Multicolinealidad y Sobreajuste en Modelos Jerárquicos Bayesianos para Datos Espaciales. Una Aplicación al Modelado de Servicios Ecosistémicos a Escala de Paisaje.

Tesista: Guillermina Senn

Directora: Monica Balzarini

Co-director: Raúl Macchiavelli

Proyecto de tesis para optar al grado de
Magíster en Estadística Aplicada.

Departamento de Estadística
Universidad Nacional de Córdoba
Argentina
Octubre 2021

Multicolinealidad y sobreajuste en modelos jerárquicos bayesianos para datos espaciales. Una aplicación al modelado de servicios ecosistémicos a escala de paisaje por Guillermina Senn se distribuye bajo una Licencia Creative Commons Atribución 4.0 Internacional.

Resumen

En esta tesis utilizamos modelos jerarquicos bayesianos con efectos aleatorios normales independientes e idénticamente distribuidos, $iCAR$ y BYM2 para modelar una respuesta continua en función de covariables georreferenciadas regionalmente. La naturaleza multicolineal de las covariables y la flexibilidad que los modelos jerárquicos bayesianos pueden introducir a través del efecto aleatorio hicieron necesario preguntarnos qué procedimientos seguir para realizar un correcto modelado de la respuesta. En primer lugar realizamos un análisis de sensibilidad para elegir distribuciones a priori de la precisión del efecto aleatorio y del parámetro de mezcla del BYM2 que no fuercen el sobreajuste. En todos los casos utilizamos distribuciones a priori penalizadas por complejidad controladas por U y α. Los resultados mostraron que aumentos de U y α en la distribución a priori de la precisión se tradujeron en mayor variabilidad y escalas más finas del efecto aleatorio. Particularmente en el modelo BYM2 observamos que la escala del efecto aleatorio resultó muy sensible a pequeños cambios en las distribuciones a priori de la precisión, y que la distribución a priori de la precisión pareció tener más peso en la determinación de la complejidad del modelo que la distribución del parámetro de mezcla. En segundo lugar aplicamos técnicas de reducción de datos para eliminar la multicolinealidad. La técnica de reducción de multicolinealidad que produjo los modelos con menor error de predicción consistió en detectar clústers de variables colineales y resumirlos con su primera componente principal obtenida con PCA. Adicionalmente, los resultados mostraron que las estimaciones de los coeficientes de regresión de los modelos jerárquicos bayesianos entrenados con covariables multicolineales tuvieron intervalos de credibilidad más anchos y que sus medias a posteriori sufrieron cambios de magnitud y sentido, respecto a los coeficientes estimados luego de reducir la multicolinealidad en los datos, concordando con los resultados observados en la literatura para modelos de regresión ordinarios.

Abstract

In this thesis we used Bayesian hierarchical models with normally independent and identically distributed random effects, and $iCAR$ and BYM2 random effects to model a continuous response as a function of regionally georeferenced covariates. The multicollinear nature of the covariates and the flexibility that Bayesian hierarchical models can introduce to the model through the random effect made us wonder what procedures to follow in order to correctly model the response. We first performed a sensitivity analysis to choose prior distributions for the precision of the random effect and for the BYM2 mixing parameter that did not force overfitting. We used penalized complexity priors parameterized by U and α in all cases. The results showed that larger U and α in the prior for the precision produced greater variability and smaller scales in the random effect. In the BYM2 model we observed that the prior for the precision seemed to have greater influence in determining the complexity of the model than the prior for the mixing parameter, and that the scale of the random effect was highly sensitive to small changes in the prior for the precision. Second, we applied data reduction techniques to reduce multicollinearity. The multicollinearity reduction technique that produced models with the lowest prediction error consisted in detecting clusters of collinear variables and summarizing them with their first principal component obtained with PCA. Additionally, the results showed that the estimates of the regression coefficients of the hierarchical Bayesian models trained with multicollinear covariates had wider credible intervals and that the posterior mean suffered changes of magnitude and direction, with respect to the estimated coefficients after reducing the multicollinearity in the data, in agreement with the results found in the literature for ordinary regression models.

Índice

1. Análisis de regresión para datos espaciales regionales **6**

1.1. Introducción 6

1.2. Caso de ilustración: servicios ecosistémicos a escala de paisaje 7

1.3. El problema de estudio 8

1.4. Hipótesis y objetivos 9

2. Marco teórico **11**

2.1. Datos espaciales 11

2.1.1. Tipos de datos espaciales 12

2.1.2. Autocorrelación espacial 13

2.2. Inferencia bayesiana 14

2.2.1. El Teorema de Bayes 14

2.2.2. Estimación bayesiana con INLA 15

2.3. Evolución del análisis de regresión 21

2.3.1. Modelo de regresión lineal 22

2.3.2. Modelo de regresión generalizado 24

2.3.3. Modelo de regresión generalizado mixto 26

2.3.4. Modelo de regresión jerárquico bayesiano 28

2.4. Modelos de regresión jerárquicos bayesianos para datos espaciales 29

2.4.1. Efecto aleatorio no estructurado, normal e idénticamente distribuido . . . 32

2.4.2. Efecto aleatorio estructurado condicional autorregresivo intrínseco (iCAR) 32

2.4.3. Efecto aleatorio estructurado BYM 34

2.5. Elección de distribuciones a priori 35

2.5.1. Distribuciones a priori por defecto en `R-INLA` 37

2.5.2. Distribuciones a priori penalizadas por complejidad 37

2.6. Validación y comparación de modelos bayesianos 41

2.6.1. Validación del modelo 43

2.6.2. Comparación de modelos 46

2.7. Técnicas de reducción de datos 54

2.7.1. Análisis de componentes principales (PCA) 55

2.7.2. MULTISPATI 56

2.7.3. Clustering jerárquico de variables 57

3. Metodología de investigación **59**

3.1. Datos 59

3.1.1. Datos originales 60

3.1.2. Datos preprocesados 61

3.2. Procedimientos 62

3.2.1. Análisis de sensibilidad a las distribuciones a priori 62

3.2.2. Análisis del impacto de la multicolinealidad 67

3.2.3. Selección del mejor modelo predictivo 71

4. Resultados y discusión **73**

4.1. Análisis de sensibilidad 73

4.1.1. Modelo *iid* 73

4.1.2. Modelo con efectos aleatorios iCAR 76

4.1.3. Modelo con efectos aleatorios BYM2 78

4.1.4. Selección final de una única distribución a priori por modelo 83

4.2. Análisis del impacto de la multicolinealidad 85

4.2.1. Comparación de PCA y MULTISPATI pre-modelado 86

4.2.2. Comparación de PCA y MULTISPATI post-modelado 89

4.3. Elección del mejor modelo predictivo 95

5. Conclusión y comentarios finales **106**

6. Referencias **111**

7. Anexos **115**

1. Análisis de regresión para datos espaciales regionales

1.1. Introducción

Dentro de las técnicas estadísticas más tradicionales y difundidas para generar modelos predictivos se encuentra el análisis de regresión. La versión clásica del análisis de regresión, el *modelo lineal*, es una técnica de principios del siglo XIX para predecir una respuesta gaussiana a partir de múltiples variables predictoras medidas en unidades de análisis independientes. El modelo lineal puede extenderse para modelar respuestas no gaussianas a través del *modelo lineal generalizado* (Nelder y Wedderburn, 1972).

Tanto el modelo lineal como el modelo lineal generalizado fueron pensados para predecir respuestas medidas sobre unidades de análisis u observaciones independientes y por lo tanto su funcionamiento con datos correlacionados no es óptimo. Por ejemplo, el modelo lineal para respuesta normal no está preparado para estimar correctamente los parámetros y genera intervalos de confianza demasiado angostos (Cressie, 1993; Beale et al., 2010).

En datos espaciales, la cercanía espacial induce correlación entre observaciones: se espera que regiones vecinas sean más similares entre sí que regiones distantes. Por lo tanto, el modelado de datos espaciales debería hacerse con modelos de regresión capaces de acomodar la falta de independencia, más sofisticados que el modelo lineal normal o generalizado. El *modelo de regresión generalizado mixto* (Breslow y Clayton, 1993) es una extensión del generalizado que incorpora *efectos aleatorios* para absorber la variabilidad debida a la correlación entre observaciones.

La estimación del modelo lineal generalizado mixto no es sencilla y se han propuesto varios métodos de estimación, entre los que se encuentran los métodos bayesianos. En el enfoque bayesiano completo, los parámetros del modelo se consideran variables aleatorias y se les asignan distribuciones a priori. Además, se considera que la distribución de estos parámetros está controlada por parámetros de segundo nivel, denominados *hiperparámetros*, a los que también se les asignan distribuciones a priori. Esta formulación da origen al nombre del *modelo de regresión jerárquico bayesiano*, donde el término *jerárquico* hace referencia a que el modelo se escribe en etapas: la primera etapa representa a la distribución de los datos, condicional a los parámetros e hiperparámetros; la segunda etapa representa la distribución de los parámetros, condicional a los

hiperparámetros; y la tercera etapa especifica la distribución marginal de los hiperparámetros. El modelo jerárquico bayesiano es una formulación muy flexible, de la cual el modelo generalizado mixto estimado con métodos bayesianos es un caso especial.

La estimación de los modelos bayesianos requiere realizar una integración altamente dimensional que no está generalmente disponible en forma cerrada. Tradicionalmente, esta integración se hizo mediante algoritmos de muestreo denominados *Cadenas de Markov - Monte Carlo* (MCMC), que son métodos computacionalmente muy demandantes y no tan sencillos de programar. El método de *aproximaciones de Laplace integradas anidadas* (INLA, del inglés *Integrated Nested Laplace Approximations*), propuesto por Rue, Martino y Chopin en 2009, es un método de estimación para *modelos gaussianos latentes* - un subconjunto de los modelos jerárquicos bayesianos - que resulta mucho más rápido y amigable que MCMC. Si bien INLA solo puede ser utilizado para estimar modelos gaussianos latentes, en la práctica esto casi no supone una limitación porque la familia de modelos gaussianos latentes es amplia y abarca muchos de los modelos más comunes, desde los modelos lineales generalizados a los modelos espaciales o espacio-temporales complejos (Rue et al., 2009).

Las buenas propiedades computacionales de INLA y su facilidad de uso en `R` gracias a `R-INLA` (Lindgren y Rue, 2015) han puesto al alcance del usuario común la estimación de modelos jerárquicos bayesianos. La creciente disponibilidad de datos espaciales y las posibilidades computacionales que INLA nos brinda motivan nuestro interés en el estudio y aplicación de modelos de regresión jerárquicos bayesianos para datos espaciales.

1.2. Caso de ilustración: servicios ecosistémicos a escala de paisaje

En el marco de una operación conjunta entre la Universidad Nacional de Córdoba (UNC), el Consejo Nacional de Actividades Científicas y Tecnológicas (CONICET) y la Comisión de Enlace de Entidades Agropecuarias (CEEA) Regional Córdoba se construyó un Sistema de Información Geográfica (SIG) (Tomlinson, 1968) entre los años 2020 y 2021 con gran cantidad de datos georreferenciados obtenidos por sensado remoto del territorio conocido como Cuenca del Carcarañá. El territorio de la cuenca se ubica geográficamente en las Provincias de Córdoba y de Santa Fe de la República Argentina y abarca aproximadamente 48000 km^2.

El SIG alberga variables edáficas, climáticas, meteorológicas, topográficas, de vegetación y de

cobertura de suelos, además de índices ecosistémicos relacionados con la degradación del terreno, el almacenamiento de carbono y la provisión de servicios ecosistémicos. El territorio en análisis está dividido en 4676 unidades regionales de aproximadamente 1420Has de superficie cada una. Todas las variables están disponibles en el SIG a escala de dichas unidades regionales.

Con los datos disponibles interesa generar modelos estadísticos a escala de paisaje con potencial para predecir riesgos emergentes en la Cuenca del Carcarañá en relación con la degradación de recursos naturales y la producción agropecuaria. En este trabajo, modelamos específicamente el Índice de Provisión de Servicios Ecosistémicos (ESPI, del inglés *Ecosystem Services Provisioning Index*) en función del resto de las variables del SIG.

1.3. El problema de estudio

En los modelos jerárquicos bayesianos para datos espaciales la complejidad del modelo depende, entre otras cosas, de la especificación de los efectos aleatorios y no solo del número p de covariables. Como el efecto aleatorio está controlado por sus hiperparámetros, serán las distribuciones a priori de los hiperparámetros las que controlan el equilibrio entre precisión predictiva y parsimonia en estos modelos (Simpson et al., 2017).

Con el deseo de evitar la introducción de subjetividad al modelo jerárquico bayesiano, la bibliografía en general recomienda que sino se cuenta con información a priori sobre los parámetros se utilicen distribuciones a priori no informativas o poco informativas, que tengan mínima influencia sobre la inferencia (Wang et al., 2018).

En los modelos jerárquicos bayesianos con efectos aleatorios, el uso de distribuciones no informativas en los hiperparámetros del efecto aleatorio no impone ninguna restricción a la flexibilidad del efecto aleatorio. Esto significa que, a priori, se le permite al efecto aleatorio tener suficiente variabilidad como para operar a una escala espacial inferior a la de las covariables y explicar los datos mejor que ellas (Illian et al., 2010; Illian et al., 2012).

El primer problema de estudio en este trabajo consiste encontrar un método para elegir distribuciones a priori para los hiperparámetros de los efectos aleatorios de los modelos jerárquicos bayesianos en estudio que resulten lo suficientemente informativas como para no forzar el sobreajuste, y lo suficientemente vagas como para no introducir subjetividad al modelo.

El segundo problema de estudio no está relacionado con la tendencia al sobreajuste de los modelos jerárquicos bayesianos, sino con el impacto que pueda tener en su estimación la presencia de multicolinealidad en la matriz de covariables.

La multicolinealidad es una característica de los datos que aparece cuando una covariable puede predecirse casi perfectamente a partir de otras covariables. Es un hecho conocido que en el modelo lineal estimado por máxima verosimilitud o mínimos cuadrados la multicolinealidad causa inestabilidad en la estimación de los coeficientes de regresión, e intervalos de confianza para los parámetros demasiado angostos (Belsley et al., 1980). Lo mismo sucede en los modelos autorregresivos de regresión espacial (Beale et al., 2010), y en los modelos de regresión bayesiana estimados con distribuciones no informativas (Leamer et al., 1973)

La bibliografía al respecto del impacto de la multicolinealidad en los modelos jerárquicos bayesianos para datos espaciales no es clara ni abundante. Asumiendo que la literatura podría beneficiarse con nuestra contribución, en este trabajo nos proponemos estudiar el comportamiento de los modelos jerárquicos bayesianos para datos espaciales en un contexto de multicolinealidad.

1.4. Hipótesis y objetivos

El objetivo general de este trabajo es identificar técnicas estadísticas que permitan aplicar correctamente modelos de regresión jerárquicos bayesianos a datos espaciales regionales en un contexto de multicolinealidad.

Se proponen como objetivos específicos:

1. Realizar un análisis de sensibilidad para elegir distribuciones a priori para los hiperparámetros de los efectos aleatorios que no causen sobreajuste.
2. Aplicar técnicas para la reducción de la multicolinealidad apropiadas para datos espaciales y evaluar el impacto de la multicolinealidad en el ajuste de los modelos jerárquicos bayesianos.
3. Modelar la respuesta ESPI de los datos de ilustración con distintos modelos jerárquicos bayesianos para datos espaciales regionales y encontrar el mejor modelo predictivo.

Las hipótesis que sustentan estos objetivos específicos son:

1. Las distribuciones a priori de los hiperparámetros de los efectos aleatorios son cruciales para controlar el sobreajuste en los modelos jerárquicos bayesianos.

2. La presencia de multicolinealidad en los datos afectará negativamente el ajuste de los modelos jerárquicos bayesianos.

3. La incorporación explícita de la estructura de dependencia espacial en los efectos aleatorios de sitio mejora la eficacia estadística de los modelos para datos espaciales sin efectos aleatorios o con efectos aleatorios no estructurados espacialmente.

Esta tesis está organizada como sigue. En la Sección se describe la teoría de los modelos jerárquicos bayesianos para datos espaciales, de INLA, de las distribuciones a priori penalizadas por complejidad y de los métodos de validación y comparación de modelos bayesianos. En la Sección 3 se describen los datos y los procedimientos seguidos para realizar el análisis de sensibilidad a las distribuciones a priori y para verificar el impacto de la multicolinealidad en el ajuste. Los resultados del análisis de sensibilidad, del impacto de la multicolinealidad y de la elección del mejor modelo predictivo se presentan en la Sección 4. El trabajo finaliza con las conclusiones y futuras líneas de investigación en la Sección 5.

2. Marco teórico

El presente capítulo está organizado como sigue. En la Sección 2.1 se definen los datos espaciales, los tipos de datos espaciales y la autocorrelación espacial. En la Sección 2.2 se describen las bases de la estadística bayesiana junto con el método de aproximaciones de Laplace integradas anidadas que se utiliza para la estimación de los modelos. La Sección 2.3 describe los modelos de regresión jerárquicos bayesianos en general y la Sección 2.4 los modelos jerárquicos bayesianos para datos espaciales. La Sección 2.5 describe las distribuciones a priori penalizadas por complejidad. La Sección 2.6 establece los criterios de validación y comparación de modelos bayesianos utilizados en este trabajo. Finalmente, la Sección 2.7 detalla las técnicas de reducción de datos, i.e., análisis de componentes principales espacial y no espacial y clustering de variables, que se pueden utilizar para tratar la multicolinealidad.

2.1. Datos espaciales

Un dato espacial es una observación que incluye información sobre su ubicación en un espacio euclídeo de d dimensiones $\mathbb{R}^d$. A diferencia de datos de una muestra aleatoria, las observaciones de una variable en n ubicaciones espaciales distintas no corresponden a múltiples realizaciones de la misma variable aleatoria ya que no son independientes. El conjunto de datos debe pensarse como una mustra de tamaño $n = 1$ de un proceso estocástico espacial correspondiente a una variable aleatoria definida en $\mathbb{R}^2$.

Sea $\boldsymbol{s} \in \mathbb{R}^d$ una ubicación cualquiera en un espacio euclídeo en d dimensiones y sea $Z(\boldsymbol{s})$ un dato que puede medirse en esa ubicación. El conjunto de Zs que se corresponden con las ubicaciones $\boldsymbol{s}$ pertenecientes al conjunto de índices $D \subset \mathbb{R}^d$ forma un **campo o proceso aleatorio** (Cressie, 1993)

$$\{\boldsymbol{Z}(\boldsymbol{s}) : \boldsymbol{s} \in D\}. \tag{1}$$

Como D es un subconjunto de $\mathbb{R}^d$, este proceso aleatorio o estocástico se denomina *proceso espacial*, y los datos espaciales surgen como realizaciones del proceso espacial.

2.1.1. Tipos de datos espaciales

Según la naturaleza del dominio D se definen tres tipos de datos espaciales: datos de *patrones de puntos*, datos *geoestadísticos* y datos *regionales o areales*.

En los datos de patrones de puntos espaciales el conjunto de ubicaciones de los eventos D es aleatorio. Es decir, puede cambiar de realización a realización. Por ejemplo, las ubicaciones correspondientes a las caídas de rayos durante el mes de Agosto en determinado territorio corresponden a un patrón de puntos espacial (Schabenberger y Gotway, 2005).

En los datos geoestadísticos D es un subconjunto fijo de $\mathbb{R}^d$ y el índice espacial $\boldsymbol{s}$ varía de manera continua D (Cressie, 1993). Por ejemplo, el conjunto de observaciones de la variable Precipitaciones Líquidas acumuladas en el mes de agosto, medida a través de pluviómetros distribuidos regularmente sobre el territorio de interés, representa un conjunto de datos geoestadísticos.

En los datos regionales o areales D también es un subconjunto fijo de $\mathbb{R}^d$, pero es contable, por lo que el índice espacial $\boldsymbol{s}$ no puede variar de manera continua y sólo puede representar a los sitios espaciales incluidos en el subconjunto D (Cressie, 1993). Por ejemplo, la tasa anual de homicidios observada a nivel departamental para la provincia de Córdoba generaría un conjunto de datos regionales, donde las regiones serían los departamentos de la provincia. Los departamentos constituyen un conjunto contable.

Los datos regionales surgen en su mayoría como datos agregados espacialmente sobre regiones. Las regiones pueden ser irregulares, como en el caso de los departamentos de la Provincia de Córdoba, o una grilla regular, como en el caso de parcelas en un experimento. En general, es conveniente asignar a cada región una coordenada espacial representativa, como puede ser el centroide de la región, para poder realizar mediciones entre regiones (Schabenberger y Gotway, 2005)

Es posible definir un *vecindario* $N_i = \{k : k \text{ es vecino de } i\}, \quad i = 1, ..., n$ para cada región i. Es muy usual definir a N_i es como el conjunto de regiones que comparten un límite con la región i (Cressie, 1993). Hay múltiples criterios para decidir si dos regiones comparten un límite; los más comunes son el criterio *reina*, donde dos regiones son vecinas si se tocan al menos en un punto, y el criterio *torre*, donde es necesario que se toquen en al menos dos puntos que estén separados como máximo cierta distancia (Bivand et al., 2008).

Una vez definido el conjunto de vecinos para cada región, se le asigna a la relación entre la región i y una región j un *peso espacial*. Para la totalidad de regiones, se forma una *matriz* $\boldsymbol{W}$ *de pesos espaciales* donde el elemento W_{ij} corresponde al peso que tiene el enlace entre las regiones i y j. Si el elemento ij es cero, significa que las regiones i y j no son vecinas (Bivand et al., 2008).

Los pesos espaciales diferentes a cero se pueden elegir con distinto criterio. Por ejemplo, podrían ser pesos binarios, donde $W_{ij} = 1$ si i y j son vecinos. Otra opción es definir pesos normalizados por la cantidad de vecinos de la región, donde $W_{ij} = 1/|N_i|$, siendo $|N_i|$ la cardinalidad del conjunto N_i, cuando i y j son vecinas (Bivand et al., 2008).

2.1.2. Autocorrelación espacial

La *autocorrelación espacial* es la correlación entre $Z(\boldsymbol{s}_i)$ y $Z(\boldsymbol{s}_j)$; es decir, es la correlación del atributo Z con sí mismo, pero medido en ubicaciones distintas (Schabenberger y Gotway, 2005). Si un atributo no muestra autocorrelación espacial, significa que la ubicación no es relevante para determinar el valor del atributo. Es de esperar que los datos georreferenciados sí presenten autocorrelación espacial, inducida por la proximidad entre observaciones.

El *índice de Moran* (Moran, 1950) es un estadístico que mide la autocorrelación espacial para datos regionales. El I de Moran se define como

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(Z(\boldsymbol{s}_i) - \bar{Z})(Z(\boldsymbol{s}_j) - \bar{Z})}{(\sum_{i \neq j} W_{ij}) \sum_{i=1}^{n} (Z(\boldsymbol{s}_i) - \bar{Z})^2}, \tag{2}$$

donde W_{ij} es el peso espacial correspondiente al enlace entre las ubicaciones i y j.

La hipótesis nula de I es la ausencia de autocorrelación espacial. Bajo el modelo nulo, donde las $Z(\boldsymbol{s}_i)$ son independientes e idénticamente distribuidas, I es asintóticamente normal con media $E[I] = -1/(n-1)$ (Banerjee et al., 2015). Cuando $I > E[I]$, es señal de autocorrelación espacial positiva; cuando $I < E[I]$, es señal de autocorrelación espacial negativa.

