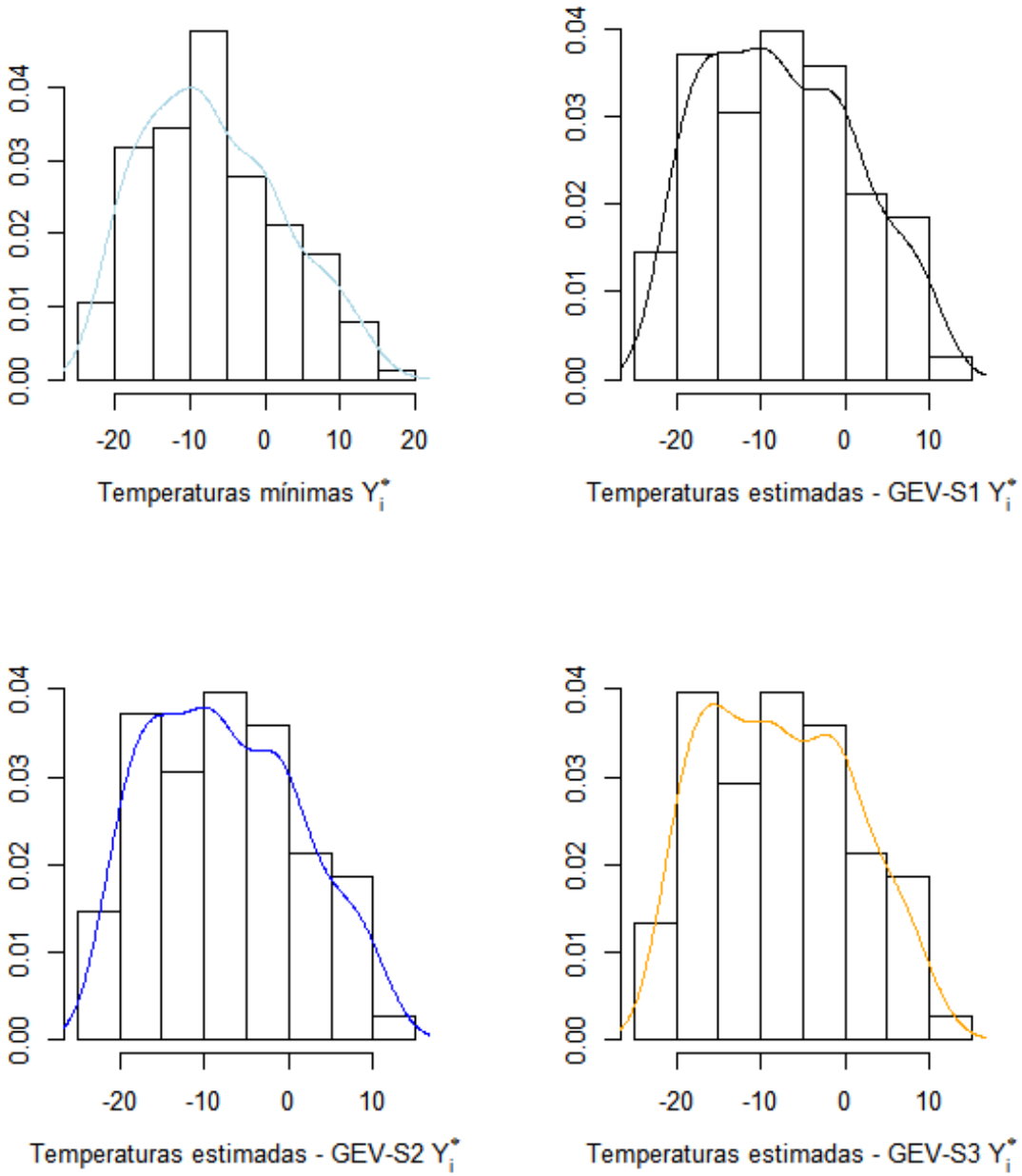

Figura 5.7: Superior Izquierda: Histograma de las temperaturas mínimas medidas en Agosto 2012. Superior Derecha: Histograma con la media estimada por el modelo GEV-S1, Inferior Izquierda: Histograma con la media estimada por el modelo GEV-S2, Inferior Derecha: Histograma con la media estimada por el modelo GEV-S3. En todos los casos la linea representa la función de densidad suavizada.