La inferencia para el I de Moran se puede hacer mediante pruebas de permutación o pruebas de Monte-Carlo, entre otras (Schabenberger y Gotway, 2005). Banerjee et al. (2015) recomiendan utilizar el enfoque Monte Carlo, que luego de una cantidad moderada de permutaciones (incluyendo al I observado) posicionará al I observado en relación al resto de las permutaciones y

permitirá determinar si es extremo o no, posiblemente a través de un p-valor.

2.2. Inferencia bayesiana

La inferencia estadística se ocupa de extraer conclusiones sobre cantidades no observadas a partir del estudio de datos medidos sobre una muestra de la población en estudio (Gelman et al., 2014). En la inferencia bayesiana, los parámetros se consideran variables aleatorias, y las conclusiones estadísticas sobre los parámetros se hacen condicionales a los datos observados y en términos de afirmaciones probabilísticas (Gelman et al., 2014). El artefacto matemático que permite calcular estas afirmaciones probabilísticas condicionales a los datos observados es el *Teorema de Bayes* (Bayes, 1763; Price, 1763; Laplace, 1774).

2.2.1. El Teorema de Bayes

Sea la función de verosimilitud de los datos L indexada por un vector de parámetros $\boldsymbol{\theta}$ con componentes θ_j:

$$L(\boldsymbol{\theta}) = p(Y = y|\boldsymbol{\theta}). \tag{3}$$

En el enfoque bayesiano, los parámetros θ_j son variables aleatorias cuya distribución a priori $\pi(\boldsymbol{\theta})$ se especifica antes de ver los datos (Blangiardo and Cameletti, 2015).

El **Teorema de Bayes** (Ecuación 4) combina la verosimilitud $L(\boldsymbol{\theta})$ con la distribución a priori $\pi(\boldsymbol{\theta})$ y genera una distribución a posteriori $\pi(\boldsymbol{\theta}|y)$ del vector de parámetros que es un compromiso entre su distribución a priori y la verosimilitud.

$$p(\boldsymbol{\theta}|y) = \frac{p(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\sum_{i=1}^{K} p(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}. \tag{4}$$

El denominador del Teorema de Bayes es la verosimilitud marginal, que se calcula sumando o integrando la verosimilitud respecto a $\boldsymbol{\theta}$.

$$p(y) = \int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} p(y|\theta)\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{5}$$

En la práctica, la distribución a posteriori se estima sin computar esta verosimilitud marginal, y el Teorema de Bayes se reescribe como

$$p(\boldsymbol{\theta}|y) \propto p(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{6}$$

Uno de los artefactos más usados en la inferencia bayesiana para realizar afirmaciones probabilísticas sobre los parámetros es el *intervalo de credibilidad*. Para un nivel de credibilidad del 95 %, el intervalo de credibilidad para θ_j es el rango de valores comprendidos entre los $(\theta_{0.025}, \theta_{0.975})$, donde

$$p(\theta_j \leq \theta_{0.025}|y) = 0.025 \quad y \quad p(\theta_j \geq \theta_{0.975}|y) = 0.025. \tag{7}$$

La interpretación del intervalo de credibilidad es directa: *para un intervalo de credibilidad del* $(1-\alpha)\,\%$*, el verdadero* θ_j *cae entre* $(\theta_{\alpha/2}$ *y* $\theta_{1-\alpha/2})$ *con una probabilidad del* $(1-\alpha)\,\%$, dados los datos.

2.2.2. Estimación bayesiana con INLA

En la estimación bayesiana, las distribuciones a priori pueden elegirse de forma que la distribución a posteriori de los parámetros esté disponible en forma cerrada. Esto ocurre, por ejemplo, cuando se usan distribuciones a priori *conjugadas*. En la gran mayoría de los modelos no es este el caso, no es es posible estudiar la distribución a posteriori analíticamente, y para conocerla hay que explorarla mediante métodos de simulación (Blangiardo y Cameletti, 2015).

Los métodos de Cadena de Markov - Monte Carlo (MCMC) son métodos computacionales probabilísticos para generar muestras de la distribución **conjunta** a posteriori (Gómez-Rubio, 2020). Estos métodos construyen una cadena de Markov cuya distribución estacionaria es igual a la distribución a posteriori que se intenta muestrear. Si la cadena de Markov cumple ciertas propiedades, las muestras obtenidas después de cierto número de iteraciones de la cadena se pueden considerar como provenientes de la distribución a posteriori. Los algoritmos más conocidos para

construir estas cadenas son el de *Metropolis-Hastings* (Metropolis et al., 1953) y el *muestreo de Gibbs* (Geman y Geman, 1984).

Los métodos MCMC producen resultados exactos pero son lentos porque deben computar la distribución a posteriori conjunta de los parámetros, que está en un espacio de altas dimensiones (Gómez-Rubio, 2020). El cálculo de la inferencia bayesiana se vio propulsado por la aparición de INLA, un método de inferencia bayesiana aproximada propuesto en 2009 por Rue, Martino y Chopin, con mejores características computacionales que MCMC.

INLA, acrónimo de *Integrated Nested Laplace Approximations* o Aproximaciones de Laplace Integradas-Anidadas, es un método computacional para realizar inferencia bayesiana para modelos modelos gaussianos latentes. Como los modelos gaussianos latentes se construyen con un campo aleatorio Gaussiano de Markov (Ver la Ecuación 13), su matriz de precisión es dispersa. INLA aprovecha esta cualidad y combina métodos numéricos para matrices dispersas, que son muy eficientes, con aproximaciones de Laplace, para estimar las distribuciones **marginales** a posteriori de los parámetros. El método resulta mucho más rápido que MCMC y, a veces, más preciso (Rue et al., 2009).

Para estimar un modelo con INLA, hay que plantearlo como un modelo gaussiano latente, que a su vez tiene una formulación basada en la formulación del modelo lineal generalizado.

Sea entonces un vector de observaciones $\mathbf{y} = (y_1, ..., y_n)$ cuya distribución pertenece a la familia exponencial y se puede caracterizar con su media $E[y_i] = \mu_i$. Además, sea un predictor lineal η_i aditivo con la forma de la ecuación (8):

$$\eta_i = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} f^{(k)}(u_{ki}) + \varepsilon_i; \quad i = 1, ..., n. \tag{8}$$

La media y el predictor lineal se relacionan a través de la función de enlace g:

$$\eta_i = g(\mu_i). \tag{9}$$

En el predictor lineal se incluyen el error aleatorio ε_i, los efectos fijos β_j y los efectos especiales $f(\cdot)$ normales. En estos efectos especiales se pueden modelar relaciones no lineales, tendencias temporales, efectos aleatorios y efectos espaciales, entre otros (Rue et al., 2009).

Para que el modelo anterior sea considerado un modelo gaussiano latente, los componentes del predictor lineal α, $\{\beta_k\}$, $\{f^{(j)}\}$ y $\{\varepsilon_i\}$ deben tener distribuciones a priori gaussianas. La combinación de α, $\{\beta_k\}$, $\{f^{(j)}\}$ y $\{\eta_i\}$ genera un campo latente $\boldsymbol{x} = (\alpha, \boldsymbol{\beta}, \boldsymbol{f}, \boldsymbol{\eta})$ que, condicional a los hiperparámetros $\boldsymbol{\theta_1}$, se distribuye gaussiano:

$$\boldsymbol{x}|\boldsymbol{\theta_1} \sim N(\boldsymbol{0}, \boldsymbol{Q}^{-1}(\boldsymbol{\theta_1})) \tag{10}$$

Las matriz de precisión estará parametrizada por los hiperparámetros $\boldsymbol{\theta_1}$, cuya distribución a priori $\pi(\boldsymbol{\theta_1})$ no necesariamente debe ser gaussiana.

Finalmente se define la distribución del vector de observaciones $\boldsymbol{y}$ condicional respecto al campo latente y a un segundo vector de hiperparámetros $\boldsymbol{\theta_2}$:

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta_2} \sim \prod_{i=1}^{n} \pi(y_i|x_i, \boldsymbol{\theta_2}), \tag{11}$$

donde se asume que cada observación y_i solo se corresponde con un predictor lineal, η_i, y que las observaciones son condicionalmente independientes:

De esta manera, el modelo gaussiano latente ha quedado definido en tres etapas: la distribución de los hiperparámetros, la distribución de los parámetros condicional a los hiperparámetros, y la distribución de los datos condicional a los parámetros e hiperparámetros. Es posible calcular la distribución a posteriori conjunta del campo latente y de los hiperparámetros con la Ecuación 12, que está expresada en función de un único vector de hiperparámetros $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ para simplificar la notación, siguiendo la formulación de Rue et al. (2009).

$$\begin{aligned}
\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i=1}^{n} \pi(y_i|x_i, \boldsymbol{\theta}) \\
&\propto \pi(\boldsymbol{\theta})|\boldsymbol{Q}(\boldsymbol{\theta})|^{1/2} exp\{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x}\} \prod_{i=1}^{n} \pi(y_i|x_i, \boldsymbol{\theta}) \\
&= \pi(\boldsymbol{\theta})|\boldsymbol{Q}(\boldsymbol{\theta})|^{1/2} exp\{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_{i=1}^{n} log\, \pi(y_i|x_i, \boldsymbol{\theta})\}
\end{aligned} \tag{12}$$

Los modelos latentes gaussianos que INLA resuelve deben satisfacer dos propiedades básicas (Rue et al., 2009). La primera propiedad asume que el campo latente $\boldsymbol{x}$ es un *campo aleatorio gaussiano markoviano* (GMRF, del inglés *Gaussian Markov Random Field*) con matriz de precisión dispersa $\boldsymbol{Q}(\boldsymbol{\theta})$.

Un GMRF puede definirse en función de un grafo $G = (V, E)$ donde $V = \{1, ..., n\}$ es el conjunto de nodos o vértices, y E es el conjunto de aristas tal que no hay arista entre dos nodos que son condicionalmente independientes, dados el resto de los nodos. Tomando prestada la definición de Rue y Held (2005), un vector aleatorio $\boldsymbol{x}$ es un GMRF respecto a un grafo G, con media $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{Q} > 0$, si y solo si $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$ y

$$Q_{ij} \neq 0 \iff \{i, j\} \in E \quad \textit{para cualquier} \quad i \neq j. \tag{13}$$

En datos espaciales, esta definición implica que cuantas menos conexiones haya entre nodos, o equivalentemente, cuantos menos vecinos tenga una observación, más elementos de la $\boldsymbol{Q}$ se harán cero, y mayor eficiencia computacional se obtendrá al poder realizar operaciones para matrices dispersas.

La segunda propiedad que INLA requiere es que el número de hiperparámetros $m = dim(\boldsymbol{Q})$ sea pequeño, alrededor de $m \leq 6$, para que la inferencia sea rápida computacionalmente. Por el contrario, $dim(\boldsymbol{x})$ puede estar en el orden de los cientos o cientos de miles.

El objetivo de INLA es resolver dos tareas (Blangiardo y Cameletti, 2015):

1. Encontrar la distribución marginal a posteriori de cada elemento del vector de hiperparámetros

$$\pi(\theta_j|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-j}. \tag{14}$$

   Para ello se calcula una aproximación a la distribución conjunta a posteriori de los hiperparámetros $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ que luego se integra para obtener las marginales.

2. Encontrar la distribución marginal a posteriori de cada elemento del campo latente:

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i, \boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} = \int \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \pi(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}. \tag{15}$$

Para resolver *2.* se necesitan la aproximación calculada en *1.* y la aproximación de la distribución condicional a posteriori de cada elemento del campo latente, $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$.

La Ecuación 16 muestra cómo encuentra INLA una aproximación a la distribución conjunta a posteriori del vector de hiperparámetros.

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\boldsymbol{y}) &= \pi(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y}) \cdot \frac{1}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \\
&= \frac{\pi(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})}{\pi(\boldsymbol{y})} \cdot \frac{1}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \\
&= \frac{\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{y})} \cdot \frac{1}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \\
&\propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}.
\end{aligned} \tag{16}$$

Como el denominador es la distribución condicional del campo latente, que a priori es un GMRF, es correcto aproximarlo a posteriori con una aproximación gaussiana dada por la aproximación de Laplace $\tilde{\pi}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ (Blangiardo y Cameletti, 2015).

Reemplazando el denominador por su versión aproximada y evaluando la expresión completa en la moda del campo latente para un determinado valor del vector de hiperparámetros, se obtiene una aproximación a la distribución conjunta del vector de hiperparámetros, $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$.

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \approx \left.\frac{\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\tilde{\pi}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}\right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} =: \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}). \tag{17}$$

Conociendo $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$, las posteriori de cada hiperparámetro en *1.* se encuentran por integración.

Para encontrar *2.*, la distribución marginal a posteriori de cada elemento del campo latente, hace falta la distribución condicional a posteriori del elemento del campo latente que se desea, $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$, además de la conjunta a posteriori de los hiperparámetros.

Encontrar la distribución para cada x_i a partir de la conjunta de $\boldsymbol{x}$ es complejo porque la dimensión de $\boldsymbol{x}$ podría rondar los cientos o miles y la integración de la conjunta para obtener una marginal implicaría una integración de altas dimensiones. Rue *et al.* (2009) plantean tres formas para encontrar las $p(x_i|\boldsymbol{\theta}, \boldsymbol{y})$: una *aproximación Gaussiana*, una *aproximación de Laplace*,

y una *aproximación simplificada de Laplace*. La aproximación Gaussiana no se recomienda porque no produce buenos resultados. La aproximación de Laplace se describe en la Ecuación 18.

Reescribiendo a $\boldsymbol{x}$ como $\boldsymbol{x} = (x_i, \boldsymbol{x}_{-i})$:

$$\begin{aligned}
\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y}) &= \frac{\pi((x_i, \boldsymbol{x}_{-i})|\boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \\
&= \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{\theta}|\boldsymbol{y})} \cdot \frac{1}{\pi(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \\
&\propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})}.
\end{aligned} \tag{18}$$

El denominador se puede aproximar correctamente con una aproximación de Laplace porque las $\boldsymbol{x}_{-i}|(x_i, \boldsymbol{\theta}, \boldsymbol{y})$ son en general razonablemente normales. La expresión completa se evalúa en la moda $\boldsymbol{x}^*_{-i}(x_i, \boldsymbol{\theta})$ y se obtiene finalmente la expresión aproximada de la Ecuación 19. Esta aproximación de Laplace es costosa computacionalmente porque requiere recalcular $\tilde{\pi}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})$ para cada combinación de valores de $\boldsymbol{x}$ y de $\boldsymbol{\theta}$.

$$\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \propto \left.\frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{\tilde{\pi}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})}\right|_{\boldsymbol{x}_{-i}=\boldsymbol{x}^*_{-i}(x_i, \boldsymbol{\theta})} =: \tilde{\pi}(x_i|\boldsymbol{\theta}, \boldsymbol{y}). \tag{19}$$

La aproximación simplificada de Laplace esta basada en la serie de Taylor de $\tilde{\pi}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})$. Como esta opción es más rápida que la aproximación normal y suele tener buenos resultados, es la opción por defecto en la librería `R-INLA` (Blangiardo y Cameletti, 2015).

La aproximación a la distribución marginal a posteriori para los elementos del campo latente se calcula resolviendo la integral de la Ecuación 15 a través de una suma ponderada, donde se han reemplazado las distribuciones a posteriori del integrando por sus versiones aproximadas:

$$\tilde{\pi}(x_i|\boldsymbol{y}) \approx \sum_j \tilde{\pi}(x_i|\boldsymbol{\theta}^{(j)}, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}^{(j)}|\boldsymbol{y})\boldsymbol{\Delta}_j. \tag{20}$$

Como la suma se hace sobre el espacio de $\boldsymbol{\theta}$, es necesario elegir puntos de integración $\{\boldsymbol{\theta}_k^{(j)}\}$. Dichos puntos se encuentran explorando el logaritmo de la distribución aproximada a posteriori de los hiperparámetros $log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$. Antes de comenzar a explorar la distribución se reparametriza

el espacio de los hiperparámetros para corregir la escala y la rotación y simplificar la integración numérica (Rue *et al.*, 2009).

Hay dos estrategias para explorar $log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$. La estrategia de grilla construye una grilla de puntos asociada con el grueso de la masa de $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$. Es posible usarla cuando se tienen un máximo de cuatro hiperparámetros, pero cuando el número de hiperparámetros es mayor a cuatro, se recomienda usar la estrategia de *diseño compuesto central*. El diseño compuesto central solo elige algunos puntos relevantes del espacio de $\boldsymbol{\theta}$, que en general suelen ser muchos menos puntos que los necesarios con la estrategia de grilla (Blangiardo y Cameletti, 2015). El diseño compuesto central acelera los cómputos manteniendo resultados casi equivalentes a los obtenidos con la estrategia de grilla y por lo tanto es la opción por defecto en `R-INLA` para explorar $log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$ (Rue et al., 2009). La fórmula para el cálculo de los pesos $\{\Delta_j\}$ bajo la estrategia CCD puede encontrarse en la Sección 6.5 de Rue et al. (2005).

Una vez que se tienen los puntos $\{\boldsymbol{\theta}_k^{(j)}\}$, se resuelve la suma ponderada de la Ecuación 20 para encontrar la aproximación a las marginales a posteriori del campo latente, y se utiliza un algoritmo de interpolación para encontrar la aproximación a las marginales a posteriori para los elementos del vector de hiperparámetros de la Ecuación 14. Esta interpolación se basa en los valores de $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ evaluada en los puntos $\{\boldsymbol{\theta}_k^{(j)}\}$.

2.3. Evolución del análisis de regresión

El análisis de regresión es una de las técnicas más antiguas y estudiadas para generar modelos que relacionan la respuesta con las covariables. El primer paso en los modelos de regresión fue la aparición del método de *mínimos cuadrados* (Legendre, 1806) en un tratado de geodesia. Casi 100 años después, en 1895, Sir Francis Galton describió el fenómeno de la *regresión a la media*, que se manifiesta cuando en un conjunto de datos los procesos naturales amortiguan los "outliers" extremos y los acercan a la media. Galton utilizó una *regresión lineal simple* estimada con mínimos cuadrados para resolver un problema relacionado con su estudio de la herencia genética. Finalmente, a principios del siglo XX, Fisher completó el desarrollo formal de la *regresión lineal múltiple*.

Con el correr de los años, el marco de la regresión lineal se desarrolló y amplió de manera tal que la regresión lineal ordinaria propuesta por Galton y Pearson hoy se considera un caso especial

del *modelo lineal generalizado* propuesto por Nelder y Wedderburn en 1972. El modelo lineal generalizado extiende al modelo lineal permitiendo distribuciones de la respuesta no gaussianas, y estima los parámetros con técnicas basadas en la máxima verosimilitud (Nelder y Wedderburn, 1972).

A su vez, el *modelo lineal generalizado mixto* extiende al modelo lineal generalizado para que pueda acomodar fuentes de variabilidad que no pueden ser explicadas por las covariables, tales como aquellas que surgen de la correlación entre observaciones. El modelo lineal generalizado mixto propuesto por Breslow y Clayton (1993) tiene una formulación jerárquica, donde la distribución de los datos se plantea condicional a los efectos aleatorios, y los efectos aleatorios a su vez tienen distribución normal con matriz de covarianza parametrizada por un vector de componentes de varianza desconocidos. La estimación frecuentista de los modelos lineales generalizados mixtos puede hacerse mediante métodos basados en la verosimilitud, como los iniciales cuasi-verosimilitud penalizada o marginal propuestos por Breslow y Clayton (1993) y las más modernas aproximaciones de Laplace o de cuadratura gaussiana adaptativa a la verosimilitud marginal propuestas por Pinheiro y Bates (1995; 2000).

Por otro lado, si a los parámetros del modelo se les asignan distribuciones a priori y además se asume que parte de o todos estos parámetros están controladas por hiperparámetros, y se les agrega una distribución a priori a los hiperparámetros, los modelos resultantes pueden ser estimados mediante métodos bayesianos y caen en la familia de *modelos jerárquicos bayesianos*.

Los modelos jerárquicos bayesianos proporcionan un marco general para la formulación en etapas de un modelo que se estima con métodos bayesianos. Los modelos jerárquicos bayesianos son flexibles, pueden modelar relaciones complejas, y sobre todo, pueden ser fácilmente estimados con INLA si se pueden escribir como un modelo latente gaussiano.

A continuación se describe brevemente la teoría de cada modelo mencionado hasta llegar a los modelos jerárquicos bayesianos.

2.3.1. Modelo de regresión lineal

El modelo de regresión lineal toma su nombre del concepto de *regresión*, que en estadística se usa para denotar una relación entre variables, y del termino *lineal*, que en este caso indica que

el modelo es *lineal en los parámetros* (Montgomery et al., 2012).

El modelo de regresión lineal para una observación i asume una relación entre respuesta y covariables con la forma

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_J x_{iJ} + \varepsilon_i; \quad i = 1, ..., n \tag{21}$$

donde y_i es una variable aleatoria respuesta medida en la observación i y las x_i son variables observables. En el marco de la regresión clásica frecuentista, se asume que cada covariable tiene asociado un *coeficiente de regresión* β_j, que es una cantidad fija y desconocida, y que ε_i es una variable aleatoria que representa la variabilidad de las y_i no explicada por las covariables (Casella y Berger, 2002).

La función de regresión de la población se obtiene tomando la esperanza condicional de la expresión anterior:

$$E[y_i|x_i] = \beta_0 x_{i0} + \beta_1 x_{i1} + ... + \beta_p x_{ip} \tag{22}$$

Para hacer inferencia a la población se asume que las $\varepsilon_1, ..., \varepsilon_n$ son variables aleatorias no correlacionadas tales que $E[\varepsilon_i] = 0$ y $Var(\varepsilon_i) = \sigma^2$. Bajo esta formulación, la varianza del error σ^2 es constante; es decir, es la misma para todas las observaciones i. Este supuesto se denomina *homocedasticidad*.

Para poder derivar pruebas de hipótesis e intervalos de confianza hay que especificar una distribución de probabilidad para los datos. Cuando se asume que las $y_i s$ tienen distribución normal se obtiene el *modelo condicional normal*, que es el modelo de regresión lineal mas simple. En este modelo, el supuesto de no correlación entre los errores se transforma en el supuesto de independencia y es posible especificar de manera exacta la distribución de probabilidad marginal y conjunta de las $y_i s$ (Casella y Berger, 2012):

$$y_i|\boldsymbol{\beta}, \sigma^2 \sim N(\sum_{j=1}^{p} \beta_k x_{ij}, \sigma^2), \quad i = 1, ..., n. \tag{23}$$

Los estimadores para los β_j y σ^2 se pueden obtener con los métodos de estimación por *máxima*

verosimilitud o de mínimos cuadrados.

En resumen, para que la especificación de una relación pueda hacerse correctamente con el modelo de regresión condicional normal, los datos no deben contradecir tres supuestos: (1) relación lineal entre la media y las covariables; (2) homocedasticidad en los errores; y (3) observaciones no correlacionadas. Además, para realizar inferencia aprovechando las buenas propiedades de los estimadores, se agrega el supuesto (4) de distribución normal de los errores.

2.3.2. Modelo de regresión generalizado

Si bien el término "modelo lineal generalizado" se popularizó con Nelder y Wedderburn en 1972, muchos ejemplos de modelos lineales generalizados - como análisis probit y tablas de contingencia - fueron desarrollados independientemente a lo largo del siglo XX. En su publicación de 1972, Nelder y Wedderburn notaron que todos estos modelos podían ser divididos en un componente aleatorio y en un componente sistemático con estructura lineal en los parámetros, y propusieron al modelo lineal generalizado (MLG) como marco unificador junto con un método para su estimación basado en la verosimilitud.

Gracias a su formulación, los MLG pueden verse como una extensión natural del modelo regresión lineal estándar que permite modelar distribuciones no normales de la respuesta y funciones no-lineales de la media (Agresti, 2015). El MLG consta de tres partes: el *componente aleatorio*, el *predictor lineal* o *componente sistemático* y la *función de enlace.*

El componente aleatorio (Ecuación 24) especifica la respuesta $\boldsymbol{y} = (y_1, ..., y_n)$ donde las y_i tienen la misma distribución de probabilidad en la familia exponencial y se asumen condicionalmente independientes dados los parámetros.

$$f(y_i|\theta_i, \phi) = exp\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\}, \tag{24}$$

donde ϕ se denomina *parámetro de dispersión.* Cuando ϕ es conocido, este modelo pertenece a la familia exponencial con *parámetro canónico* o *natural* θ_i. Si ϕ es desconocido, el modelo puede pertenecer o no a la familia exponencial (McCullagh y Nelder, 1989).

Las Ecuaciones 25 y 26 muestran relación entre el parámetro canónico y los momentos de la

distribución. Los detalles de la obtención de estas relaciones pueden encontrarse por ejemplo en el Capítulo 2 de McCullagh y Nelder (1989).

$$E[y_i] = \mu_i = b'(\theta_i) \tag{25}$$

y

$$Var(y_i) = b''(\theta_i)a(\phi). \tag{26}$$

En el componente sistemático (Ecuación 27), las variables explicativas $x_1, ..., x_p$ producen un predictor lineal η_i (McCullagh y Nelder, 1989) a través de una combinación lineal de los parámetros, reflejando la cualidad *lineal en los parámetros* del MLG.

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, ..., n \tag{27}$$

Finalmente, la función de enlace g, monótona y diferenciable, relaciona el predictor lineal η_i con la media $\mu_i = E[y_i]$:

$$g(\mu_i) = \eta_i \tag{28}$$

El modelo lineal normal termina siendo un MLG con enlace identidad y respuesta normal (Ecuación 29).

$$\begin{aligned} f(y_i|\mu_i, \sigma^2) &= N(\mu_i, \sigma^2) \\ \eta_i &= \sum_{j=1}^{p} \beta_j x_{ij} \\ g(\mu_i) &= \mu_i \end{aligned} \tag{29}$$

Los MLG se pueden ajustar con técnicas frecuentistas basadas en la verosimilitud o con métodos bayesianos.

2.3.3. Modelo de regresión generalizado mixto

En las ciencias aplicadas es común ver datos correlacionados debido a múltiples causas, donde algunas de las más comunes son las medidas repetidas, las observaciones agrupadas en clústers, los datos longitudinales y la dependencia espacial (Verbeke y Molenberghs, 2000).

Los modelos para datos longitudinales comenzaron a ser estudiados antes que los modelos para datos espaciales. En datos longitudinales, existe tanto el concepto de *réplicas* como el de *distancia*. El concepto de réplicas hace referencia a que se toman múltiples mediciones de la misma unidad, mientras que la distancia se refiere a la distancia temporal entre estas mediciones.

En los datos espaciales no hay réplicas, pero el concepto de distancia (espacial en este caso) está presente, por lo que los modelos desarrollados para datos longitudinales podrían haberse adaptado a datos espaciales, dejando de lado los aspectos que se refieren a las réplicas (Verbeke y Molenberghs, 2000). No obstante, los modelos para datos longitudinales y para datos espaciales se desarrollaron por separado.

Diggle sugiere, en la ultima publicacion del compendio de Gregoire et al. (1997), que el desarrollo independiente se debió a que la pregunta científica que motivó el desarrollo de la metodología difiere según el tipo del dato: en las aplicaciones espaciales importa el análisis de una única realización de un proceso subyacente, mientras que en los estudios lontigudinales en general aparecen datos replicados y/o en el contexto de un experimento.

El *modelo de regresión generalizado mixto* (MLGM) fue propuesto en 1993 por Breslow y Clayton como un marco unificador para modelos de regresión generalizados capaces de tratar con datos correlacionados. Específicamente, el MLGM es una extensión del MLG que incorpora efectos aleatorios al predictor lineal para tratar las variabilidad que no puede ser explicada por las covariables. La formulación del modelo lineal generalizado mixto es muy flexible: *lineal generalizado* hace referencia a su capacidad para modelar respuestas no normales con una estructura lineal en los parámetros, mientras que *mixto* permite acomodar tanto los contextos longitudinal como espacial (Gregoire et al., 1997).

El MLGM se formula como sigue. Sea $\boldsymbol{b}$ un vector de efectos aleatorios de dimensión q y $\boldsymbol{z}_i$ el vector de covariables asociado a dichos efectos. Dado $\boldsymbol{b}$, las y_i se consideran independientes con distribución en la familia exponencial con media $E[y_i|\boldsymbol{b}]$ y varianza $Var(y_i|\boldsymbol{b}) = \phi a_i V(\mu_i)$. En la

expresión de la varianza, a_i es una constante conocida, ϕ es el parámetro de dispersión y $V(\cdot)$ es la función de varianza (Breslow y Clayton, 1993).

Dado el predictor lineal $\eta_i = \boldsymbol{x}_i^{'}\boldsymbol{\alpha} + \boldsymbol{z}_i^{'}\boldsymbol{b}$, la media condicional $\mu_i = E[y_i|\boldsymbol{b}]$ se relaciona con η_i a través de la función de enlace $g(\mu_i) = \eta_i$. El modelo se completa especificando una distribución normal multivariada para los efectos aleatorios $\boldsymbol{b} \sim N(0, \boldsymbol{D}(\boldsymbol{\theta}))$ donde $\boldsymbol{\theta}$ es un vector de componentes de varianza desconocidos.

La Ecuación 30 muestra la formulación de un MLGM con respuesta normal, también llamado *modelo lineal mixto.*

$$\begin{aligned} y_i|\boldsymbol{\alpha}, \boldsymbol{b}, \sigma^2 &\sim N(\boldsymbol{x}_i^{'}\boldsymbol{\alpha} + \boldsymbol{z}_i^{'}\boldsymbol{b}, \sigma^2) \\ \eta_i &= \boldsymbol{x}_i^{'}\boldsymbol{\alpha} + \boldsymbol{z}_i^{'}\boldsymbol{b} \\ g(\mu_i) &= \mu_i \\ \boldsymbol{b} &\sim N(0, \boldsymbol{D}(\boldsymbol{\theta})) \end{aligned} \tag{30}$$

La inferencia para los efectos fijos en los GLMM requiere realizar una integración multidimensional sobre la distribución de los efectos aleatorios, lo que representa una dificultad computacional. Por ende, se han planteado múltiples enfoques para resolver los MLGM (Schabenberger y Gotway, 2005):

- Estimación por cuasi-verosimilitud. En este caso no se requiere la verosimilitud completa, solo los primeros dos momentos.
- Estimación con pseudo-verosimilitud, que involucra una expansión en series de Taylor.
- Estimación por cuasi-verosimilitud penalizada o marginal (Breslow y Clayton, 1993).
- Estimación por máxima verosimilitud, mediante por ejemplo aproximaciones de Laplace (Pinheiro y Bates, 1995; 2000) y métodos de cuadratura gaussiana adaptativa (Pinheiro y Bates, 1995).
- Modelos jerárquicos bayesianos.

2.3.4. Modelo de regresión jerárquico bayesiano

El término *jerárquico* se encuentra en la literatura asociado a dos interpretaciones. La primera de ellas hace referencia a la estructura jerárquica o multinivel de los datos, como podría ser el caso de alumnos anidados en escuelas o de medidas repetidas para el mismo sujeto. La segunda interpretación se refiere a la **formulación condicional** de un modelo en etapas. En este trabajo, cuando hablamos de *jerárquico* nos referimos a la segunda interpretación, dejando el término *multinivel* para la estructura de datos agrupados.

La formulación del modelo jerárquico bayesiano (MJB) suele hacerse en tres etapas (Albert, 2009):

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\phi},\boldsymbol{\theta} &\sim f(\boldsymbol{y}|\boldsymbol{\phi},\boldsymbol{\theta}) \\ \boldsymbol{\phi}|\boldsymbol{\theta} &\sim \pi(\boldsymbol{\phi}|\boldsymbol{\theta}) \\ \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}) \end{aligned} \tag{31}$$

La primera etapa especifica la verosimilitud de los datos independientes condicionalmente a los parámetros $\boldsymbol{\phi}$ e hiperparámetros $\boldsymbol{\theta}$ dados; la segunda etapa especifica la distribución a priori de los parámetros en función de los hiperparámetros, y la tercera etapa especifica la distribución a priori de los hiperparámetros (Allenby et al., 2005; Albert, 2009). En general, la tercera etapa debe estar presente para que el modelo se considere un MJB.

En el caso del MLGM estimado con métodos bayesianos, donde todos los parámetros se consideran variables aleatorias con distribución a priori, su formulación condicional permite plantearlo como un modelo jerárquico bayesiano de manera natural. En estos modelos, los hiperparámetros suelen ser los parámetros de la matriz de covarianza de los efectos aleatorios.

La Ecuación 32 muestra la formulación del MLGM normal planteado como un modelo jerárquico bayesiano, asumiendo que el efecto aleatorio $\boldsymbol{b}$ es independiente de los errores aleatorios y del efecto de las covariables fijas.

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{b}, \sigma^2, \boldsymbol{\theta} &\sim N(\boldsymbol{x}_i^{'}\boldsymbol{\alpha} + \boldsymbol{z}_i^{'}\boldsymbol{b}, \sigma^2\boldsymbol{I} + \boldsymbol{D}(\boldsymbol{\theta})) \\
\boldsymbol{\alpha}, \boldsymbol{\sigma^2} &\sim \pi(\boldsymbol{\alpha}, \sigma^2)) \\
\boldsymbol{b} &\sim N(0, \boldsymbol{D}(\boldsymbol{\theta})) \\
\boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta})
\end{aligned}
\tag{32}
$$

El motivo por el que los modelos jerárquicos bayesianos están en boga en la actualidad es que recién en los años 1970 con la aparición de los algoritmos de MCMC se hizo posible realizar el cálculo de las distribuciones a posteriori para modelos complejos. Antes solo era posible para los casos más sencillos, que tenían distribuciones a posteriori con forma analítica. Aún así, MCMC es computacionalmente demandante y difícil de implementar para el investigador no especializado, por lo que el surgimiento de INLA en 2009 aceleró y facilitó la inferencia en estos modelos.

2.4. Modelos de regresión jerárquicos bayesianos para datos espaciales

En el análisis de datos espaciales es común ver las etapas del modelo jerárquico bayesiano organizadas levemente diferente a la estructura típica verosimilitud - parámetros - hiperparámetros de la Ecuación 31. Por ejemplo, Banerjee et al. (2005) y Schabenberger y Gotway (2005) plantean siguiente el modelo genérico para modelos jerárquicos bayesianos para datos espaciales:

$$[\,datos\,|\,proceso,\ parámetros\,]\,[\,proceso\,|\,parámetros\,]\,[\,parámetros\,].$$

En este modelo, la primera etapa especifica la distribución de los datos $\boldsymbol{Z}(\boldsymbol{s})$ condicional a un *proceso espacial* $\boldsymbol{S}(\boldsymbol{s})$ y al vector $\boldsymbol{\phi}$ de parámetros e hiperparámetros. El vector $\boldsymbol{\phi}$ incluye los efectos de las covariables, los parámetros de la verosimilitud y el vector de parámetros $\boldsymbol{\theta}$ de la matriz de covarianzas del efecto espacial $\boldsymbol{S}(\boldsymbol{s})$, que son los que se suelen denominar *hiperparámetros*. La segunda etapa modela el proceso espacial $\boldsymbol{S}(\boldsymbol{s})$ condicional a $\boldsymbol{\theta}$. La tercera etapa especifica la distribución a priori de $\boldsymbol{\phi}$ (Banerjee et al., 2005).

La Ecuación 33 plantea el modelo de regresión lineal clásico estimado con métodos bayesianos para datos espaciales. Si bien se les asignan distribuciones a priori a los parámetros $\boldsymbol{\beta}$ y σ^2, este

modelo no se puede plantear bajo la estructura del modelo *jerárquico* bayesiano en tres etapas porque no hay efectos aleatorios.

$$\begin{aligned} \boldsymbol{Z}(\boldsymbol{s}) &= \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{e}(\boldsymbol{s}) \\ \boldsymbol{e}(\boldsymbol{s}) &\sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \\ \boldsymbol{\beta}, \sigma^2 &\sim \pi(\boldsymbol{\beta}, \sigma^2) \end{aligned} \tag{33}$$

En general, el consenso es que este último modelo solo es válido para observaciones independientes. Schabenberger y Gotway (2005) explican que el modelo lineal normal con errores independientes solo puede incorporar la estructura espacial en la función de medias; por lo tanto, si esta función no logra explicar toda la heterogeneidad espacial, los residuos mostrarán un patrón distinto al aleatorio, y se obtendrán efectos indeseados en los resultados, tales como estimadores ineficientes de los parámetros. Dichos autores sugieren que, sin embargo, podría ser correcto aplicar este modelo a datos espaciales en ciertos casos, por ejemplo, en el contexto de un experimento, donde la dependencia espacial se haya neutralizado por diseño a través de técnicas de aleatorización.

Los GLMM, que sí pueden manejar correlación espacial, se pueden formular como modelos jerárquicos bayesianos de manera natural gracias a su formulación condicional original. En el GLMM para datos espaciales, $\boldsymbol{Z}(\boldsymbol{s})$ tiene distribución en la familia exponencial condicional a $\boldsymbol{S}(\boldsymbol{s})$ y las observaciones se asumen independientes (Schabenberger y Gotway, 2005).

La media de los datos $\boldsymbol{\mu}(\boldsymbol{s}) = E[\boldsymbol{Z}(\boldsymbol{s})|\boldsymbol{S}(\boldsymbol{s})]$ se define condicional y se relaciona con las covariables y con el proceso espacial $\boldsymbol{S}(\boldsymbol{s})$ a través de la función de enlace:

$$g[\boldsymbol{\mu}(\boldsymbol{s})] = \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{S}(\boldsymbol{s}), \tag{34}$$

de manera que $\boldsymbol{S}(\boldsymbol{s})$ es un intercepto aleatorio en la ubicación $\boldsymbol{s}$.

Al asumir datos condicionalmente independientes, la autocorrelación espacial solo entra al modelo a través de $\boldsymbol{S}(\boldsymbol{s})$. Generalmente, se asume que $\boldsymbol{S}(\boldsymbol{s})$ es un campo aleatorio gaussiano con media cero y función de covarianzas $\sigma_s^2 \rho_s(\boldsymbol{s}_i - \boldsymbol{s}_j)$ y la función de correlación $\rho_s(\boldsymbol{s}_i - \boldsymbol{s}_j)$ se modela con q parámetros incluidos en el vector $\boldsymbol{\theta_s}$ (Schabenberger y Gotway, 2005):

$$Corr[\boldsymbol{S}(\boldsymbol{s}_i), \boldsymbol{S}(\boldsymbol{s}_j)] = \rho_s(\boldsymbol{s}_i - \boldsymbol{s}_j; \boldsymbol{\theta}) \tag{35}$$

No es obligatorio que la estructura de covarianzas de $\boldsymbol{S}(\boldsymbol{s})$ tenga la forma de la Ecuación 35, pero sí que sea una matriz definida positiva (Schabenberger y Gotway, 2005).

La Ecuación 36 muestra la formulación jerárquica bayesiana en tres etapas del MLGM para datos espaciales, donde f pertenece a la familia exponencial.

$$\begin{aligned} \boldsymbol{Z}(\boldsymbol{s})|\boldsymbol{S}(\boldsymbol{s}), \boldsymbol{\phi} &\sim f(g^{-1}(\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{S}(\boldsymbol{s}))) \\ \boldsymbol{S}(\boldsymbol{s})|\boldsymbol{\theta} &\sim N(\boldsymbol{0}, \boldsymbol{\Sigma_s}(\boldsymbol{\theta})) \\ \boldsymbol{\phi} &\sim \pi(\boldsymbol{\phi}) \end{aligned} \tag{36}$$

Si f es gaussiana, la Ecuación 36 se transforma en el modelo jerárquico bayesiano de la Ecuación 37, donde $\boldsymbol{\beta}$ es el vector de efectos de las covariables y σ^2 representa la varianza del error de las observaciones.

$$\begin{aligned} \boldsymbol{Z}(\boldsymbol{s})|\boldsymbol{S}(\boldsymbol{s}), \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 &\sim N(\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{S}(\boldsymbol{s}), \sigma^2\boldsymbol{I}) \\ \boldsymbol{S}(\boldsymbol{s})|\boldsymbol{\theta} &\sim N(\boldsymbol{0}, \boldsymbol{\Sigma_s}(\boldsymbol{\theta})) \\ \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 &\sim \pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2), \end{aligned} \tag{37}$$

La elección de la función de covarianzas $\boldsymbol{\Sigma_s}(\boldsymbol{\theta})$ dependerá de si se quiere darle un enfoque geoestadístico o para datos regionales al problema. En el enfoque geoestadístico se suele utilizar $\boldsymbol{\Sigma_s}(\boldsymbol{\theta}) = \sigma_s^2 \rho_s(\boldsymbol{s}_i - \boldsymbol{s}_j; \boldsymbol{\theta})$, donde la función de correlación $\rho_s(\cdot)$ se especifica siguiendo algún modelo de correlación en función de la distancia para datos geoestadísticos.

Para modelar los efectos espaciales en datos regionales se puede plantear la función de covarianzas en función de la distancia entre centroides de las áreas, lo que equivale a utilizar la especificación de covarianzas como en el enfoque geoestadístico, o se puede especificar la covarianza en función de la estructura de vecindad. En esta tesis solo trabajamos con estructuras de vecindarios.

A continuación se definen tres formas de especificar los efectos aleatorios en el modelo jerárquico bayesiano para respuesta normal y datos espaciales regionales.

2.4.1. Efecto aleatorio no estructurado, normal e idénticamente distribuido

Cuando $\boldsymbol{\Sigma_s(\theta)} = \sigma_u^2 \boldsymbol{I}$ en la Ecuación 37, se obtiene un modelo lineal mixto con respuesta normal común, pero aplicado a datos espaciales y estimado con métodos bayesianos:

$$\begin{aligned} \boldsymbol{Z}(\boldsymbol{s})|\boldsymbol{S}(\boldsymbol{s}), \sigma_u^2, \boldsymbol{\beta}, \sigma^2 &\sim N(\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{S}(\boldsymbol{s}), \sigma^2 \boldsymbol{I}) \\ \boldsymbol{S}(\boldsymbol{s})|\sigma_u^2 &\sim N(\boldsymbol{0}, \sigma_u^2 \boldsymbol{I}) \\ \sigma_u^2, \boldsymbol{\beta}, \sigma^2 &\sim \pi(\sigma_u^2, \boldsymbol{\beta}, \sigma^2). \end{aligned} \tag{38}$$

El efecto aleatorio se distribuye condicionalmente normal, con media cero, varianza σ_u y matriz de correlación identidad, lo que implica que los efectos aleatorios se asumen *independientes e idénticamente distribuidos* (*iid*) en cada región. El efecto aleatorio no tiene estructura espacial explícita y por ello se lo denomina *no estructurado*.

2.4.2. Efecto aleatorio estructurado condicional autorregresivo intrínseco (iCAR)

Besag (1974) propuso modelar la variabilidad espacial en cada región incluyendo un efecto espacial estructurado $\boldsymbol{u}$ en el predictor lineal tal que

$$\eta_i = \sum_{k=1}^{p} x_{ik}\beta_k + u_i, \tag{39}$$

donde $\sum_{k=1}^{p} x_{ik}\beta_k$ es la estructura de medias que contiene los efectos fijos de las covariables y u_i es el efecto espacial estructurado. El modelo Besag le asigna una estructura *condicional autorregresiva* a $\boldsymbol{u}$:

$$u_i|\boldsymbol{u}_{-i} \sim N(\mu_i + \sum_{j=1}^{n} r_{ij}(u_j - \mu_j), s_i^2) \tag{40}$$

donde la media del área es μ_i y la varianza $s_i^2 = \frac{\sigma_u^2}{N_i}$ es una fracción de del parámetro de varianza σ_u^2. Definiendo la precisión como la inversa de la varianza $\tau = 1/\sigma^2$, resulta que a mayor cantidad de vecinos, menor varianza y mayor precisión tiene el efecto espacial en la región.

El coeficiente r_{ij} indica la proximidad espacial:

$$r_{ij} = \phi W_{ij} = \phi \frac{a_{ij}}{N_i} \tag{41}$$

donde $a_{ij} \neq 0$ solo si las áreas i y j son vecinas. Es decir, r_{ij} permitirá que solo los vecinos contribuyan a la media de la distribución del efecto espacial en una región, y a la vez fijará el peso de cada contribución, donde este peso está basado en la cantidad de vecinos de la región.

El parámetro ϕ controla qué tan propia es la distribución de $\boldsymbol{u}$. Para que la distribución sea propia, ϕ debe tomar valores entre $min\{\lambda_i\}$ y $max\{\lambda_i\}$, $i = 1, ..., n$, donde los λ_i son los eigenvalores de la matriz de pesos espaciales $\boldsymbol{W}$ (ver Sección 2.1). En este caso se puede escribir la especificación *condicional autorregresiva propia* como

$$\boldsymbol{u} \sim N(\boldsymbol{\mu}, (\boldsymbol{I} - \phi \boldsymbol{W})^{-1} \boldsymbol{S}^2) \tag{42}$$

Como ϕ puede ser difícil de estimar, la especificación anterior no suele utilizarse tanto y en su lugar se usa una versión simplificada que fija $\phi = 1$ (Blangiardo y Cameletti, 2015). En este caso la distribución de $\boldsymbol{u}$ no será propia, porque al elegir $\phi = 1$ su matriz de covarianzas no es definida positiva. Esta especificación se llama *condicional autorregresiva intrínseca* o *iCAR*, del ingles *intrinsic conditional autorregesive.*

La matriz de covarianzas $iCAR$ genera una distribución condicional para u_i

$$u_i | \boldsymbol{u}_{-i} \sim N(\mu_i + \frac{1}{N_i} \sum_{j=1}^{n} a_{ij}(u_j - \mu_j), s_i^2) \tag{43}$$

donde $r_{ij} = \phi W_{ij} = \frac{a_{ij}}{N_i}$. Como no hay una distribucion conjunta propia para $\boldsymbol{u}$, se podria agregar una constante a los u_i sin cambiar la distribucion, y para solucionar este aspecto se suele agregar una restriccion como $\sum_{i=1}^{n} u_i = 0$. En general se suele asumir $\mu_i = 0$ (Blangiardo y Cameletti, 2015).