Según todos los criterios de selección de modelos, el modelo GEV-S2 se ajusta mejor a los datos, por lo tanto se continua el análisis para este modelo. En la Figura 5.8 se muestra las

densidades marginales a posteriori de parámetros e hiperparámetros estimados así como los límites del intervalo de crédibilidad al 95 % del modelo GEV-S2. Para este modelo como $\xi = -0.154 < 0$, la distribución tiene una cola hacia la izquierda, y el mínimo de la temperatura mínima media para cada estación es calculada por $E(Y_i) = -(\eta_i + (\Gamma(1-\xi)-1)\frac{1}{\xi\sqrt{\tau s}})$. Como la estimación a posteriori de la media de τ es grande, entonces el valor esperado de las temperaturas mínimas es aproximadamente igual al predictor lineal. Esto implica por cada incremento en una unidad de la altitud, el mínimo de la temperatura mínima media aumenta en 0.0043 grados.

Con respecto a los parámetros del efecto espacial, el rango efectivo es igual a $r_* = 1.6241$, por lo tanto la temperatura mínima en una estación i depende espacialmente de todas las estaciones que se encuentran a una distancia máxima de 1.62 grados, mientras que la varianza marginal es $\sigma = 3.15$, la cual es mayor comparada a la variabilidad restante obtenida calculando $V(Y_i)$, esto implica que efectivamente hay una autocorrelación espacial entre la temperatura mínima entre estaciones.

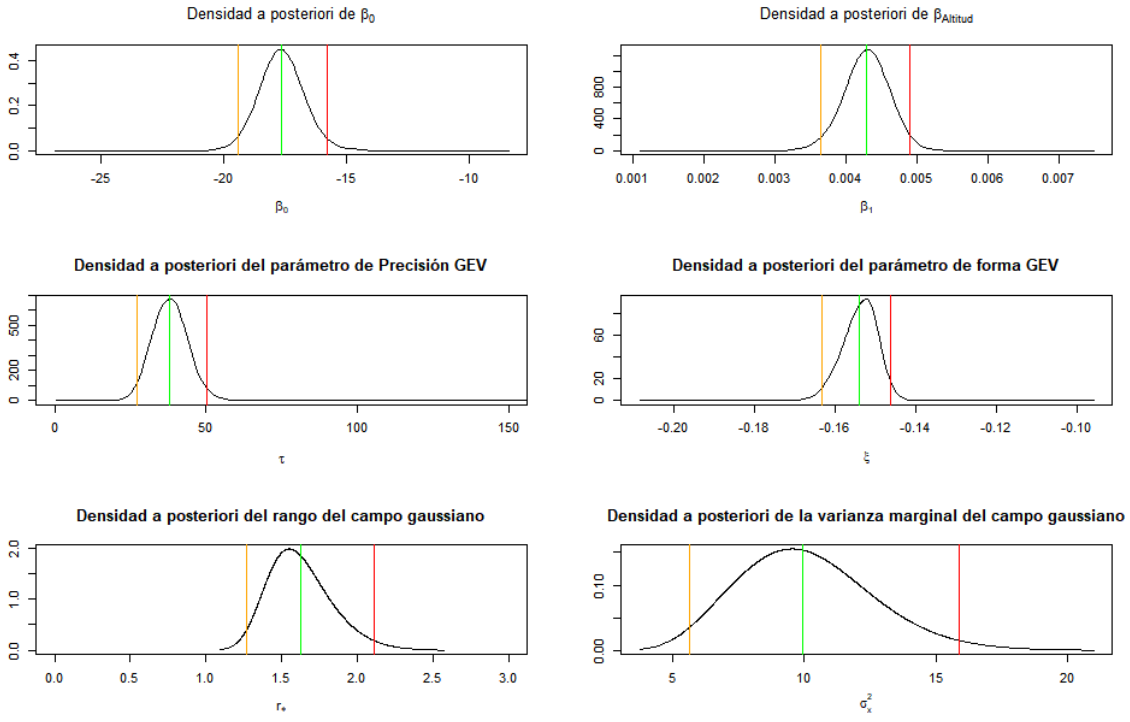

Figura 5.8: Gráficos de densidad de las marginales a posterior de los hiperparámetros del modelo geoestadístico GEV-S1, la línea verde la media estimada, la línea naranja el límite inferior del intervalo de crédibilidad y la línea roja el límite superior del intervalo de crédibilidad.

Por último, en la Figura 5.9 se muestran el comparativo del mapa con los valores reales (Y_i) y el mapa con las predicciones del modelo propuesto GEV-S1, el cual tiene el efecto espacial. Se puede observar que se estima bastante bien las temperaturas mínimas, sobre todo los valores extremos mínimos. No se puede realizar la estimación en la zona norte de la selva porque no hay estaciones metereológicas en uso, y como en estas zonas las temperaturas son más elevadas no son de relevancia en este estudio.

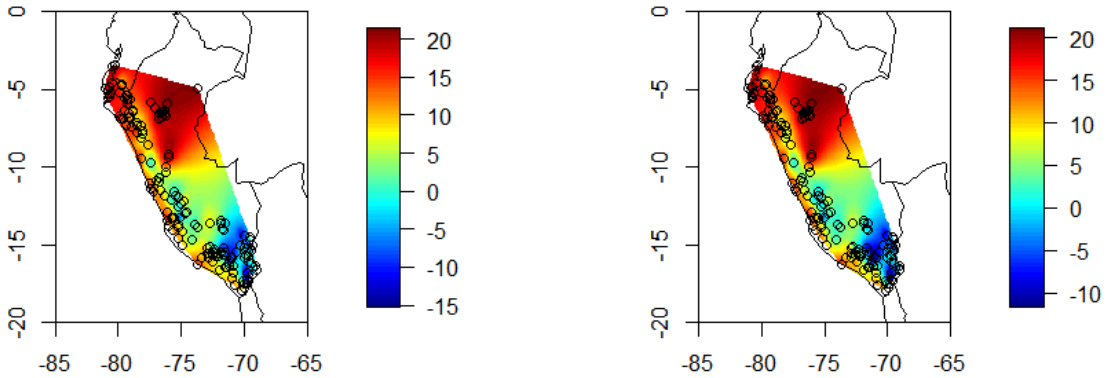

Figura 5.9: Izquierda: Mapa de temperaturas mínimas reales medidas en Agosto 2012 en el Perú. Derecha: Mapa de temperaturas mínimas estimadas con modelo GEV-S1.

Capítulo 6

Conclusiones

6.1. Comentarios finales

Se ha desarrollado un modelo generalizado de valores extremos mínimos con adición de un efecto espacial estimado bajo inferencia bayesiana, cuyo objetivo fue el de estimar las temperaturas mínimas diarias en el Perú, usando el método de aproximación de Laplace integrada y anidada (INLA). Según Blangiardo y Cameletti (2015) esta es una técnica con costos computacionales bajos que permite modelar procesos gaussianos jerárquicos. El modelo propuesto ha sido evaluado en comparación con uno que no incluya el efecto espacial, concluyendo que el primero obtiene estimaciones que se ajustan mejor a los datos reales. La principal ventaja de este modelo frente a un modelo geoestadístico clásico el cual únicamente estimaria las temperaturas mínimas en el Perú, es que el modelo geoestadístico GEV para mínimos, es capaz de estimar temperaturas mínimas extremas, es decir, las mínimas temperaturas mínimas en el Perú. Este tipo de estimaciones son de vital importancia para saber cuál es la temperatura mínima extrema que podría ocurrir en las regiones más afectadas por friajes.