La especificación del modelo jerárquico bayesiano con efecto aleatorio espacial $iCAR$ se muestra en la Ecuación 44.

$$
\begin{aligned}
Z(\boldsymbol{s}_i)|\boldsymbol{u}, \tau_u, \boldsymbol{\beta}, \sigma^2 &\sim N(\boldsymbol{x}(\boldsymbol{s}_i)^T\boldsymbol{\beta} + u(\boldsymbol{s}_i), \tau^{-1}\boldsymbol{I}) \\
u_i|\boldsymbol{u}_{-i} &\sim N(\frac{1}{N_i}\sum_{j=1}^{n} a_{ij}u_j, \frac{1}{\tau_u N_i}) \\
\tau_u, \boldsymbol{\beta}, \tau &\sim \pi(\tau_u, \boldsymbol{\beta}, \tau).
\end{aligned} \tag{44}
$$

2.4.3. Efecto aleatorio estructurado BYM

El modelo BYM (por Besag, York y Mollié, 1991) es una mejora al modelo con efecto espacial $iCAR$ que considera una posible variabilidad individual $\boldsymbol{v}$ en cada región, además de la variabilidad espacial estructurada $\boldsymbol{u}$ proveniente de similitudes entre vecinos. Para incluir la variabilidad individual se agrega un intercepto aleatorio v_i no estructurado en cada región al predictor lineal, que resulta ser

$$
\eta_i = \sum_{k=1}^{p} x_{ik}\beta_k + u_i + v_i. \tag{45}
$$

Los efectos espaciales $\boldsymbol{u_i}$ se modelan con la especificación iCAR, mientras que el efecto aleatorio no estructurado $\boldsymbol{v}$ se asume iid, es decir, intercambiable entre regiones y con distribución

$$
v_i \sim N(0, \tau_v^{-1}). \tag{46}
$$

La especificación del BYM como un modelo jerárquico bayesiano se muestra en la Ecuación 47.

$$
\begin{aligned}
Z(\boldsymbol{s}_i)|\boldsymbol{u}, \boldsymbol{v}, \tau_u, \tau_v, \boldsymbol{\beta}, \sigma^2 &\sim N(\boldsymbol{x}(\boldsymbol{s}_i)^T\boldsymbol{\beta} + u(\boldsymbol{s}_i) + v(\boldsymbol{s}_i), \tau^{-1}\boldsymbol{I}) \\
u_i|\boldsymbol{u}_{-i} \sim N(\frac{1}{N_i}\sum_{j=1}^{n} a_{ij}u_j, \frac{1}{\tau_u N_i}); \quad v_i &\sim N(0, \tau_v^{-1}) \\
\tau_u, \tau_v, \boldsymbol{\beta}, \tau &\sim \pi(\tau_u, \tau_v, \boldsymbol{\beta}, \tau).
\end{aligned} \tag{47}
$$

Hay dos problemas con el modelo BYM y la elección de distribuciones a priori. En primer lugar, el componente espacial no está escalado. La varianza marginal luego de incorporar la restricción $\sum u_i = 0$ no está estandarizada, lo que implica que no se pueden usar las mismas distribuciones

a priori para distintos grafos, porque la varianza generalizada depende del grafo (Simpson et al., 2017). En segundo lugar, el componente espacial $\boldsymbol{u}$ no se puede ver independientemente del componente no estructurado $\boldsymbol{v}$ y por lo tanto las distribuciones a priori para sus precisiones deberían ser dependientes (Simpson et al., 2017).

El modelo *BYM2* (Riebler et al., 2016) resuelve ambos problemas mediante una reparametrización de los efectos aleatorios. El BYM2 define un componente espacial estructurado $\boldsymbol{u}^*$ donde la varianza generalizada es 1 y se ha calculado como la media geométrica de las varianzas marginales. El predictor lineal en el BYM2 se escribe como

$$\eta_i = \sum_{k=0}^{p} \boldsymbol{x}_k^{'} \beta_k + \frac{1}{\sqrt{\tau_s}} (\sqrt{1-\phi} v_i + \sqrt{\phi} u_i^*) \tag{48}$$

donde τ_s es la precisión marginal del efecto aleatorio y $0 \leq \phi \leq 1$ controla la cantidad de efecto espacial: cuando $\phi = 0$ no hay efecto espacial y cuando $\phi = 1$ todo el efecto aleatorio es espacial.

La contribución marginal a la precisión de $\boldsymbol{u}^*$ y $\boldsymbol{v}$ es $\frac{1}{\tau_s}$. La proporción de esta varianza explicada por el efecto espacial es ϕ, y la proporción explicada por el efecto no estructurado es $1 - \phi$. Si τ_s se mantiene constante, a medida que ϕ aumenta mayor parte de la variabilidad será explicada por el efecto espacial y menor por el efecto no estructurado.

La reparametrización propuesta por el BYM2 permite utilizar distribuciones a priori independientes para τ_s y ϕ, y permite ver al modelo como una extensión flexible de dos modelos base (Simpson et al., 2017).

2.5. Elección de distribuciones a priori

La aplicación del Teorema de Bayes a la inferencia bayesiana ha sido cuestionada porque implica incorporar información subjetiva al modelo a través de las distribuciones a priori, ya que estas representan el conocimiento del investigador sobre los parámetros previo a ver los datos. Por otro lado, las distribuciones a posteriori pueden ser muy sensibles a la elección de la distribución a priori, especialmente si se tienen pocos datos o si las distribuciones a priori son muy informativas. En los modelos complejos, como lo son los modelos jerárquicos bayesianos, donde podría haber mas parámetros que observaciones, es la especificación de las distribuciones a priori la que

controla el equilibrio entre precisión predictiva y parsimonia (Simpson et al., 2017) y por tanto el sobreajuste.

Resulta entonces que la elección de las distribuciones a priori es vital. Hay dos aspectos de ellas a considerar: el tipo de distribución, que dependerá de la naturaleza del parámetro, y los hiperparámetros, que manejan el nivel de información que se tendrá para los parámetros (Blangiardo y Cameletti, 2015).

Cuando el investigador no tiene información sobre los parámetros en estudio y quiere dejar que los datos hablen por sí mismos, la distribución a priori debería asignar igual densidad a todos los posibles valores del parámetro. Este tipo de distribuciones a priori se denominan *no informativas*.

Como es difícil encontrar una buena distribución a priori no informativa para todos los posibles valores del parámetro que no haga que la distribución a posteriori se vuelva impropia, una estrategia consiste en utilizar distribuciones *vagas*, que son distribuciones no informativas en el subconjunto de los parámetros donde la verosimilitud no es cero. Por ejemplo, se podría utilizar una distribución normal con media cero y varianza muy grande para una media o un parámetro de regresión, y una $Gamma(a, b)$ con a, b pequeños para el parámetro de precisión de la matriz de covarianzas (Blangiardo y Cameletti, 2015).

En los modelos jerárquicos bayesianos para datos espaciales la elección de las distribuciones a priori para los hiperparámetros cobra especial importancia porque define el grado de suavizado del efecto espacial. Por ejemplo, en el modelo $iCAR$ para datos regionales la distribución a priori de los efectos espaciales se define condicional a un parámetro de precisión τ_u (Ver Ecuación 43). Si a τ_u se le asigna una distribución a priori $Gamma(a, b)$, el grado de suavizado del efecto espacial estará gobernado por el parámetro de forma a de la distribución $Gamma$ (Illian et al., 2012).

Una de las ideas principales de Simpson et al. (2017) es que una distribución a priori que no ponga suficiente masa en el *modelo base* fuerza el sobreajuste. Por ejemplo, en un modelo con efectos aleatorios *iid* o iCAR, el modelo base sería el modelo sin efectos aleatorios, lo que se traduce en una precisión infinita para el efecto aleatorio. Todas las distribuciones $Gamma(a, b)$ con valor esperado $a/b < \infty$ utilizadas como distribuciones a priori para esta precisión apoyarán la tendencia al sobreajuste del modelo, porque asignan cero probabilidad a priori a la inexistencia del efecto aleatorio (Simpson et al., 2017).

Simpson, Rue, Riebler, Martins y Sørbye (2017) proponen un marco de construcción de distribuciones a priori informativas para modelos jerárquicos bayesianos que le permiten al usuario controlar la complejidad o sobreajuste de un modelo penalizando su desviación respecto a un modelo simple o base. Las distribuciones a priori construidas bajo este marco se denominan *distribuciones a priori penalizadas por complejidad.*

A continuación se describen las distribuciones a priori por defecto en `R-INLA` y las distribuciones penalizadas por complejidad de Simpson et al. (2017).

2.5.1. Distribuciones a priori por defecto en `R-INLA`

En INLA el campo latente debe ser un GMRF (Ver Ecuación 10). Por lo tanto, los componentes del campo latente de un modelo ajustado con INLA deben tener distribuciones a priori gaussianas. En los modelos de este trabajo, el campo latente comprende los parámetros de las covariables.

`R-INLA` asigna por defecto una distribución a priori $N(0, 10E5)$ para los parámetros de regresión y una distribución a priori $Gamma(1, 10E-5)$ para la precisión del error aleatorio de las observaciones. En ambos casos es posible modificar los parámetros de las distribuciones, y en el caso de la precisión del error, también es posible modificar la familia de la distribución.

En el caso de modelos con efectos espaciales especificados en función de una precisión, `R-INLA` asigna por defecto una distribución a priori $Gamma(1, 10E-5)$ para las precisión. En este caso también es posible modificar la familia y la parametrización de esta distribución.

2.5.2. Distribuciones a priori penalizadas por complejidad

En el marco de las distribuciones penalizadas por complejidad, el modelo base se define formalmente como el modelo más simple en la clase de modelos $p(\boldsymbol{x}|\boldsymbol{\xi})$, donde $\boldsymbol{x}$ son los datos y $\boldsymbol{\xi}$ es un parámetro de flexibilidad, que suele ser un escalar no-negativo. En esta especificación, el modelo base corresponde al modelo con $\boldsymbol{\xi} = 0$, es decir, a un modelo con cero flexibilidad. Por lo tanto, $p(\boldsymbol{x}|\boldsymbol{\xi})$ se puede interpretar como una extensión flexible del modelo base, donde valores crecientes de $\boldsymbol{\xi}$ permiten desviaciones cada vez mayores del modelo base.

Se define la distancia entre dos modelos con densidades f y g como

$$d(f||g) = \sqrt{2KLD(f||g)} \tag{49}$$

donde KLD es la divergencia de Kullback-Leibler (Simpson et al., 2017). La distancia d se interpreta como una medida de complejidad del modelo f comparado con el g. Como su nombre lo indica, las distribuciones penalizadas por complejidad buscan *penalizar* la complejidad, y por lo tanto penalizarán a los modelos en función de su distancia d al modelo base.

La penalización a la complejidad se hace con una *tasa de decaimiento constante*, lo que significa que la tasa de penalización no varía con la diferencia en complejidad, es decir, no varía para distintas áreas del espacio de parámetros (Simpson et al., 2017). Esto permite definir una distribución a priori exponencial en la escala de d

$$\pi(d) = \lambda exp(-\lambda d) \tag{50}$$

que se puede llevar fácilmente a la escala del parámetro de flexibilidad (univariado, en este caso), ξ:

$$\pi(\xi) = \lambda e^{-\lambda d(\xi)} \left| \frac{\partial d(\xi)}{\partial \xi} \right| \tag{51}$$

La tasa de penalización está controlada por λ a través de la siguiente condición:

$$p(Q(\xi) > U) = \alpha \tag{52}$$

donde $Q(\xi)$ es una función del parámetro de flexibilidad ξ, U es un límite superior definido por el usuario que marca el 'evento de cola' de la distribución, y α es el peso que el usuario le asigna a este evento. De esta manera, modificando U y α el usuario puede controlar la cantidad de masa que se le da a la cola de la distribución a priori.

La derivación de las distribuciones a priori penalizadas por complejidad es distinta para cada tipo de parámetro. En la publicación original de Simpson et al. (2017) se pueden encontrar los detalles. Aquí solo mencionamos los resultados para el **parámetro de precisión** τ de los efectos aleatorios, y para el **parámetro de mezcla** ϕ del modelo BYM.

Distribución para la precisión del efecto aleatorio

En un modelo con efectos aleatorios $\boldsymbol{u}$ normales con precisión τ el modelo base corresponde a la ausencia de efectos aleatorios, es decir, a efectos aleatorios que tienen variabilidad nula. Esto se expresa matemáticamente como $\sigma = 0$ en la escala de la desviación estándar o como $\tau = \infty$ en la escala de la precisión del efecto aleatorio, porque $\tau = 1/\sigma^2$.

La condición de la Ecuación 52 se transforma en

$$Prob(\frac{1}{\sqrt{\tau}} > U) = Prob(\sigma > U) = \alpha, \tag{53}$$

la tasa de penalización para la desviación estándar σ es

$$\lambda = \frac{-ln(\alpha)}{U}. \tag{54}$$

y la distribución penalizada por complejidad para τ tiene la forma de la Ecuación 55.

$$\pi(\tau) = \frac{\lambda}{2}\tau^{-3/2}exp(-\lambda\tau^{-1/2}), \quad \tau > 0, \quad \lambda > 0 \tag{55}$$

La Figura 1 muestra la forma de la distribución penalizada por complejidad para la precisión para distintos valores de U y α. La Figura 24b muestra cuatro valores de U para $\alpha = 0.5$, y la Figura 24d muestra los mismos cuatro valores de U para $\alpha = 0.8$.

Aumentar α hace que la densidad de las curvas se apile cerca de $\tau = 0$ y asigne a priori mayor probabilidad a valores pequeños de τ y valores grandes de σ. Estas densidades a priori permiten que el modelo sea muy flexible - que el efecto aleatorio absorba mucha variabilidad. Aumentar U también hace que las curvas pongan mayor densidad en valores cercanos a cero y por tanto también se traduce en permitir a priori modelos más flexibles. El comportamiento visto es resultado de la proposición de la Ecuación 53, que maneja la forma de las densidades.

Distribución para el parámetro de mezcla del modelo BYM2

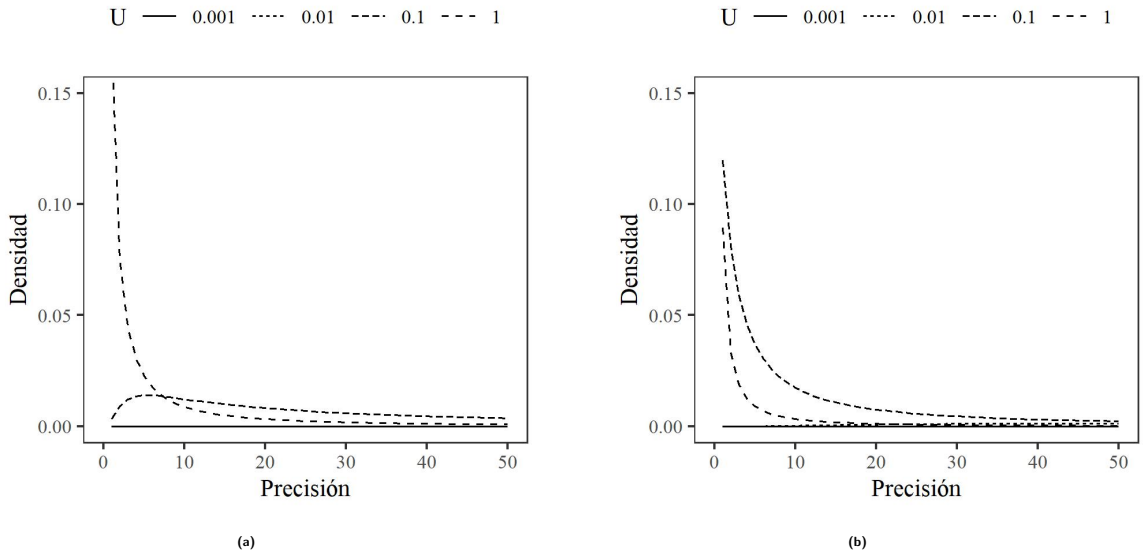

Figura 1: Gráficos de la densidad de la distribución a priori penalizada por complejidad para la precisión para distintos valores de U y (a) $\alpha = 0.5$, (b) $\alpha = 0.8$.

En BYM2 hay dos hiperparámetros cuya distribución a priori se puede modificar: la precisión marginal τ y el parámetro de mezcla ϕ. En BYM2 las distribuciones a priori se formulan en dos niveles. Primero se define la distribución para τ, que controla la contribución a la varianza marginal de los efectos aleatorios $\boldsymbol{v}$ y $\boldsymbol{u*}$. Segundo, se distribuye la varianza marginal entre los efectos espaciales y los efectos no estructurados mediante la distribución a priori de ϕ.

El modelo más simple es el *modelo base de primer nivel*, que asume precisión infinita o ausencia de variabilidad espacial. La forma de la distribución a priori de la precisión τ ya se describió en la Ecuación 55. La distribución de ϕ no es importante en el modelo base. Sin embargo, cuando hay una desviación del modelo base a favor de un modelo mas flexible, significa que hay algo de variabilidad espacial, y la distribución de ϕ empieza a influir.

Condicional a una precisión marginal fija, el *modelo base de segundo nivel* es el modelo que asume variabilidad espacial independiente, específica de la región, lo que equivale a $\phi = 0$. El grado de penalización de la distribución de ϕ se controla con la condición

$$P(\phi < U) = \alpha, \tag{56}$$

que se interpreta como: "la probabilidad de que el parámetro de mezcla ϕ sea menor a U es α". Manteniendo fija la cota superior U, cuando aumenta α, aumenta la confianza en la proposición de que ϕ está acotado por debajo de U.

La distribución a priori penalizada por complejidad para ϕ depende del grafo definido por las regiones en los datos y no tiene forma analítica (Riebler et al., 2016). En INLA es posible obtenerlas a partir del comando *INLA:::inla.pc.bym.phi*.

La Figura 2 muestra la forma que toma la distribución a priori penalizada por complejidad para los datos de ilustración de esta tesis. Las Figuras 2a, 2b y 2c muestran las curvas para distintos U fijando $\alpha = 0.01$, $\alpha = 0.5$ y $\alpha = 0.8$ respectivamente.

El valor de α es muy importante para determinar la forma de la distribución. Cuando α es alto, el nivel de confianza en el conocimiento a priori es alto, y el conocimiento a priori en este caso dice que el parámetro de mezcla es menor a cierto umbral U. Cuanto mayor sea ϕ, mayor porcentaje de la variabilidad será explicada por el efecto aleatorio espacial. Por lo tanto, si se limita el valor de ϕ con la cota superior U, se está restringiendo el porcentaje de variabilidad estructurada espacialmente. Con $\alpha = 0.5$ y $\alpha = 0.8$ se consiguen distribuciones que ponen mucho peso en $\phi = 0$ (ver Figuras 2b y 2c). Solo con $\alpha = 0.01$ fue posible lograr distribuciones a priori que pongan mucho peso en $\phi = 1$ (ver Figura 2a).

2.6. Validación y comparación de modelos bayesianos

En el proceso de modelado se busca evitar dos tipos principales de error estadístico: sobreajuste y subajuste. Ambos generan predicciones de mala calidad: el modelo que sobreajusta ha aprendido demasiado de los datos, y el que subajusta, demasiado poco (McElreath, 2020). Resulta entonces esencial evaluar la plausibilidad del modelo luego de la estimación, para examinar el ajuste y buscar indicios de violación de algún supuesto. Este proceso recibe el nombre de *validación del modelo*.

La *comparación de modelos* es la etapa que permite comparar el ajuste y capacidad predictiva de un modelo con la de otros, con fines de encontrar el "mejor modelo". La etapa de comparación no proporciona información absoluta sobre la calidad del modelo, sino relativa a la calidad de otros modelos competidores.

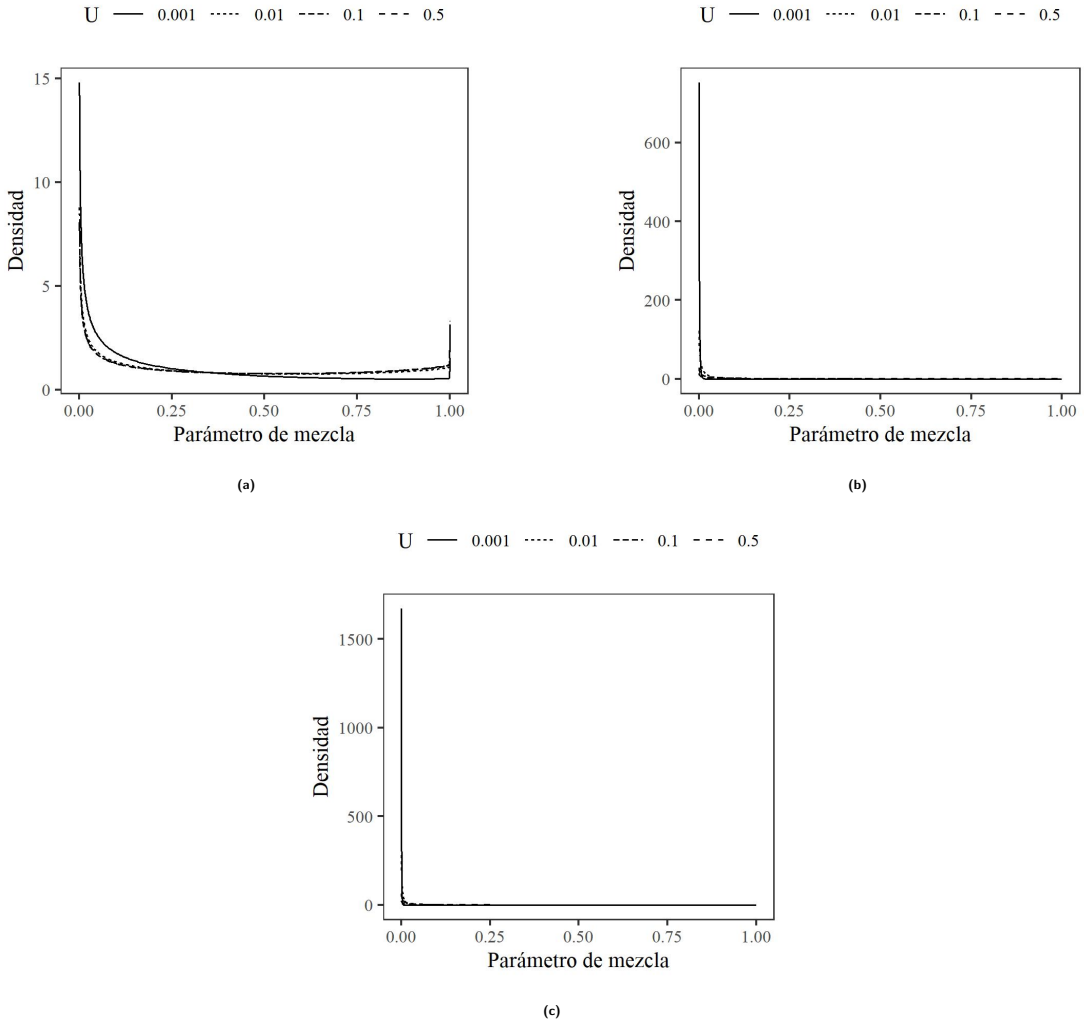

Figura 2: Gráficos de la densidad de la distribución a priori penalizada por complejidad para el parámetro de mezcla ϕ para distintos valores de U y (a) $\alpha = 0.01$, (b) $\alpha = 0.5$ y (c) $\alpha = 0.8$.

Una manera de evaluar un modelo es a través de la *precisión predictiva* (Gelman, 1996). Idealmente, la precisión predictiva se estima con datos que no se hayan usado para entrenar al modelo. En su defecto, cuando no se dispone de nuevos datos, se la estima con los datos de entrenamiento, pero corrigiendo por el sesgo en el que se incurre al evaluar las predicciones del modelo con mismos datos que se usaron para entrenarlo (Gelman et al., 2014). Dos enfoques que utilizan datos de entrenamiento para el cálculo de la precisión predictiva son los *criterios de información* y las *técnicas de validación cruzada*.

A continuación se describen las técnicas de validación y comparación más frecuentemente utilizadas en los modelos bayesianos.

2.6.1. Validación del modelo

La validación del modelo es el paso de verificación del ajuste del modelo a los datos. Durante la validación se analiza la plausibilidad del modelo a través de los supuestos asumidos en los parámetros y en la verosimilitud, las distribuciones a priori, la estructura jerárquica y las variables incluidas.

Es importante resaltar que, a diferencia con los modelos de regresión estimados con métodos frecuentistas, en la estimación bayesiana se obtiene una distribución para los predichos $p(y_i^*|\boldsymbol{y})$ en lugar de un predicho puntual $\hat{y}$, que se denomina *distribución predictiva a posteriori*. En caso de ser necesario, se la puede resumir con estadísticos resumen como la media, mediana o moda para obtener una medida escalar de cada predicción.

En los modelos de regresión bayesianos, una manera informal de chequear el ajuste es mediante chequeos gráficos a posteriori, y una manera formal consiste en utilizar técnicas basadas en la *distribución predictiva a posteriori*.

Chequeos gráficos a posteriori

En el enfoque frecuentista, el gráfico de validación básico es el gráfico de *residuos versus predichos*. En el caso bayesiano este gráfico también es válido, pero para calcular el residuo como la diferencia entre las observación y la predicción es necesario resumir la distribución a posteriori del predicho con un estadístico resumen, que suele ser la moda. La interpretación del gráfico de residuos

versus predichos es análoga al enfoque frecuentista: los puntos deben distribuirse aleatoriamente alrededor del cero de los residuos, y la varianza debe ser constante a lo largo del eje de los predichos. En general, no deben observarse patrones sistemáticos en la nube de puntos.

Otros gráficos de validación secundarios pueden ser los *mapas de residuos*, para examinar visualmente la presencia de autocorrelación espacial, *histogramas de residuos*, para evaluar su normalidad, o *residuos versus covariables*, para encontrar indicios de relaciones no lineales.

Técnicas basadas en la distribución predictiva

En este caso se dividen los datos $\boldsymbol{y}$ en dos conjuntos: un conjunto para validar el modelo, y otro para entrenarlo. Hay dos maneras de dividir los datos. La primera consiste en usar el enfoque del *chequeo predictivo a posteriori*, donde cada observación se utiliza simultáneamente tanto para entrenamiento como para validación. La segunda consiste en usar técnicas de *validación cruzada*, donde cada observación puede pertenecer solo a uno de los dos grupos a la vez.

En el *chequeo predictivo a posteriori* (Gelman et *al.*, 1996) y_i^* es una *réplica*. Una réplica es una simulación u observación nueva obtenida con el mismo modelo que generó a los datos $\boldsymbol{y}$. En el chequeo predictivo a posteriori se calculan dos cantidades: la *distribución predictiva a posteriori* marginal $\pi(y_i^*|\boldsymbol{y})$ y el *p-value predictivo a posteriori* $p(y_i^* \leq y_i|\boldsymbol{y})$.

La distribución predictiva a posteriori de y_i^* cuando $\boldsymbol{y}$ y y_i^* son *intercambiables* (similares) se define como

$$\begin{aligned}
\pi(y_i^*|\boldsymbol{y}) &= \frac{\pi(y_i^*, \boldsymbol{y})}{\pi(\boldsymbol{y})} \\
&= \frac{1}{\pi(\boldsymbol{y})} \int \pi(y_i^*|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \frac{1}{\pi(\boldsymbol{y})} \int \pi(y_i^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{y})d\boldsymbol{\theta} \\
&= \int \pi(y_i^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}.
\end{aligned} \tag{57}$$

Cada $\pi(y_i^*|\boldsymbol{y})$ es la verosimilitud de obtener una observación y_i^* habiendo observado a $\boldsymbol{y}$. Si hay muchas $\pi(y_i^*|\boldsymbol{y})$ pequeñas, sera señal de un modelo no tan bueno, ya que una verosimilitud pequeña indica que las observaciones vienen de las colas de la distribución asumida y posiblemente

sean outliers (Blangiardo y Cameletti, 2015).

Los p-values predictivos a posteriori o *ppp* se obtienen como

$$ppp_i = p[T(y_i^*) \leq T(y_i|\boldsymbol{y})] \tag{58}$$

donde T es un estadístico de testeo tal como se lo define en la literatura frecuentista. Si y_i es un escalar continuo, un estadístico de testeo natural es $T(y_i) = y_i$, resultando

$$ppp_i = p(y_i^* \leq y_i|\boldsymbol{y}). \tag{59}$$

El *ppp* se interpreta directamente como la probabilidad de observar los datos y_i^* dado el modelo propuesto. Un valor razonable para los *ppp* caerá entre 0.05 y 0.95; si un *ppp* se acerca a cero o a uno, es indicio de que el patrón observado es poco probable de obtener con el modelo propuesto (Gelman et *al.* , 2014).

En la *validación cruzada* y_i^* representa a una observación i de la muestra, y $\boldsymbol{y}_{-i}$ es el vector que contiene todas las observaciones presentes menos la i. El modelo se entrena utilizando solo los datos en $\boldsymbol{y}_{-i}$ y por lo tanto la distribución a posteriori de los parámetros se obtiene condicional a $\boldsymbol{y}_{-i}$ y no a $\boldsymbol{y}$.

Una cantidad relacionada con los p-values predictivos a posteriori son las *transformadas integrales de probabilidad* o *PIT*, del inglés *probability integral transform* (Dawid, 1984). Los PIT se obtienen como los *ppp*, pero con validación cruzada para los mismos datos de la muestra. El PIT de cada observación de una distribución continua se define como

$$PIT_i = p(y_i^* \leq y_i|\boldsymbol{y}_{-i}) \tag{60}$$

donde $\boldsymbol{y}_{-i}$ es el vector que contiene todas las observaciones de la muestra menos la i. Los PIT en un modelo calibrado con datos continuos se deberían distribuir uniformemente entre 0 y 1, aunque la uniformidad de los PIT no valida automáticamente el modelo (Gelman et *al.*, 2014).

Una medida alternativa es la *ordenada predictiva condicional* o *CPO*, del ingles *Conditional*

Predictive Ordinate (Petit, 1990), definida para cada observación i como

$$CPO_i = \pi(y_i^*|\boldsymbol{y}_{-i}) \tag{61}$$

Los CPO_i se suelen usar para detectar outliers, ya que serán cercanos a cero para observaciones difíciles de obtener con el modelo propuesto.

2.6.2. Comparación de modelos

Es natural que aparezcan múltiples modelos candidatos para describir un conjunto de datos que difieren en complejidad y en precisión, y se haga necesario compararlos para encontrar el mejor. El desarrollo de los criterios de comparación basados en la evaluación de la precisión predictiva está íntimamente relacionado con la teoría de la información. A continuación se describe el razonamiento detrás del desarrollo de los criterios de información comúnmente utilizados en la comparación de modelos bayesianos. Para realizar esta descripción nos hemos apoyado principalmente en el Capítulo 7 del libro de Richard McElreath (2020).

Para calcular la precisión predictiva se busca una definición de *precisión* que haga que la precisión se maximice en el verdadero modelo, porque sería la mejor medida posible. La única definición de precisión que cumple con esta característica es la probabilidad conjunta de predecir correctamente todos los eventos, es decir, la *verosimilitud* (McElreath, 2020). El verdadero modelo tendrá máxima verosimilitud.

Generalmente se usa el logaritmo de la verosimilitud en lugar de la verosimilitud original $p(\boldsymbol{y}|\boldsymbol{\theta})$ para medir la precisión predictiva. En este contexto, a la log-verosimilitud se la renombra como *regla de score logarítmica* o *densidad predictiva logarítmica*, y se la abrevia como *lpd*, del inglés *log predictive density*.

$$lpd = log[p(\boldsymbol{y}|\boldsymbol{\theta})] \tag{62}$$

Como el logaritmo es una función monótona, el verdadero modelo p (que es en realidad una distribución de probabilidad) tendrá máxima log-verosimilitud y realizará predicciones perfectas.

Cualquier otro modelo q realizará predicciones no tan perfectas, y por lo tanto podemos medir la distancia entre la predicción perfecta de p y la predicción candidata de q.

Una distribución de probabilidad p contiene cierta cantidad de *información* o, equivalentemente, de *incertidumbre*. La información se define como la reducción en la incertidumbre que se gana al observar la realidad. En otras palabras, si se observa la ocurrencia de un evento poco frecuente, se obtiene mucha información - una gran reducción en la incertidumbre. Pero si se observa la ocurrencia de eventos muy comunes o frecuentes, no se obtiene demasiada información, porque era de esperar que ocurriesen.

La reducción en la incertidumbre se puede medir con las probabilidades p_i de los eventos i descriptas por la distribución p. La función más común para medir la reducción en la incertidumbre es la *entropía de la información*, que fue propuesta por Claude Shannon en 1940 en el contexto de la teoría de la información. La entropía de la información se describe como

$$H(p) = -E[log(p_i)] = -\sum_{i=1}^{n} p_i log(p_i) \tag{63}$$

donde n es la cantidad de eventos, cada uno con probabilidad p_i. De esta manera, la reducción en la incertidumbre contenida en una distribución de probabilidad p es su entropía, y equivale a la log-probabilidad promedio de un evento de esa distribución (McElreath, 2020).

Si se usa un modelo competidor q para hacer predicciones, será interesante saber qué tan lejos caen sus predicciones de las realizadas por el verdadero modelo p. Esta distancia puede calcularse en términos de la incertidumbre como una diferencia entre entropías: la del verdadero modelo p, y la entropía cruzada que surge de usar a q para predecir a p. Esta diferencia de entropías se llama *divergencia de Kullback-Leibler* y se define como

$$\begin{aligned} D_{KL}(p,q) &= H(p,q) - H(p) \\ &= \sum_i p_i(log(p_i) - log(q_i)) \\ &= \sum_i p_i log(\frac{p_i}{q_i}) \end{aligned} \tag{64}$$

La D_{KL} se interpreta como la incertidumbre adicional introducida al usar q para predecir p, y

no es más que la diferencia en log-probabilidad promedio entre p y q. Cuánto más distinto sea q de p, mayor será la divergencia. Por lo tanto, en la presencia de múltiples modelos candidatos se calculan las divergencias para cada modelo y el modelo más cercano al verdadero será aquel que minimice la divergencia (McElreath, 2020).

En la realidad, el verdadero modelo p nunca es conocido. Afortunadamente no es necesario conocer la *ubicación* absoluta de p para comparar modelos y alcanza con conocer qué *tan lejos* están los modelos candidatos q y r entre sí. Esto es así porque cuando se restan las divergencias entre cada modelo candidato y el verdadero p, se cancelan los términos que exigen conocer la distribución p:

$$\begin{aligned} D_{KL}(p,q) - D_{KL}(p,r) &= \sum_i p_i(log(p_i) - log(q_i)) - \sum_i p_i(log(p_i) - log(r_i)) \\ &= \sum_i p_i(log(r_i)) - \sum_i p_i(log(q_i)) \\ &= E[log(r_i)] - E[log(q_i)] \\ &= H(p,q) - H(p,r) \end{aligned} \tag{65}$$

Los términos de log-probabilidad promedio de los modelos candidatos, $E[log(q_i)]$ para q y $E[log(r_i)]$ para r, se pueden obtener con el estimador por método de los momentos $\frac{1}{n}\sum_i log(q_i)$.

Multiplicando el estimador anterior por n se obtiene una puntuación total $S(q)$ para un modelo q:

$$S(q) = \frac{1}{n}\sum_i log(q_i) \tag{66}$$

Esta puntuación o *score* total no es más que la regla de score logarítmica o *lpd* (ver Ecuación 62), que es la mejor medida de la precisión predictiva que se puede tener:

$$\begin{aligned} lpd &= log[p(\boldsymbol{y}|\boldsymbol{\theta_q})] \\ &= \sum_{i=1}^{n} log[p(y_i|\boldsymbol{\theta_q})] \\ &= S(q), \end{aligned} \tag{67}$$

donde se asumen observaciones independientes.

La versión bayesiana de la *lpd* es la *densidad predictiva logarítmica puntual* o *lppd* (del inglés *log pointwise predictive density*). Dados los datos $\boldsymbol{y}$ y una distribución a posteriori representada por el vector de parámetros $\boldsymbol{\Theta}$,

$$lppd(\boldsymbol{y}, \boldsymbol{\Theta}) = \sum_i log \frac{1}{S} \sum_s p(y_i|\Theta_s) \tag{68}$$

donde S es la cantidad de muestras y $\boldsymbol{\Theta_S}$ son los valores de los parámetros en la muestra s. Básicamente, para cada observación se calcula su probabilidad *promedio* (promediada para todos los posibles valores de los parámetros en $\boldsymbol{\Theta_S}$) antes de calcularle el logaritmo y sumar las probabilidades logarítmicas resultantes.

El problema con la *lpd* y la *lppd* es que siempre mejorarán al aumentar la complejidad del modelos si se las calcula con los datos de entrenamiento, por lo que lo ideal es calcularlas con datos nuevos, "fuera de la muestra". Como en muchos casos no es posible conseguir datos fuera de la muestra, se han desarrollado estrategias para estimar la precisión predictiva con los datos de entrenamiento. Dos estrategias muy conocidas son los métodos de **validación cruzada** y los **criterios de información**.

Los criterios de información son medidas que aproximan la precisión predictiva utilizando la *lppd* y corrigiendo el sobreajuste resultante de utilizar los datos de entrenamiento mediante una penalización en función de la complejidad del modelo. Los métodos de validación cruzada corrigen el sobreajuste dividiendo la muestra en submuestras de entrenamiento y validación.

Criterios de información

En la estadística es usual encontrar una modificación de la *lppd* llamada *deviance*. La deviance no es más que la *lppd* multiplicada por -2 por razones históricas. Sea entonces la deviance $D(\boldsymbol{\theta})$ calculada con los datos de entrenamiento tal que

$$D(\boldsymbol{\theta}) = -2lppd$$

La idea detrás de los criterios de información es que se puede aproximar la precisión predictiva fuera de la muestra con la precisión predictiva calculada con los datos de entrenamiento, penalizando esta última con un termino de corrección por sobreajuste.

El primer criterio de información fue el *Criterio de información de Akaike*, propuesto por Akaike en 1973. Akaike entendió que en las regresiones lineales ordinarias, la diferencia en la deviance calculada con datos de entrenamiento y calculada con datos de testeo era aproximadamente siempre el doble de la cantidad de parámetros p (McElreath, 2020) y propuso

$$AIC = D(\boldsymbol{\theta}) + 2p \tag{69}$$

Por lo tanto, en las regresiones lineales ordinarias, la cantidad de parámetros es una medida natural de la tendencia al sobreajuste del modelo. Sin embargo, el AIC como aproximación a la precisión predictiva fuera de la muestra solo es confiable cuando las distribuciones a priori no dominan la verosimilitud, cuando la distribución a posteriori es aproximadamente normal multivariada, y cuando $n >> p$ (McElreath, 2020).

El *criterio de información de deviance* o *DIC* (Spiegelhalter et *al.*, 2002) (del inglés *Deviance Information Criterion*) es una generalización del AIC que permite evaluar la precisión fuera de la muestra cuando se usan distribuciones a priori informativas, pero mantiene la exigencia respecto a la posteriori normal y $n >> p$ (McElreath, 2020).

El DIC se calcula sumándole el numero efectivo de parámetros p_D a la deviance media a posteriori $\bar{D}$:

$$DIC = \bar{D} + p_D, \tag{70}$$

donde $\bar{D} = E_{\boldsymbol{\theta}|\boldsymbol{y}} D(\boldsymbol{\theta})$ y el numero efectivo de parámetros se calcula como

$$\begin{aligned} p_D &= E_{\boldsymbol{\theta}|\boldsymbol{y}} D(\boldsymbol{\theta}) - D(E_{\boldsymbol{\theta}|\boldsymbol{y}} \boldsymbol{\theta}) \\ &= \bar{D} - D(\bar{\boldsymbol{\theta}}) \end{aligned} \tag{71}$$

En el contexto bayesiano resulta necesario utilizar la deviance media a posteriori, porque la

deviance a posteriori es una distribución de probabilidad y no un estimador puntual.

El *criterio de información ampliamente aplicable* o *WAIC* (Widely Applicable Information Criterion), propuesto en 2013 por S. Watanabe, es un criterio de información que funciona con distribuciones a priori informativas y con distribuciones a posteriori no normales, a diferencia del AIC y DIC.

$$WAIC(\boldsymbol{y}, \boldsymbol{\theta}) = -2(lppd - \sum_i var_\theta \; log[p(y_i|\theta)]) \tag{72}$$

donde $\boldsymbol{y}$ son las observaciones y $\boldsymbol{\theta}$ son los parámetros que definen la distribución a posteriori. El segundo término dentro del paréntesis es el término de penalización por sobreajuste, que no es más que la suma sobre todas las observaciones de la varianza de la probabilidad logarítmica. A menor varianza, menor penalización por sobreajuste.

El término de penalización en WAIC a veces se denomina *número efectivo de parámetros*, p_{WAIC}, por analogía con el AIC (McElreath, 2020), aunque en modelos distintos a las regresiones lineales normales no sea el número de parámetros lo que aumente la tendencia del modelo al sobreajuste.

Otras dos medidas de capacidad predictiva que son útiles para los modelos de regresión con respuesta normal son la *raíz del error cuadrático medio* (RMSE, del inglés *Root Mean Square Error* y el *error medio absoluto* (MAE, del inglés *Mean Absolute Error*). Las fórmulas del $RMSE$ y del MAE son

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{73}$$

y

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}, \tag{74}$$

donde y_i representa la observación y $\hat{y}_i$ representa el predicho puntual para esa observación. En este trabajo calculamos el predicho puntual como la moda a posteriori de la distribución del predicho.

Tanto $RMSE$ como MAE miden el error de predicción en unidades de la cantidad que se desea estimar, pero por el cuadrado, $RMSE$ es más sensitiva a "outliers". Por otro lado, MAE es más sencilla de comprender porque representa la magnitud promedio de los residuos.

Métodos de validación cruzada

La validación cruzada intenta proveer una medida de precisión predictiva evaluada fuera de la muestra utilizando solo datos de entrenamiento. Para ello se dividen los datos de entrenamiento en k particiones de igual tamaño, se vuelven a unir $k-1$ de estas particiones para entrenar el modelo, y se calcula la medida de precisión predictiva sobre la partición k restante. El procedimiento se repite k veces, evaluando al modelo cada vez en una partición distinta. La estimación de la precisión predictiva fuera de la muestra será el promedio de las k medidas calculadas. También es posible calcular la precisión predictiva dentro de la muestra, evaluándola sobre las $k-1$ particiones que se usan en cada iteración para entrenar al modelo y promediando las medidas obtenidas.

El caso especial de validación cruzada más conocido es la validación cruzada *leave-one-out*, donde $k = n$. Los CPO_i definidos en la Sección 2.6.1 son verosimilitudes marginales de cada observación obtenidas por validación cruzada *leave-one-out*. El producto de los CPO_i se puede interpretar como una *pseudo-verosimilitud marginal* y por lo tanto sirve como medida resumen de ajuste del modelo calculada mediante validación cruzada (Carlin y Louis, 2009). Por motivos computacionales se suele tomar el logaritmo esta medida, propuesta originalmente por Geisser y Eddy (1979) y bautizada *pseudo-verosimilitud marginal logarítmica* o LPML (del inglés *log pseudo marginal likelihood*). La LPML termina siendo una estimación de la *lppd* (ver Ecuación 68) que evita mediante validación cruzada el sesgo por sobreajuste. Una mayor LPML indica mejor ajuste.

$$LPML = log\,\{\prod_i p(y_i^*|\boldsymbol{y}_{-i})\} = \sum_i log\,CPO_i \tag{75}$$

En el caso de modelos para datos espaciales, existe el inconveniente de que las observaciones no son independientes, y los métodos de validación cruzada requieren particiones independientes de los datos. Se han propuesto varios enfoques para realizar validación cruzada con datos espaciales

que tienen en cuenta la falta de independencia inducida por la cercanía. Por ejemplo, la validación cruzada espacial *leave-one-out* o *SLOOCV*, del inglés *Spatial Leave-one-out Cross-Validation* propone eliminar de la partición de entrenamiento no solo la observación i sino también todas las vecinas en el rango de la autocorrelación espacial. De esta manera asegura la independencia entre la partición de entrenamiento y la de testeo.

Otra estrategia para realizar validación cruzada espacial es la denominada *bloqueo espacial*, que consiste en dividir el territorio en análisis en m bloques espaciales de igual tamaño, que a su vez se asignan a k particiones de los datos, con $m \leq k$ (Valavi, Elith, Lahoz-Monfort y Guillera-Arroita, 2018). El paquete 'blockCV' de 'R' (Valavi et al., 2018) permite implementar fácilmente SLOOCV y bloqueo espacial, entre otras opciones. Este paquete permite un alto grado de customización en la división del territorio. Por ejemplo, en el caso del bloqueo espacial es posible elegir el tamaño de los bloques espaciales, la cantidad de particiones y la forma de asignar los bloques a las particiones, entre otras opciones.

2.7. Técnicas de reducción de datos

La *reducción de datos* es la acción de reducir el espacio de covariables para disminuir el número de parámetros a estimar, sin distorsionar la inferencia estadística para los parámetros. En algunos casos, descartar covariables resulta útil porque se logran modelos más parsimoniosos. En otros casos, resulta imprescindible para eliminar covariables con mala calidad predictiva.

Harrell (2015) propone cuatro métodos de reducción de datos en el contexto de la estimación frecuentista:

1. Usar la literatura para eliminar variables no importantes.
2. Eliminar variables con distribuciones muy angostas.
3. Eliminar covariables que no estén presentes en un gran número de sujetos.
4. Usar algún *método estadístico de reducción de datos.*

Un motivo para querer aplicar métodos estadísticos de reducción de datos es la presencia de *multicolinealidad.* La multicolinealidad, que es una característica de los datos y no del modelo, aparece cuando al menos una covariable puede ser predicha muy bien a partir de otras.

En la estimación frecuentista del modelo lineal, la multicolinealidad genera una matriz mal condicionada en el proceso de estimación que causa inestabilidad y mayor varianza en los coeficientes de regresión estimados (Belsley et al., 1980). En otras palabras, cuando hay multicolinealidad, la inferencia clásica, que no considera información a priori, falla como método para interpretar la evidencia de los datos (Leamer, 1973).

El efecto de la multicolinealidad en modelos de regresión para datos espaciales también ha sido en cierta medida estudiado con anterioridad. Por ejemplo, Beale et al. (2010), en un estudio con datos simulados, compararon dos modelos estimados con GLS que incluyeron la estructura espacial en el término de error, y encontraron evidencia de que las variables espaciales altamente correlacionadas entre sí causaron estimaciones imprecisas de los coeficientes de regresión.

Dentro del contexto de reducción de datos, la multicolinealidad se puede eliminar (1) descartando manualmente las covariables colineales o (2) aplicando técnicas estadísticas que resuman la

información repetida en las covariables colinealies de manera que las columnas de la matriz de predictoras tengan un grado de ortogonalidad aceptable (Harrell, 2015). En ambos casos, la matriz de covariables se manipula en un contexto no supervisado, es decir, sin utilizar la información en la respuesta.

La técnica de reducción de datos más famosa es sin dudas el *análisis de componentes principales* (PCA, del inglés *Principal Component Analysis*), formalizada por Hotelling en 1933. PCA realiza una transformación lineal de una matriz de variables que genera variables sintéticas ubicadas en las direcciones de mayor variabilidad de los datos originales. Como estas direcciones son ortogonales entre sí, se genera una nueva matriz de variables con columnas ortogonales, eliminando por completo la multicolinealidad.