6.2. Sugerencias para investigaciones futuras

- Se sugiere continuar con la investigación incluyéndose efectos espacio-temporales, dado que el tiempo es un factor importante para predecir temperaturas y por otro lado las nevadas y friajes en el Perú ocurren de manera estacionaria.
- Realizar la aplicación con mayor número de covariables que puedieran explicar también las temperaturas mínimas en la región.

Bibliografía

Banerjee, S., Gelfand, A. y Carlin, B. (2003). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC.

Blangiardo, M. y Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*, John Wiley & Sons, Ltd.

Blangiardo, M., Cameletti, M., Baio, G. y Rue, H. (2013). Spatial and spatio-temporal models with r-INLA, *Spatial and Spatio-temporal Epidemiology* **4**: 33–49.

CENEPRED (2018). Escenarios de riesgo por bajas temperaturas 2018, *Technical report*. https://cenepred.gob.pe/web/escenario-de-riesgo-trimestral-por-bajas-temperaturas/.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer London.

De, D. K., Ming-Huiand, C. y Chang, H. (1997). Bayesian approach for nonlinear random effects models, *Biometrics* **53**(4): 1239–1252.

Dyrrdal, A. V., Lenkoski, A., Thorarinsdottir, T. L. y Stordal, F. (2014). Bayesian hierarchical modeling of extreme hourly precipitation in norway, *Environmetrics* **26**(2): 89–106.

Fisher, R. A. y Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Mathematical Proceedings of the Cambridge Philosophical Society* **24**(2): 180–190.

Fuglstad, G.-A., Simpson, D., Lindgren, F. y Rue, H. (2017). Constructing priors that penalize the complexity of gaussian random fields, *Journal of the American Statistical Association* **114**(525): 445–452.

Geisser, S. y Eddy, W. F. (1979). A predictive approach to model selection, *Journal of the American Statistical Association* **74**: 153–160.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum dune serie aleatoire, *The Annals of Mathematics* **44**(3): 423.

Held, L., Schordle, B. y Rue, H. (2010). *Statistical Modelling and Regression Structures, chapter Posterior and crossvalidatory predictive checks: A comparison of MCMC and INLA*, Heidelberg: Physica-Verlag.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements, *Quarterly Journal of the Royal Meteorological Society* **81**(348): 158–171.

Kotz, S. y Nadarajah, S. (2000). *Extreme Value Distributions*, Published by Imperial College Press and distributed by World Scientific Publishing CO.

Lindgren, F. y Rue, H. (2015). Bayesian spatial modelling with r-inla, *Journal of Statistical Software* **63**(19).

Lindgren, F., Rue, H. y Lindstrom, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: The spde approach, *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **73**(4): 423–498.

Rachmawati, R. N., Djuraidah, A., Fitrianto, A. y Sumertajaya, I. M. (2018). Spatio-temporal models using r-inla with generalized extreme value distribution in hierarchical bayes regression, *International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET)* **4**(4): 1129–1143.

Rue, H. y Held, L. (2005). *Gaussian Markov Random Fields*, Chapman and Hall/CRC.

Rue, H., Martino, S. y Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(2): 319–392.

Simpson, D., Rue, H., Riebler, A., Martins, T. G. y Sorbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors, *Statistical Science* **32**(1): 1–28.

Spiegelhalter, D. J., Best, N., Carlin, B. y Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society, Series B* **64**: 583–640.

Tierney, L. y Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* **81**(393): 82–86.

Von Mises, R. (1936). La distribution de la plus grande de n valeur, *Rev Math Union Interbalcanique* **1**: 271–294.

Whittle, P. (1963). Stochastic processes in several dimensions, *Bull. Inst. Internat. Statist.* **40**: 974–994.