PCA no es capaz de incorporar la información de autocorrelación espacial y por ello se han sugerido modificaciones al procedimiento para tratar con datos espaciales. *MULTISPATI* (Dray et al., 2008) es el nombre dado a uno de estos algoritmos, que incorpora la autocorrelación espacial medida con el índice de Moran y puede verse como PCA adaptado para datos espaciales.

Los algoritmos de clústering jerárquico no reducen la multicolinealidad pero sí pueden determinar las variables que forman grupos multicolineales, permitiendo la posterior aplicación de PCA y MULTISPATI a los grupos multicolineales en lugar de a la matriz original de predictoras.

A continuación se describen con mayor detalle los algoritmos detrás de PCA, MULTISPATI y clústering jerárquico de variales.

2.7.1. Análisis de componentes principales (PCA)

El *Análisis de componentes principales* o *PCA*, del inglés *Principal Component Analysis* es una técnica no supervisada de reducción de datos basada en álgebra de matrices y formulada por Hotelling en 1933.

Calcular las componentes principales de un conjunto de p variables sirve para encontrar una base del espacio de las p columnas de la matriz de correlación o covarianza. Como estas son matrices simétricas y definidas positivas, la descomposición espectral genera eigenvalores reales positivos y eigenvectores ortogonales. Los eigenvectores representan las direcciones de variabilidad de los datos y los eigenvalores representan la varianza en cada dirección. Por lo tanto, si se los

ordena por orden decreciente de los eigenvalores, el primer eigenvector indica la dirección de máxima variabilidad, el segundo eigenvector la segunda dirección de máxima variabilidad, y así sucesivamente hasta el p eigenvector.

Cada dirección de variabilidad es llamada *componente principal*, y por ser ortogonales los eigenvectores, las componentes no están correlacionadas y la información que proporciona una componente no está en parcialmente contenida en ninguna otra componente.

Matemáticamente, PCA resuelve la ecuación característica de la matriz de covarianzas $\boldsymbol{\Sigma}$ o de correlación $\boldsymbol{R}$

$$det(\boldsymbol{\Sigma} - \lambda \boldsymbol{I}) = 0 \tag{76}$$

utilizando su descomposición en valores singulares o descomposición espectral para encontrar sus eigenvalores y eigenvectores.

El eigenvector $\boldsymbol{u}_i$ asociado a una componente principal i estará formado por p elementos, donde cada elemento representará la correlación lineal entre una variable y la componente principal i. La varianza de la componente principal será λ_i, y la proporción de varianza explicada por esta componente principal será $\lambda_i / \sum_{j=1}^{p} lambda_j$. El cuadrado de las correlación de una variable con una componente representará el porcentaje de la variabilidad de la componente explicada por esa variable. Uno de los objetivos principales de encontrar las componentes principales de un conjunto de datos suele ser reducir sus dimensiones, explicándolos solo con sus direcciones de mayor variabilidad. Dependiendo del área de aplicación, se considera necesario explicar un porcentaje mínimo de la variabilidad, donde 60 % suele ser un valor usual. Para ello se retienen componentes principales hasta que la variabilidad acumulada explicada por las componentes retenidas sea superior al porcentaje mínimo requerido.

2.7.2. MULTISPATI

Si bien PCA podría aplicarse para resumir datos georreferenciados, no incorporaría directamente la información proporcionada por la estructura de dependencia espacial. MULTISPATI es un método de análisis multivariado espacial basado en el I de Moran que introduce una restricción espacial a los métodos multivariados clásicos. MULTISPATI fue propuesto por Dray, Saïd

y François en 2008 y puede verse como una generalización de un método multivariado anterior, propuesto por Wartenberg en 1985. El objetivo principal de MULTISPATI es encontrar componentes principales que optimicen el producto de la varianza total y el indice de Moran como medida de autocorrelación espacial. La ventaja de MULTISPATI respecto a PCA es que las componentes principales espaciales de MULTISPATI maximizan la autocorrelación espacial entre los sitios.

El desarrollo matemático de MULTISPATI es como sigue. Sea $\boldsymbol{x}$ un vector compuesto por n mediciones en n ubicaciones espaciales y sea $\boldsymbol{W}$ la matriz de pesos espaciales estandarizada por filas definida en la Sección 2.1 tal que se pueda usar para calcular el I de Moran definido en la misma sección. Cuando se usa esta $\boldsymbol{W}$, el I de Moran se puede ver como el producto de un escalar de suavizados espacial por la correlación de Pearson entre una variable y su *vector de lag espacial* $\tilde{\boldsymbol{x}} = \boldsymbol{W}\boldsymbol{x}$ (Dray et al., 2008).

Para plantear MULTISPATI, sus autores se basan en el marco general del análisis multivariado que dice que aplicar un método multivariado como PCA o MULTISPATI a una matriz de covariables $\boldsymbol{X}$ equivale a analizar el triplete estadístico $(\boldsymbol{X}, \boldsymbol{Q}, \boldsymbol{D})$, donde $\boldsymbol{Q}_{pxp}$ y $\boldsymbol{D}_{nxn}$. El análisis del triplete estadístico consiste en encontrar las direcciones $\boldsymbol{u}_i$ que maximicen la forma cuadrática $\boldsymbol{u}_i^T\boldsymbol{Q}\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X}\boldsymbol{Q}\boldsymbol{u}_i$. En la práctica, los vectores $\boldsymbol{u}_j$ son los eigenvectores derechos de $\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X}\boldsymbol{Q}$, donde $1 \leq j \leq r$ y r es el rango de la $\boldsymbol{X}$. (Dray et al., 2008).

MULTISPATI equivale a diagonalizar el triplete estadístico $(\boldsymbol{X}, \boldsymbol{Q}, \boldsymbol{E})$, donde

$$\boldsymbol{E} = \frac{1}{2}(\boldsymbol{W}^T\boldsymbol{D} + \boldsymbol{D}\boldsymbol{W}). \tag{77}$$

En otras palabras, MULTISPATI equivale a maximizar la cantidad $\boldsymbol{u}_i^T\boldsymbol{Q}\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{X}\boldsymbol{Q}\boldsymbol{u}_i$ bajo la restricción $\boldsymbol{u}_1^T\boldsymbol{Q}\boldsymbol{u}_1 = 1$, donde $\tilde{\boldsymbol{X}} = \boldsymbol{W}\boldsymbol{X}$ es la matriz de lag espacial. La cantidad a maximizar resulta ser el producto escalar entre la combinación lineal $\boldsymbol{X}\boldsymbol{Q}\boldsymbol{u}_1$ de las variables originales y una combinación lineal de variables lagueadas $\boldsymbol{W}\boldsymbol{X}\boldsymbol{Q}\boldsymbol{u}_1$.

2.7.3. Clustering jerárquico de variables

Los algoritmos de clustering intentan segmentar o agrupar una colección de objetos en subconjuntos denominados *clusters* tal que los objetos dentro de un cluster sean más similares entre sí

que los objetos pertenecientes a distintos clústers (Hastie, Tibshirani y Friedman, 2009). Todas las técnicas de clustering se basan en el conocimiento de la similaridad entre los k objetos que se quieren agrupar. Para representar las similaridades se suele utilizar una matriz $\boldsymbol{D}_{k\times k}$, donde cada elemento d_{ij} guarda la proximidad entre los objetos i y j (Hastie et al., 2009).

Hay muchas formas de definir la matriz $\boldsymbol{D}$ y también muchos tipos de clustering, entre ellos los algoritmos de *clustering jerárquico*. En los algoritmos de clustering jerárquico la medida de similaridad se calcula entre grupos disjuntos de objetos y se basa en las similaridades de a pares entre los objetos de cada grupo. Estos algoritmos parten de tantos clústers como objetos haya, y producen representaciones jerárquicas donde los clústers en cada nivel de la jerarquía se crean uniendo los clústers del nivel inmediatamente inferior (Hastie et al., 2009). Más detalles de los algoritmos de clustering jerárquico se pueden revisar en el Capítulo 14 de Hastie et al. (2009).

Si bien el clustering se suele aplicar para agrupar puntos del espacio de las observaciones, se pueden aplicar los mismos algoritmos al espacio de las covariables. Las técnicas de clustering estadístico de variables se pueden así utilizar para identificar dimensiones independientes dentro de un conjunto de variables. Por ejemplo, una forma de realizar clustering de variables es aplicando un algoritmo de clustering jerárquico a una matriz de similaridad apropiada, por ejemplo, calculada con correlaciones cuadradas de Pearson o de Spearman (Harrell, 2015).

3. Metodología de investigación

Este capítulo está organizado como sigue. En la Sección 3.1 se describen los datos originales que utilizaremos para ilustrar la metodología, junto con el pre-procesamiento aplicado a los mismos previo a la etapa de modelado. En la Sección 3.2 se describen: en primer lugar, los pasos seguidos para llevar a cabo el análisis de sensibilidad a las distribuciones a priori de los hiperparámetros del efecto aleatorio; en segundo lugar, los pasos seguidos para analizar el impacto de la multicolinealidad en el ajuste de los modelos; y en tercer y último lugar, los criterios de selección del mejor modelo predictivo.

3.1. Datos

En el año 2020 inició una operación conjunta entre la Facultad de Ciencias Agropecuarias de la UNC (FCA-UNC), la UFYMA (CONICET) y el Instituto de Investigación de la Comisión de Enlace de Entidades Agropecuarias (CEEA) Regional Córdoba, con fines de investigar interdisciplinariamente y utilizando datos geoposicionados la región de la Cuenca del Río Carcarañá. El territorio en análisis se ubica geográficamente en las Provincias de Córdoba y de Santa Fe de la República Argentina y abarca aproximadamente 48000 km^2 (Figura 3a).

El objetivo principal de esta colaboración es construir conocimiento para la gestión ambiental y productiva integral del territorio de dicha cuenca, a través modelos estadísticos que permitan identificar asociaciones con potencial para predecir riesgos emergentes en relación con la degradación de recursos naturales y la producción agropecuaria

Las actividades realizadas bajo esta colaboración hasta el día de la fecha incluyeron la sistematización de datos espaciales georreferenciados del territorio de la cuenca, su posterior depuración y re-escalamiento, la construcción de un Sistema de Información Geográfica (SIG) y el modelado estadístico de los datos del SIG mediante modelos de aprendizaje automático e interpolación *kriging*.

A continuación se describen las características de los datos originales disponibles en el SIG y el preprocesamiento adicional al que se sometieron los datos en este trabajo.

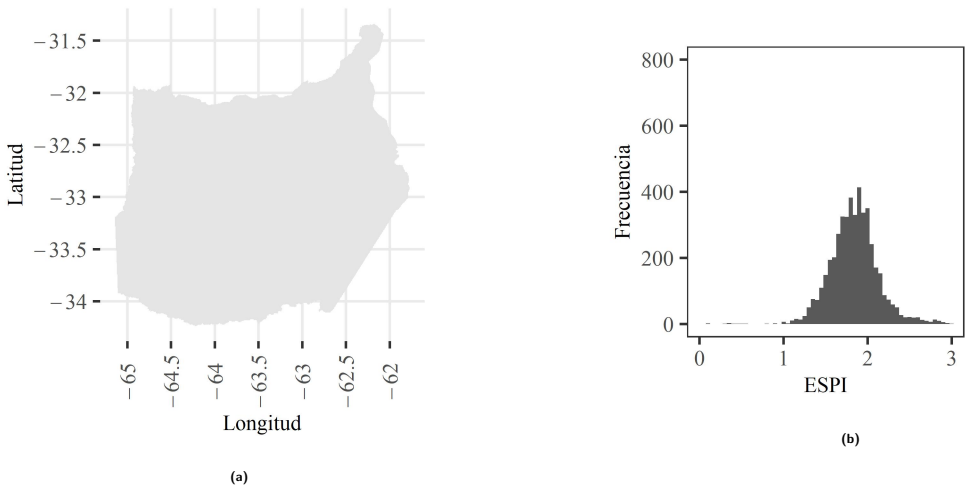

Figura 3: (a) Territorio de la Cuenca del Carcarañá. (b) Histograma del ESPI.

3.1.1. Datos originales

El SIG alberga variables edáficas, climáticas, topográficas, de vegetación y de cobertura de suelos, georreferenciadas, para el territorio de la Cuenca del Río Carcarañá. En el SIG también se dispone de índices que describen servicios ecosistémicos, tales como el índice de fertilidad del suelo, un índice de provisionamiento de servicios ecosistémicos, el almacenamiento de carbono en suelo y en vegetación, y la dinámica de la productividad terrestre.

En esta tesis se utilizó como respuesta la tendencia a largo plazo del *índice de provisionamiento de servicios ecosistémicos* (ESPI, del inglés *Ecosystem Services Provisioning Index*), que es un indicador construido siguiendo la metodología de Teich et al. (CITA). La tendencia a largo plazo del ESPI está basada en la dinámica de series temporales largas de las diferencias normalizadas del índice de vegetación (NDVI, del inglés *Normalized-Difference Vegetation Index*) y es una variable continua, positiva, que fluctúa aproximadamente entre 0 y 3, donde mayores valores indican mejor provisionamiento de servicios ecosistémicos. La Figura 3b muestra un histograma del ESPI.

Las variables del SIG están disponibles para realizar modelado geoestadístico o regional sobre el territorio de la cuenca. En el caso regional, las regiones fueron obtenidas mediante un algoritmo de clustering espacial que se alimentó con las variables NDVI, elevación, materia orgánica, arcilla, temperatura y precipitación. El resultado del clustering fue la división del territorio en 4676

regiones de aproximadamente 1420Has de superficie promedio, que se supusieron homogéneas en cuanto a clima, suelo, topografía y vegetación. El razonamiento detrás de la partición del territorio en clústeres fue que las unidades resultantes pueden resultar mas convenientes para el modelado que la división del territorio mediante una grilla regular. En esta tesis todas las variables se trabajaron medidas a escala de estas unidades regionales.

3.1.2. Datos preprocesados

A partir de aquí, se utilizará el término *covariables* para referirse a las variables del SIG disponibles a nivel de región que se utilizaron como predictoras en los modelos.

Pre-selección de variables

De las covariables disponibles en el SIG, un gran porcentaje eran covariables formadas por distintos estadísticos resumen de la misma cantidad. Tomando como ejemplo la variable **arcilla**, se contaba con la media, el desvío estándar, algunos cuantiles y la moda del porcentaje de arcilla medido en cada región. En esos casos solo se retuvo la media y se descartaron el resto de los estadísticos resumen.

Se encontaron unas 20 covariables con valores faltantes, donde la covariable con mayor cantidad de observaciones faltantes tuvo como máximo 406 observaciones faltantes sobre un total de 4676. Se decidió no eliminar covariables en base a la presencia de observaciones faltantes porque 406 observaciones representan menos del 10 % de las observaciones y por lo tanto se prefirió tener valores faltantes a descartar la covariable completa. Como R-INLA reemplaza por defecto los valores faltantes en las covariables por cero, si las covariables entran estandarizadas al modelo, los valores faltantes terminan siendo imputados con la media de la variable.

Se eliminaron las covariables con distribuciones demasiado angostas. En este grupo se incluyeron las covariables que tenían valor cero en más del 95 % de las regiones o que no presentaban suficiente variabilidad.

Se detectaron las covariables perfectamente colineales o *aliased* entre sí mediante una regresión lineal múltiple con el paquete `LM` de R, y se eliminaron hasta dejar solo una de las covariables.

También se eliminaron las casi perfectamente colineales, con coeficiente de correlación de Pearson mayor a 0.99, hasta dejar una sola del grupo colineal. Los grupos colineales en este caso se detectaron con el coeficiente de correlación de Pearson.

Como la respuesta ESPI es un índice calculado a partir del NDVI, se descartó la covariable NDVI y otras covariables también construidas a partir de imágenes del NDVI. Además, se descartaron todas las variables que podrían considerarse variables respuesta, ya que en su mayoría son índices sintéticos construidos a partir de otras covariables, y por lo tanto podrían contener información repetida con las covariables que se incluyeron como regresoras. El número de covariables al final de la etapa de preselección de variables fue de 36. La Tabla 9 en el Anexo A muestra la lista de covariables preseleccionadas y su distribución en términos de estadísticos resumen. Antes de comenzar la etapa de modelado se estandarizaron todas las covariables restándoles su media y dividiéndolas por su desvío estándar.

3.2. Procedimientos

3.2.1. Análisis de sensibilidad a las distribuciones a priori

El *análisis de sensibilidad* es el procedimiento mediante el cual el investigador intenta comprender cuánto cambia la inferencia a posteriori de un modelo bayesiano ante perturbaciones en el modelo (Gelman et al., 2014), tales como cambios en la verosimilitud, en las distribuciones a priori y en las covariables incluidas, entre otros aspectos. En el caso de los modelos jerárquicos bayesianos, interesa particularmente la posible influencia de las distribuciones a priori de los hiperparámetros en la tendencia al sobreajuste del modelo (ver Sección 2.5).

En esta tesis realizamos un análisis de sensibilidad para las distribuciones a priori de los hiperparámetros de los efectos aleatorios en los modelos *iid*, $iCAR$ y BYM2, utilizando en todos los casos distribuciones a priori penalizadas por complejidad. Para hacer el análisis de sensibilidad nos basamos en los trabajos de Beguin et al. (2012) e Illian et al. (2010), pero a diferencia de estos autores utilizamos las distribuciones a priori penalizadas por complejidad de Simpson et al. (2017).

El objetivo final del análisis de sensibilidad fue elegir una única distribución a priori para cada modelo que permita suficiente variabilidad en el efecto aleatorio pero que no genere un modelo

cuya flexibilidad sea solo consecuencia de la distribución a priori y no se vea respaldada por los datos.

La distribución a priori final de cada modelo se eligió en dos etapas:

1. *Pre-selección de las distribuciones a priori.* Se ajustaron modelos nulos (solo intercepto) para cada posible distribución a priori, y se analizo el comportamiento del WAIC, del cociente datos-parámetros y de la escala del efecto aleatorio en función del grado de información de la distribución a priori. Se seleccionaron dos o tres distribuciones a priori candidatas por cada tipo de modelo.

2. *Selección de la distribución a priori final.* Se ajustaron modelos incluyendo covariables para las dos o tres distribuciones a priori candidatas, y se eligió la distribución a priori que mejores resultados obtuvo en las métricas de capacidad predictiva $RMSE$ y MAE calculadas con validación cruzada espacial.

A continuación se describen los modelos ajustados, los criterios de comparación, y la elección las distribuciones a priori.

Modelos ajustados

En la pre-selección de distribuciones a priori se ajustaron *modelos nulos* que modelaron la respuesta solo en función del intercepto, imitando el procedimiento descripto por Illian et al. (2010). La Tabla 1 describe los modelos ajustados. En todos los casos la verosimilitud se asumió normal y se utilizaron las distribuciones a priori por defecto de `R-INLA` para los coeficientes de regresión y la varianza de las observaciones.

Todas las distribuciones a priori se construyeron con el marco de distribuciones a priori penalizadas por complejidad y su nivel de información fue manejado por U y α (ver Sección 2.5.2).

El BYM2 tuvo dos hiperparámetros, la precisión y el parámetro de mezcla. Suponiendo 50 distribuciones para el parámetro de precisión y 50 para el parámetro de mezcla, la cantidad total de combinaciones a probar en el BYM2 hubiese sido de $50 \times 50 = 2500$. Como el costo computacional de ajustar 2500 modelos se consideró demasiado alto, se permitieron 50 configuraciones

Tabla 1: Distribuciones a priori probadas en el análisis de sensibilidad.

Tipo de Modelo	U_τ	α_τ	U_ϕ	α_ϕ	Cantidad de Modelos
iid	0.001, 0.005, 0.01, 0.05, 0.1	0.05, 0.1, 0.25, 0.5, 0.65, 0.75, 0.95			35
$iCAR$	0.0001, 0.0005, 0.001, 0.005, 0.01	0.05, 0.15, ..., 0.95			50
BYM2	0.0001, 0.0005, 0.001, 0.005, 0.01	0.05, 0.15, ..., 0.95	0.5	0.05, 0.5, 0.75, 0.95	200

U_τ y α_τ controlan la distribución a priori de la precisión τ; U_ϕ y α_ϕ controlan la distribución a priori del parámetro de mezcla ϕ del modelo BYM2.

a priori para la precisión pero solo cuatro configuraciones a priori para el parámetro de mezcla ϕ, aquellas resultantes de combinar los valores de $\alpha \in (0.05, 0.5, 0.75, 0.95)$ con $U = 0.5$. Estas cuatro distribuciones permitieron expresar la confianza a priori que se tuvo en que el porcentaje de variabilidad espacial fuese inferior al 50 %. En total, se ajustó el modelo BYM2 con $50 \times 4 = 200$ posibles distribuciones a priori.

Para elegir la distribución a priori final de cada modelo ajustamos (1) el MLR, (2) el modelo *iid* con sus distribuciones preseleccionadas, (3) el modelo $iCAR$ con sus distribuciones preseleccionadas, y (4) el modelo BYM2 con sus distribuciones preseleccionadas. En este caso no se especificaron modelos nulos, como en el análisis de sensibilidad, sino que se modeló la respuesta $ESPI$ en función de las covariables del Conjunto 1 (las covariables originales) y términos cuadráticos para las covariables Elevación, Recurrencia y VUT, que mostraron relaciones polinómicas con los residuos del MLR.

Criterios de comparación de resultados

La preselección de distribuciones a priori se hizo evaluando manualmente el comportamiento del WAIC, del cociente datos-parámetros y de la escala del efecto aleatorio a medida que aumentó el nivel de información de las distribuciones a priori penalizadas por complejidad. El cociente datos-parámetros se calculó como n/p_{WAIC}, donde p_{WAIC} fue el número de parámetros efectivos calculados con la fórmula del WAIC. Este procedimiento se basó en los trabajos de Beguin et al.(2012) y de Illian et al. (2010), pero a diferencia de Beguin et al. (2012), que utilizaron el criterio de información DIC, elegimos utilizar el WAIC porque se consideró más adecuado para modelos jerárquicos bayesianos con distribuciones a priori informativas que el DIC (ver Sección

2.6.2).

Para comprender el comportamiento del WAIC y de n/p_{WAIC} con el nivel de información de las distribuciones se graficaron sus curvas en función de α, para cada nivel de U. Se eligió el eje de las abscisas para α porque de esta forma la visualización resultó más clara, ya que α está acotada entre 0 y 1, y U no lo está. Si se desea evaluar el efecto de distintos valores de U para un mismo nivel de confianza α, se puede trazar una línea vertical imaginaria en el nivel de α deseado y examinar los U que caen en la intersección.

Para comprender el comportamiento de la escala del efecto aleatorio se graficaron mapas de la media a posteriori de los efectos aleatorios en cada región y se comparó la escala observada del efecto aleatorio con la escala de las covariables. Para la comparación se eligieron las covariables Pendiente, Elevación, Recurrencia, Actividades Estivales, P (fósforo) y Nt (nitrógeno), que fueron las que se habían considerado con potencial predictivo en trabajos previos para los mismos datos de ilustración.

Finalmente, a partir del análisis manual del comportamiento del WAIC, del n/p_{WAIC} y de la escala del efecto aleatorio, se eligieron dos o tres distribuciones candidatas para cada tipo de modelo.

Para reducir las candidatas a una única distribución por cada tipo de modelo se ajustaron modelos con la formula $ESPI \sim$ *covariables*, y se analizó la tendencia al sobreajuste de cada modelo observando el $RMSE$ y MAE calculados con validación cruzada espacial (Figura 4) para los datos de entrenamiento y testeo. Este procedimiento fue una propuesta original de esta tesis y fue motivado por el deseo de contar con una medida distinta al WAIC para evaluar la capacidad predictiva del modelo, ya que sospechamos que el WAIC puede favorecer a los modelos que más sobreajustan.

Para la validación cruzada espacial se dividió el territorio en 25 bloques espaciales. Cada bloque contuvo en promedio 187 unidades regionales. Los bloques espaciales se acomodaron en 10 particiones, donde cada partición estuvo formada por dos o mas bloques espaciales. La Figura 4 muestra la división del territorio en bloques espaciales, donde los bloques que comparten el color se han asignado a la misma partición.

Figura 4: División del territorio en los bloques espaciales que forman las diez particiones.

Selección de distribuciones a priori

En la preselección de las distribuciones se intentaron seleccionar dos o tres distribuciones por cada modelo que penalizaran lo suficiente al modelo como para evitar que el efecto aleatorio capturase toda la variabilidad, pero no tanto como para eliminarlo por completo.

El gráfico de WAIC versus α para distintos valores de U tuvo fines orientativos y se utilizó para identificar posibles valores de U y α para los que el WAIC descendió abruptamente. Como WAICs menores indican mejor ajuste, se tomó un descenso abrupto en el WAIC como una alerta de sobreajuste.

En el gráfico n/p_{WAIC} versus α se buscó comprender para qué valores de U y α se obtuvo un cociente $n/p_{WAIC} < 20$, considerando que las distribuciones a priori que generen un cociente $n/p_{WAIC} < 20$ favorecen el sobreajuste. Este criterio fue propuesto por Beguin et al. (2012). En este trabajo tomamos el umbral de 20 de manera orientativa y aproximada, es decir, no descartamos distribuciones a priori sólo por generar un cociente $n/p_{WAIC} < 20$ en el análisis de sensibilidad.

En los mapas del efecto aleatorio se intentó identificar las distribuciones a priori que generaron escalas del efecto aleatorio menores a la escala espacial de las covariables. Se consideró que una

distribución a priori favorecía el sobreajuste si la escala espacial observada en el mapa del efecto aleatorio era menor a la escala espacial observada en las covariables elegidas. Este procedimiento se basó en el trabajo de Illian et al. (2010), que indican que para evitar el sobreajuste y obtener un modelo que describa una tendencia interpretable, el efecto espacial debe operar en una escala espacial similar a la de las covariables. Si el efecto espacial opera a una escala menor que la escala de las covariables, podría explicar los datos mejor que ellas (Illian et al., 2010; Illian et al., 2012).

Para elegir la distribución a priori final por cada tipo de modelo se seleccionaron aquellas que generaron modelos con $RMSE$ y MAE bajos tanto en la muestra de entrenamiento como en la de testeo. Se consideró que aquellas con grandes diferencias entre entrenamiento y testeo podrían generar mayor tendencia al sobreajuste. El resultado final fue una única distribución a priori penalizada por complejidad para cada tipo de modelo que no debería introducir mayor flexibilidad al modelo que la apoyada por los datos.

3.2.2. Análisis del impacto de la multicolinealidad

Para analizar el impacto de la multicolinealidad se entrenaron modelos con distintos conjuntos de datos donde en cada conjunto se aplicó una técnica distinta para reducir la multicolinealidad, y luego se compararon los resultados y la performance de los modelos. Las técnicas aplicadas fueron clustering jerárquico de covariables para detectar grupos de variables colineales, y PCA y MULTISPATI para resumir la información de las covariables. La aplicación de las técnicas generó cuatro conjuntos de datos alternativos, además de los datos originales, que se describen detalladamente en la siguiente sección.

Una vez generados los conjuntos alternativos de datos, se analizaron las posibles diferencias entre la aplicación de PCA y MULTISPATI debidas al distinto manejo de la información espacial entre técnicas. El análisis se hizo a nivel de *pérdida de inercia*, *ganancia en información espacial* y cargas de las covariables en las componentes principales.

A continuación se ajustó cada tipo de modelo (MLR, *iid*, $iCAR$ y BYM2) con los cuatro conjuntos alternativos y con los datos originales, y se realizó una comparación de coeficientes de regresión y capacidad predictiva entre modelos. Esta comparación tuvo dos propósitos: (a) evaluar cambios debidos a entrenar el mismo modelo con conjuntos de datos cuya multicolinealidad se trató de distinta forma, y (b) evaluar cambios en el impacto de la multicolinealidad entre los modelos

espaciales y los no espaciales.

Generación de conjuntos alternativos de covariables

En total se trabajó con cinco conjuntos alternativos de covariables.

Al Conjunto 1 o "conjunto control" no se le aplicó ninguna técnica de reducción de multicolinealidad y por lo tanto presentó la multicolinealidad original de los datos preprocesados. Estuvo formado por las 36 covariables originales (ver Tabla 9) y versiones cuadráticas de las covariables Elevación, Pendiente y VUT.

El Conjunto 2 se generó aplicando PCA a las 36 covariables del Conjunto 1 y reteniendo las primeras componentes principales capaces de explicar el 80 % de la variabilidad en los datos. No se incorporaron términos polinómicos de ninguna componente principal porque no se observaron relaciones no lineales importantes entre la respuesta y las covariables. El Conjunto 3 se generó exactamente igual que el Conjunto 2, pero aplicando MULTISPATI. Los Conjuntos 2 y 3 se denominaron "PCA80" y "MULTISPATI80" respectivamente.

Para crear los Conjuntos 4 y 5 primero fue necesario identificar los grupos de covariables colineales. La identificación se hizo con el algoritmo de clústering jerárquico mencionado en la Sección 2.7.3, implementado en la función `varclust` de la librería `Hmisc` de `R` (Harrell, 2015). La matriz de similaridad se calculó con el coeficiente de correlación de Pearson.

Una vez identificados los clústeres de variables colineales, se les aplicaron PCA y MULTISPATI a cada clúster por separado para resumir la información contenida en ellos. Se consideró que las variables en un clúster colineal pueden ser resumidas por su primera componente principal siempre y cuando esta explique más del 70 % de la variabilidad total.

El Conjunto 4 se generó reemplazando cada grupo colineal por su primera componente principal obtenida con PCA. sino fue posible resumir un grupo colineal con su primera componente, se mantuvieron sin cambios las variables del grupo colineal en la base de datos, porque se consideró que no fueron lo suficientemente colineales como para ser resumidas con una única componente sintética. Si los clusteres mostraron relaciones polinomicas con la respuesta, se agregaron terminos polinomicos de los clusteres al conjunto de datos. El Conjunto 5 se generó igual que el Conjunto

4, pero aplicando MULTISPATI en lugar de PCA. Los conjuntos 4 y 5 se denominaron "PCA-hclust" y "MULTISPATI-hclust" respectivamente.

En los cinco conjuntos se estandarizaron todas las covariables antes de proceder al ajuste de los modelos.

Criterios de comparación pre-modelado

Para comprender si hubieron diferencias sustanciales entre la reducción de multicolinealidad lograda con PCA y la lograda con MULTISPATI debido a distinto manejo de la información espacial entre técnicas se compararon:

1. Las cargas de las covariables en las primeras componentes PCA versus las cargas en las primeras componentes MULTISPATI. Se consideró que cargas similares en magnitud y signo implicaron resultados similares respecto a la descomposición de la variabilidad.

2. Los índices de autocorrelación de Moran de las primeras componentes PCA versus los índices en las primeras componentes principales MULTISPATI. Se analizó la ganancia en información espacial ente las componentes de PCA y MULTISPATI PCA, donde siendo I el índice de Moran se definió la ganancia en información espacial GIE_i para la componente i como

$$GIE_i = (I_{PCA_i} - I_{MULTISPATI_i})/I_{MULTISPATI_i}. \tag{78}$$

   Se consideró que ganancias cercanas a cero implicaron que no había diferencias en la captación de la información espacial entre PCA y MULTISPATI.

3. La variabilidad explicada por cada componente. Se analizó la pérdida de inercia entre las componentes de PCA y MULTISPATI, donde siendo v la variabilidad se definió a pérdida de inercia PI_i en la componente i como

$$PI_i = (v_{PCA_i} - v_{MULTISPATI_i})/v_{PCA_i}. \tag{79}$$

   Se consideró que pérdidas cercanas a cero implicaron que no había diferencias en la captación de la inercia entre PCA y MULTISPATI.

Modelos ajustados

Para analizar el impacto de la multicolinealidad en la capacidad predictiva y en los coeficientes de regresión se ajustaron cuatro tipos de modelos de regresión bayesianos:

1. Regresión lineal múltiple sin efectos aleatorios ("MLR").
2. Regresión lineal con efectos aleatorios *iid* a nivel de región ("modelo *iid*").
3. Regresión lineal con efectos aleatorios $iCAR$ a nivel de región ("modelo iCAR").
4. Regresión lineal con efectos aleatorios BYM a nivel de región ("modelo BYM2")

En todos los casos:

- La estimación se hizo con INLA a través del paquete `R-INLA` (Ver Sección 2.2.2).
- La verosimilitud se modeló gaussiana.
- Se mantuvo la elección de distribuciones a priori por defecto de `R-INLA` para los coeficientes de regresión y para la varianza del error de las observaciones (ver Sección 2.5.1).
- Para los hiperparámetros de los efectos aleatorios se utilizó la distribución a priori seleccionada en el análisis de sensibilidad (ver 3.2.1).

Los cuatro modelos MLR, *iid*, $iCAR$ y BYM2 se ajustaron con los cinco conjuntos de datos "Control", "PCA80", "MULTISPATI80", "PCA-hclust" y "MULTISPATI-hclust", haciendo un total de 20 modelos.

Criterios de comparación post-modelado

Para evaluar el impacto de la multicolinealidad en el ajuste se compararon los cambios en los coeficientes de regresión y en la capacidad predictiva de cada modelo, medida con $RMSE$ y MAE, al entrenarlos con los conjuntos de datos alternativos que difirieron en el tratamiento de la multicolinealidad.

Los cambios en los coeficientes de regresión se analizaron comparando dentro de cada modelo las diferencias entre los coeficientes obtenidos con los conjuntos "Control", "PCA-hclust" y "MULTISPATI-hclust". Además, solo se miraron los coeficientes de las covariables que no participaron de los grupos colineales. Los conjuntos "PCA80" y "MULTISPATI80" no se incluyeron en esta comparación porque no tienen covariables en común con los conjuntos "Control", "PCA-hclust" y "MULTISPATI-hclust". Los cambios en los coeficientes se analizaron a nivel de magnitud y sentido de la media a posteriori, y de ancho del intervalo de credibilidad.

Los cambios en la capacidad predictiva se analizaron comparando dentro de cada modelo las diferencias en $RMSE$ y MAE obtenidas al entrenar con los cinco conjuntos de datos.

Este enfoque nos permitió concluir sobre el impacto de las técnicas de control de la multicolinealidad sobre los coeficientes de regresión y la capacidad predictiva en general, descontando el efecto "tipo de modelo".

Para evaluar si el efecto de la multicolinealidad en los coeficientes de regresión y en la capacidad predictiva fue distinto en los modelos de regresión con efecto aleatorio espacial respecto a los modelos de regresión sin efecto aleatorio espacial, se analizó si el comportamiento general de los modelos MLR e *iid* fue distinto al comportamiento de los modelos $iCAR$ y BYM al entrenar los modelos con los conjuntos de datos alternativos.

3.2.3. Selección del mejor modelo predictivo

El último paso consistió en elegir el mejor modelo predictivo de los 20 modelos generados al entrenar cada tipo de modelo con cada uno de los cinco conjuntos de datos alternativos.

La plausibilidad de los modelos se evaluó mediante el análisis de residuos, de CPO_i y de PIT_i (ver Sección 2.6.1). Respecto a los residuos se analizaron las nubes de residuos versus predichos, la autocorrelación espacial residual con el índice de Moran, y la normalidad de los residuos a través de histogramas. Adicionalmente se calculó mediante validación cruzada espacial la correlación r de Pearson entre los residuos y los predichos; se consideró la obtención de diferencias sustanciales entre los valores obtenidos para las particiones de entrenamiento y de testeo r_E y r_T como una alerta de sobreajuste.

El ajuste se evaluó con DIC, WAIC y LPML calculados de manera tradicional, y con $RMSE$ y

MAE de entrenamiento y testeo calculados con validación cruzada espacial.

4. Resultados y discusión

Este capítulo está organizado como sigue. Primero, se comentan y discuten los resultados del análisis de sensibilidad que se hizo para elegir distribuciones a priori que no fuercen el sobreajuste en los modelos con efectos aleatorios *iid*, $iCAR$ y BYM2 . Segundo, se describen el impacto observado de la multicolinealidad en el ajuste de los modelos MLR, *iid*, $iCAR$ y BYM2 y las diferencias observadas entre las técnicas utilizadas para tratar la multicolinealidad, y se discute el efecto de la multicolinealidad en general para los modelos jerárquicos bayesianos y en particular para los modelos espaciales. Tercero y último, se describe la plausibilidad y capacidad predictiva de los modelos finales y la posterior elección del mejor modelo predictivo para la respuesta ESPI en el territorio de la Cuenca del Carcarañá.

4.1. Análisis de sensibilidad

En los modelos *iid* e $iCAR$ se analizó la distribución a priori de la precisión del efecto aleatorio, y en el modelo BYM2, la de la precisión y la del hiperparámetro de mezcla. En todos los casos se utilizaron distribuciones a priori penalizadas por complejidad para los hiperparámetros, formuladas en función de los parámetros U y alpha (ver Sección 2.5.2).

4.1.1. Modelo *iid*

Para el modelo *iid* se analizó la sensibilidad a cambios en la distribución a priori penalizada por complejidad para la precisión τ del efecto aleatorio. La distribución a priori se expresó en función de U y de α (ver Ecuación 55). En general, en el modelo *iid* la variabilidad del efecto aleatorio aumentó con U y con α, pero no se observó un efecto de suavizado espacial, que asociamos con la ausencia de estructura espacial explícita en el efecto aleatorio *iid*.

La Figura 5 muestra el WAIC en función de α para distintos valores de U. Los modelos ajustados con $U = 0.5$ obtuvieron el mínimo WAIC, cercano a -25000, para todos los α. Por el contrario, los modelos ajustados con $U = 0.001$ obtuvieron el mayor WAIC, cercano a 1000. Los valores intermedios $0.001 < U < 0.5$ mostraron en general curvas decrecientes para el WAIC decreciente, comenzando cerca de 1000 y disminuyendo a medida que aumentó α. La gran diferencia en WAIC

entre los modelos con U altos y U bajos sugiere que el WAIC alcanzado con U altos puede corresponder al WAIC de un modelo que ajusta perfectamente a los datos.

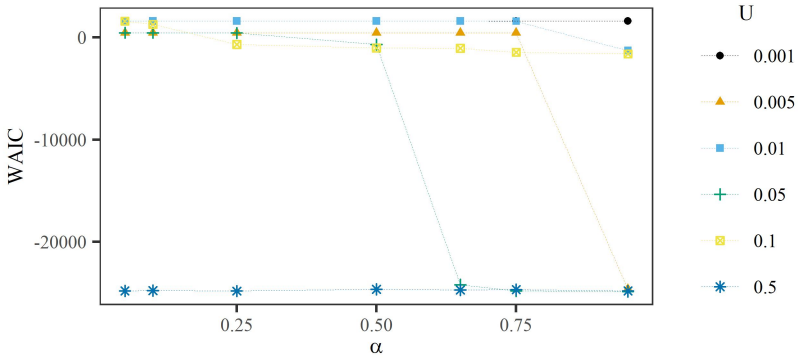

Figura 5: WAIC en funcion del parámetro α para seis valores distintos de U en el modelo *iid*.

La Figura 6 muestra el cociente entre la cantidad de observaciones ($n = 4676$) y el número de parámetros efectivos calculados con WAIC, p_{WAIC}. La línea gruesa gris representa el umbral $n/p_{WAIC} > 20$, sugerido por Beguin et al. (2012) para detectar modelos con tendencia al sobre-ajuste. Los valores altos $U \geq 0.05$ generaron modelos lo suficientemente complejos como para lograr $n/p_{WAIC} < 20$ en casi todo el recorrido α. Los valores mas pequeños $U \leq 0.01$ mostraron un decaimiento gradual del cociente datos-parámetros, pero que se mantuvo por encima de 20 en prácticamente todos los valores de α.

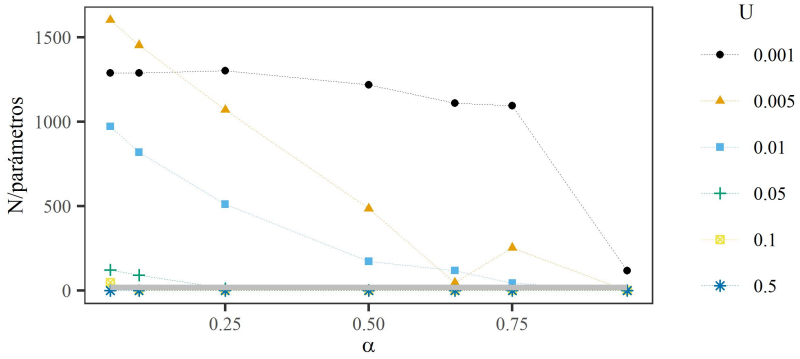

Figura 6: Cociente entre la cantidad de regiones $n = 4676$ y p_{WAIC} para el modelo *iid*.

Los resultados observados en los gráficos del WAIC y del cociente datos-parámetros versus α sugieren la misma conclusión: valores muy bajos de U no generaron modelos lo suficientemente complejos, valores muy altos de U generaron modelos demasiado complejos, y valores intermedios

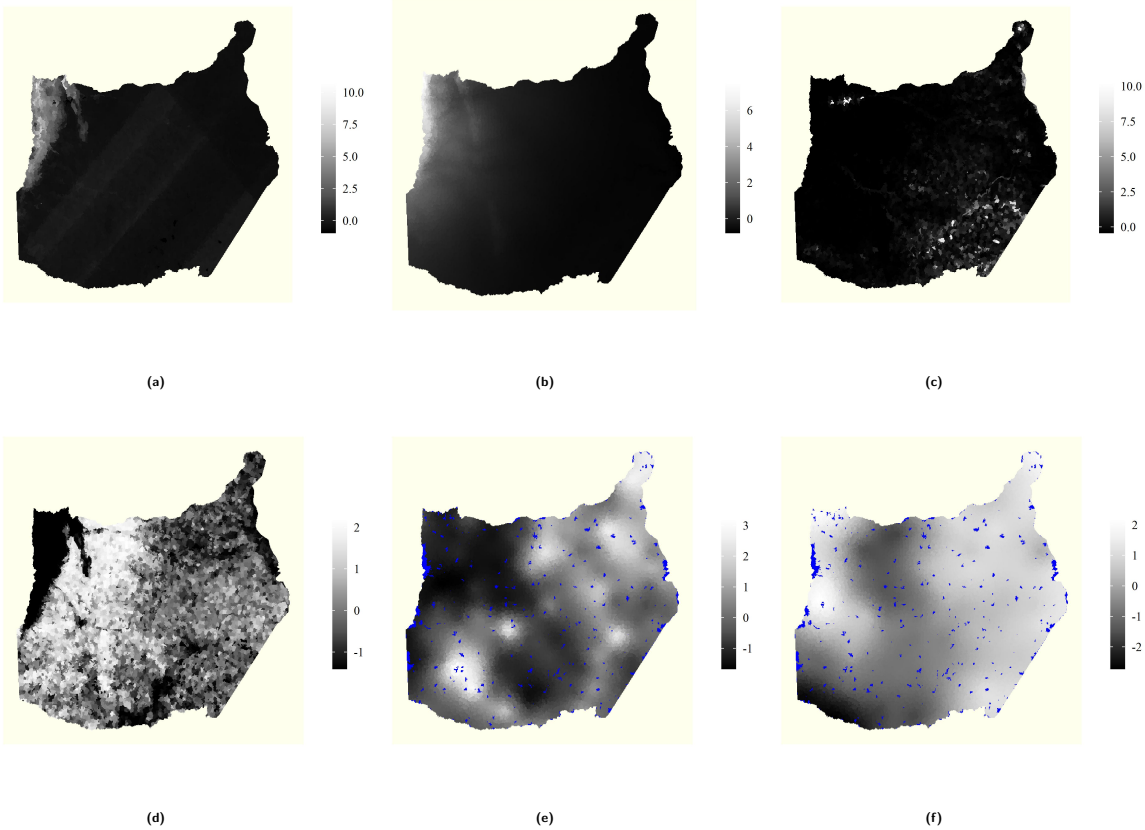

Figura 7: Mapas de covariables. Las covariables están estandarizadas y el color azul indica observaciones faltantes. (a)Pendiente. (b) Elevación. (c) Recurrencia. (d)Actividades estivales. (e) P. (f) Nt.

de U como $U = 0.005$ y $U = 0.01$ podrían ser candidatos para obtener un buen ajuste.

Además de examinar el WAIC y el cociente n/p_{WAIC}, comparamos la escala del efecto aleatorio generada por cada distribución a priori con la escala de las covariables Pendiente, Elevación, Recurrencia, Actividades Estivales, P y Nt, consideradas con alto potencial predictivo. La Figura 7 muestra mapas de las covariables y la Figura 8 muestra mapas de la media a posteriori del efecto aleatorio para los valores intermedios $U = 0.005$, $U = 0.01$ y $U = 0.05$.

El efecto aleatorio tuvo baja variabilidad para casi todos los valores de $U = 0.005$ y $U = 0.01$ y por lo tanto los mapas correspondientes mostraron una escala inferior a la escala de las covariables. La excepción fue ($U = 0.01$, $\alpha = 0.75$) (Figura 8f), que permitió suficiente variabilidad como para resultar en una escala similar a la de la Pendiente y la Elevación, pero más gruesa que

la escala de las Actividades Estivales. Los valores de $U = 0.05$ mostraron en general escalas comparables o más finas que $(U = 0.01,\ \alpha = 0.75)$. Con $(U = 0.05,\ \alpha = 0.5)$ la escala se observó más fina que todas las covariables.

Se eligieron dos distribuciones candidatas a generar modelos que no favorezcan el sobreajuste: $(U = 0.01,\ \alpha = 0.75)$, porque generó una escala comparable a las covariables, y $(U = 0.05,\ \alpha = 0.25)$, que tuvo una escala ligeramente más fina que la mayoría de las covariables pero que estuvo a la altura de las Actividades Estivales.

4.1.2. Modelo con efectos aleatorios iCAR

Para el modelo $iCAR$ se analizó la sensibilidad a cambios en la distribución a priori penalizada por complejidad para la precisión τ del efecto aleatorio.

En el modelo $iCAR$ la variabilidad del efecto aleatorio aumentó con U y con α igual que en el modelo *iid*. Sin embargo, a diferencia del modelo *iid*, donde se observaron cambios abruptos y WAICs constantes para algunos valores de U, en el gráfico del WAIC versus α (Figura 9) se observó un decaimiento gradual del WAIC en función de α para todos los valores de U.

La Figura 10 muestra el cociente n/p_{WAIC}. Casi todos los valores de U generaron, para cualquier α, modelos lo suficientemente complejos como para que el cociente quede por debajo de 20, lo que nos hizo sospechar que quizá el umbral de $n/p_{WAIC} = 20$ es demasiado estricto para este modelo jerárquico complejo.