Anexo A: Generación de datos del GEV- para el mínimo

Se usa el método de la transformación inversa para generar una muestra aleatoria de $\tilde{M}_n = min\{Y_1, \cdots, Y_n\}$ con fda $F_{\tilde{M}_n}(y)$:

CASO 1: $(\xi \neq 0)$

$$F_{\tilde{M}_n}(y) = P(\tilde{M}_n \leq y) \quad = \quad 1 - \exp\left\{-\left(1 - \xi\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right)^{-1/\xi}\right\} \quad ; \xi \neq 0,\ 1 - \xi\left(\frac{y-\tilde{\mu}}{\sigma}\right) > 0$$

Sea $U \sim U(0,1)$, se asume que $u = F_Y(y)$, luego,

$$\begin{aligned}
F_{\tilde{M}_n}(y) &= 1 - \exp\left\{-\left(1 - \xi\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right)^{-1/\xi}\right\} \\
u &= 1 - \exp\left\{-\left(1 - \xi\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right)^{-1/\xi}\right\} \\
1 - u &= \exp\left\{-\left(1 - \xi\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right)^{-1/\xi}\right\} \\
\log(1-u) &= -\left(1 - \xi\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right)^{-1/\xi} \\
-\log(1-u) &= \left(1 - \xi\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right)^{-1/\xi} \\
(-\log(1-u))^{-\xi} &= 1 - \xi\left(\frac{y-\tilde{\mu}}{\sigma}\right) \\
\xi\left(\frac{y-\tilde{\mu}}{\sigma}\right) &= 1 - (-\log(1-u))^{-\xi} \\
(y-\tilde{\mu}) &= \frac{\sigma}{\xi}(1 - (-\log(1-u))^{-\xi}) \\
y &= \tilde{\mu} + \frac{\sigma}{\xi}(1 - (-\log(1-u))^{-\xi})
\end{aligned}$$

CASO 2: $(\xi = 0)$

$$F_{\tilde{M}_n}(y) = P(\tilde{M}_n \leq y) \quad = \quad 1 - \exp\left\{-\exp\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right\}$$

Sea $U \sim U(0,1)$, se asume que $u = F_Y(y)$, luego,

$$
\begin{aligned}
F_{\tilde{M}_n}(y) &= 1 - \exp\left\{-\exp\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right\} \\
u &= 1 - \exp\left\{-\exp\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right\} \\
1-u &= \exp\left\{-\exp\left(\frac{y-\tilde{\mu}}{\sigma}\right)\right\} \\
\log(1-u) &= -\exp\left(\frac{y-\tilde{\mu}}{\sigma}\right) \\
-\log(1-u) &= \exp\left(\frac{y-\tilde{\mu}}{\sigma}\right) \\
\log((-\log(1-u))) &= \left(\frac{y-\tilde{\mu}}{\sigma}\right) \\
y-\tilde{\mu} &= \sigma\log((-\log(1-u))) \\
y &= \tilde{\mu} + \sigma\log((-\log(1-u)))
\end{aligned}
$$

Anexo B: Código para la generación de datos del GEV- para el mínimo

```
############################
## simulation from gev - min
############################

qgevmin <-
function(p,shape=-1,scale=1,location=0)
{
if (shape == 0) {
  xF <- location + scale * log(-log(1-p))
} else {
  xF <- location+scale/shape*(1-(-log(1-p))^(-shape))
}

return(xF)
}

rgevmin <-
function(n,shape=-1,scale=1,location=0)
qgevmin(runif(n),shape,scale,location)

n=1000
b_0 <- -17 # intercept
b_1 <- 3 # coefficient for covariate
set.seed(8)
covariate <- rnorm(n)
lin.pred <- b_0 + b_1 * covariate

s <- 1
tau <- 2
s.y <- 1/sqrt(s*tau) # true scale
xi.gev <- -0.3 # true shape
library(evd)
set.seed(12)
y.gev <- rgevmin(n = length(lin.pred), loc = lin.pred,
  shape = xi.gev, scale = s.y)

summary(y.gev)
hist(y.gev)
```