La Figura 11 muestra la escala espacial del efecto aleatorio para $U = 0.0001$, $U = 0.0005$ y $U = 0.001$ y distintos valores de α. Se observó suavizado espacial y aumento gradual de la escala del efecto espacial, comportamiento que no se observó en el modelo *iid*, y que coincide con el decaimiento gradual del WAIC de la Fig. 9. Los mapas para $U = 0.0001$ en la primera fila de la Fig. 11 mostraron escalas demasiado gruesas excepto quizá $\alpha = 0.65$ (Fig. 11c). Los mapas para $U = 0.0005$ en la segunda fila mostraron en general escalas comparables a las de las covariables. Los gráficos de la tercera fila para $U = 0.001$ mostraron escalas comparables o más finas que las covariables.

Finalmente se decidió preseleccionar las distribuciones a priori $(U = 0.0001,\ \alpha = 0.65)$, y $(U = 0.0005,\ \alpha = 0.15)$, que mostraron escalas un poco más finas que la Pendiente y la Elevación, pero

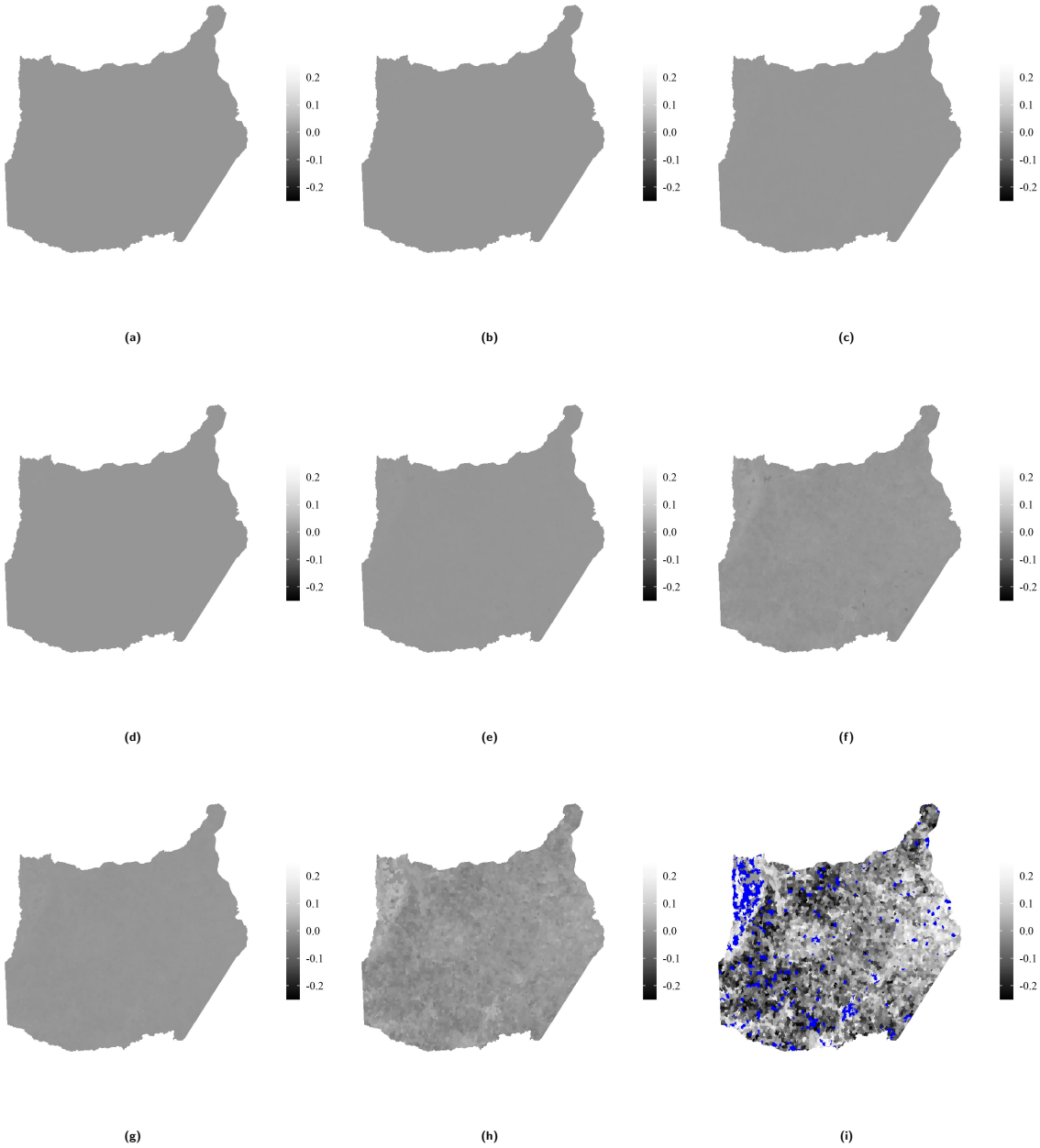

Figura 8: Media a posteriori del efecto aleatorio en el modelo *iid*. Primera fila: $U = 0.005$ con (a) $\alpha = 0.25$, (b) $\alpha = 0.5$ y (c) $\alpha = 0.75$. Segunda fila: $U = 0.01$ con (d) $\alpha = 0.25$, (e) $\alpha = 0.5$ y (f) $\alpha = 0.75$. Tercera fila: $U = 0.05$ con (g) $\alpha = 0.1$, (h) $\alpha = 0.25$ y (i) $\alpha = 0.5$. El color azul significa que la media del efecto aleatorio observado quedó fuera de la escala de colores.

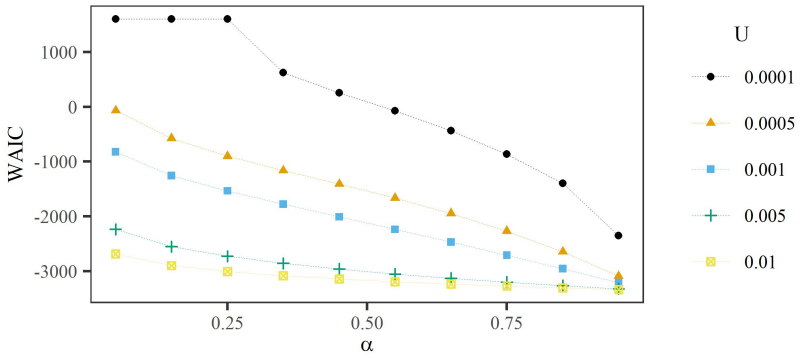

Figura 9: WAIC en función de α para el efecto aleatorio iCAR.

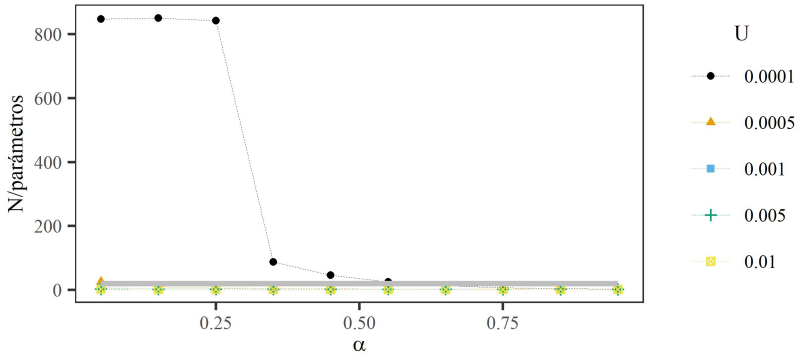

Figura 10: Cociente n/p_{WAIC} en función de α para el efecto aleatorio iCAR.

un poco más gruesas que las Actividades Estivales. Además se agregó ($U = 0.0001,\ \alpha = 0.75$) para verificar una escala un poco más fina que ($U = 0.0001,\ \alpha = 0.65$).

4.1.3. Modelo con efectos aleatorios BYM2

El análisis de sensibilidad para el modelo BYM2 se hizo en términos de las distribuciones a priori penalizadas por complejidad de los dos hiperparámetros del BYM2, la precisión τ y el parámetro de mezcla ϕ. Para no confundir el U de la precisión con el U de ϕ, se denominaron ($U_1,\ \alpha_1$) a los parámetros de la distribución de la precisión, y ($U_2,\ \alpha_2$) a los de la distribución de ϕ. Se probaron 50 opciones de distribuciones a priori para la precisión, y 4 opciones de distribuciones para ϕ, lo que hizo un total de 200 distribuciones. Las 4 opciones para ϕ se obtuvieron fijando $U2 = 0.5$ y variando α_2 (ver Sección 3.2.1 para la justificación de esta elección).

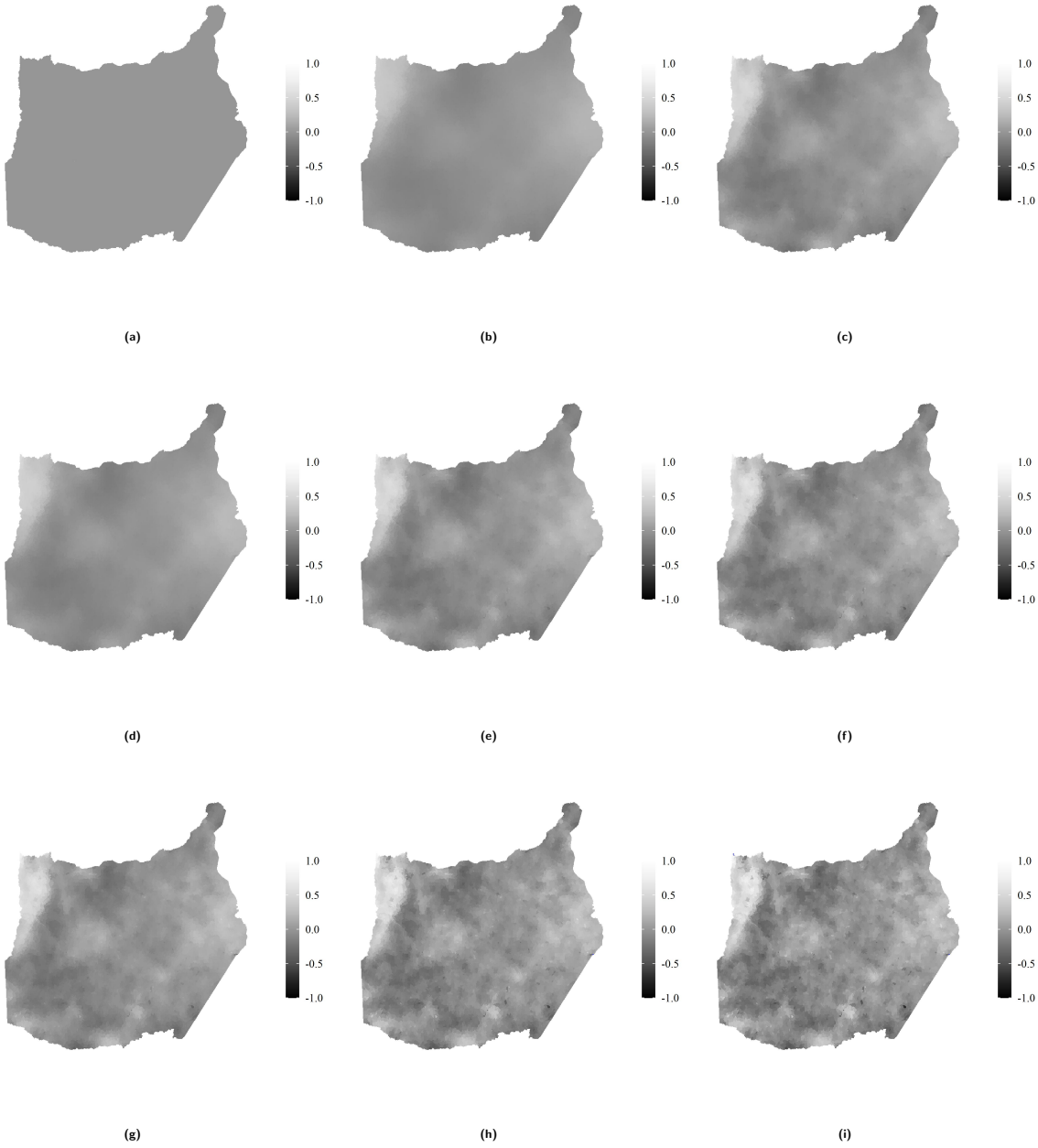

Figura 11: Media a posteriori del efecto aleatorio iCAR. *Primera fila*: $U = 0.0001$ con (a) $\alpha = 0.25$, (b) $\alpha = 0.45$ y (c) $\alpha = 0.65$. *Segunda fila*: $U = 0.0005$ con (d) $\alpha = 0.05$, (e) $\alpha = 0.15$ y (f) $\alpha = 0.25$. *Tercera fila*: $U = 0.001$ con (g) $\alpha = 0.05$, (h) !‘$\alpha = 0.15$ y (i) $\alpha = 0.25$.

Para analizar el efecto "marginal" de la distribución a priori de la precisión en el WAIC se graficó el WAIC promedio versus α_1 para distintos valores de U_1 (Figura 12), donde el WAIC se promedió sobre las posibles opciones de $(U_2,\ \alpha_2)$.

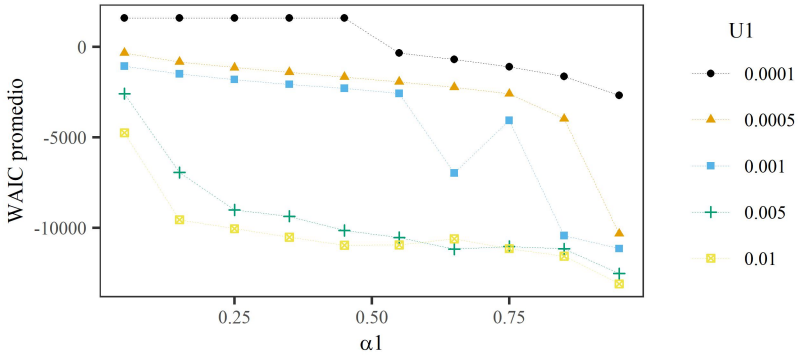

Figura 12: WAIC promedio en función de α_1 para distintos valores de U_1. El promedio se calcula sobre los posibles $(U_2,\ \alpha_2)$.

Se observó la misma tendencia decreciente que en los modelos *iid* e *iCAR* del WAIC en función del aumento de α_1 y de U_1. Sin embargo, las curvas del WAIC promedio en el BYM2 se vieron menos suaves que las curvas del WAIC en el efecto aleatorio iCAR. Suponemos que esto pudo deberse a la incorporación del efecto aleatorio *iid* en el modelo BYM, ya que en las curvas WAIC para el modelo *iid* también se observaron cambios abruptos en el WAIC.

La Figura 13 muestra el cociente n/p_{WAIC} versus α_1. Análogamente al caso del WAIC promedio, p_{WAIC} también se promedió sobre los posibles $(U2,\ \alpha_2)$ para poder analizar el efecto marginal de la distribución de la precisión. Se observó que sólo $U_1 = 0.0001$ generó modelos lo suficientemente sencillos como para obtener $n/p_{WAIC} > 20$, pero también se notó un cambio muy abrupto en el cociente, que pasó de 1500 en $\alpha_1 = 0.45$ (lo que equivale a unos 3 parámetros efectivos con $n = 4676$) a ser menor a 20 en $\alpha_1 = 0.55$ (lo que equivale a al menos 200 parámetros). El resto de los U_1 en todos los casos generaron modelos con $n/p_{WAIC} < 20$.

El hecho de que el cociente haya sido prácticamente siempre $n/p_{WAIC} < 20$ nos hizo sospechar que en el modelo BYM2, al igual que en el modelo con efectos aleatorios iCAR, el umbral de 20 para el cociente datos-parámetros puede ser demasiado estricto. La Figura 14 muestra el mismo gráfico que la Figura 13 pero eliminando el valor $U1 = 0.0001$, para poder apreciar mejor las curvas de $U1$ que generaron siempre cocientes inferiores a 20.Se observó un decaimiento bastante intenso del cociente n/p_{WAIC} con el aumento de α_1.

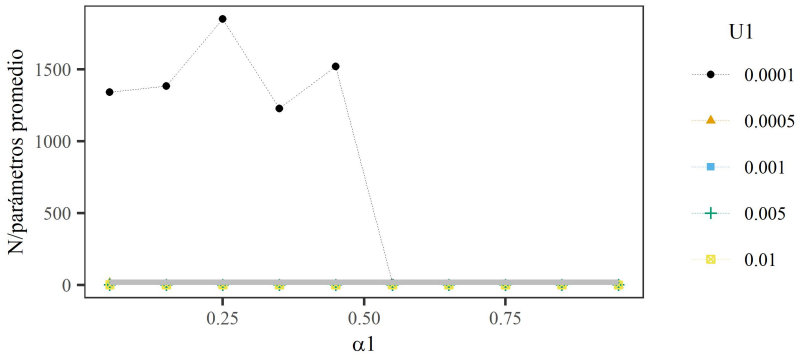

Figura 13: Cociente entre la cantidad de regiones $n = 4676$ y el p_{WAIC} promedio para el efecto aleatorio BYM2. El p_{WAIC} promedio se calculó promediando sobre los posibles $(U2,\ \alpha_2)$.

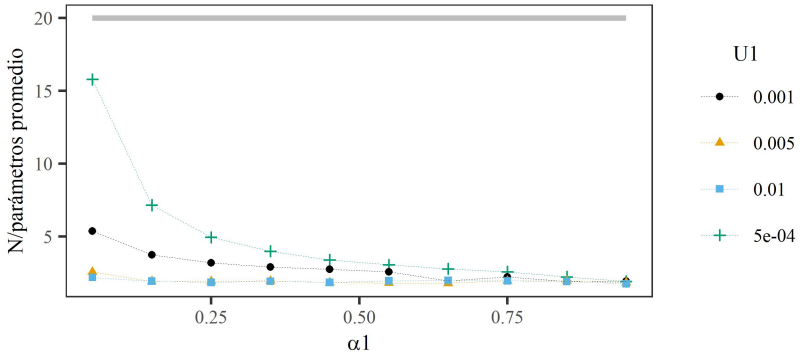

Figura 14: Cociente entre la cantidad de regiones $n = 4676$ y la cantidad de parámetros efectivos promedio para el efecto aleatorio BYM2 eliminando la curva de $U_1 = 0.0001$. El promedio del número de parámetros efectivos se calculó promediando sobre los posibles $(U2,\ \alpha_2)$.

Luego de analizar el comportamiento promedio del WAIC y del cociente n/p_{WAIC} para distintos valores de U_1, revisamos el comportamiento de $n/p_{WAIC} < 20$ pero desdoblado para los distintos valores de $(U_2,\ \alpha_2)$ con los que se combinó cada $U1$. Las Figuras 15 y 16 muestran estas curvas desdobladas para $U1 = 0.0001$ y $U1 = 0.0005$.

Si bien en el caso $U1 = 0.0001$ se observaron diferencias entre las curvas para distintos α_2, no se pudo extrapolar un comportamiento general para el efecto de α_2 en la determinación del número de parámetros efectivos. Es decir, los valores del cociente no fueron sistemáticamente menores o mayores a medida que cambio α_2. La curva más estable pareció ser la de $(U2 = 0.5,\ \alpha_2)$. En el caso de $U1 = 0.0005$ no se observaron diferencias en el cociente datos-parámetros generado por distintas α_2.

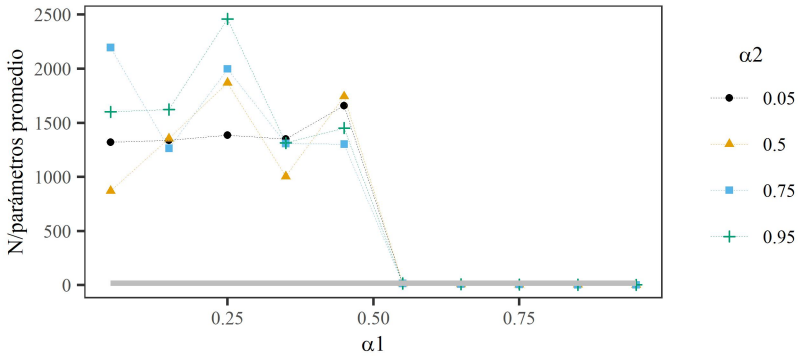

Figura 15: Cociente datos-parámetros en función de α_1 para $U1 = 0.0001$, $U2 = 0.5$ y distintos valores de α_2. Modelo BYM2.

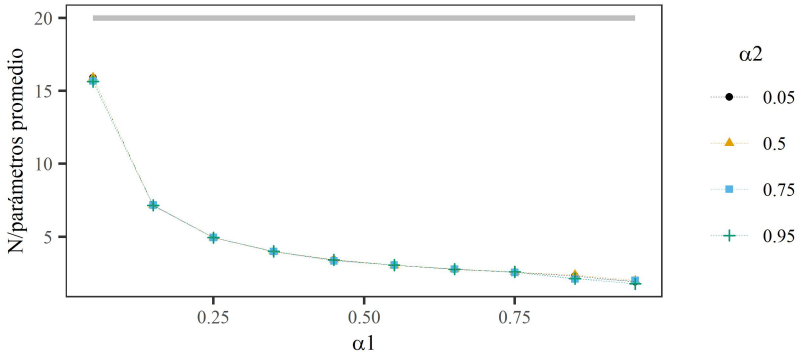

Figura 16: Cociente datos-parámetros en función de α_1 para $U1 = 0.0005$, $U2 = 0.5$ y distintos valores de α_2. Modelo BYM2.

Con base en la falta de un patrón claro en las curvas de la Figura 15 y en la ausencia de diferencias para distintos α_2 en la Figura 15, sospechamos que la elección de la distribución a priori de ϕ podría no ser tan importante como la distribución a priori de la precisión τ a la hora de determinar la variabilidad a posteriori del efecto aleatorio. Por este motivo el análisis continuó fijando $(U_2 = 0.5,\ \alpha_2 = 0.05)$ en todos los casos.

La Figura 17 muestra la media a posteriori del efecto aleatorio para $U1 = 0.0001$, $U1 = 0.0005$ y $U1 = 0.001$. Se observó menor suavización en el efecto aleatorio que en el modelo iCAR. Además, se observó que pequeños cambios en α_1 para el mismo $U1$ generaron cambios abruptos en la escala del efecto aleatorio. Este comportamiento se vio, por ejemplo, en el cambio de escala entre las Figuras 17a y 17b, que corresponden a $U = 0.0001$ con $\alpha_1 = 0.45$ y $\alpha_1 = 0.55$. Esta situación coincide los cambios abruptos observados en las curvas WAIC de la Fig. 12 y creemos

que puede estar motivada por la inclusión del efecto *iid*.

Con $U_1 = 0.001$ se obtuvieron escalas más finas que con $U_1 = 0.0005$ pero en ambos casos se logró un mejor control de la escala del efecto que con $U_1 = 0.0001$. Se decidió pre-seleccionar las distribuciones $(U1 = 0.0005,\ \alpha_1 = 0.25)$, $(U1 = 0.0005,\ \alpha_1 = 0.4)$, $(U1 = 0.0005,\ \alpha_1 = 0.5)$ y $(U1 = 0.001,\ \alpha_1 = 0.1)$.

4.1.4. Selección final de una única distribución a priori por modelo

Para elegir una única distribución a priori por modelo dentro de las candidatas preseleccionadas se ajustaron modelos no nulos, que modelaron al ESPI en función las covariables incluidas en el conjunto de datos "Control", y se eligió el mejor modelo a través de medidas comparativas.

Con la comparación de modelos intentamos detectar las distribuciones a priori que permitieron mejor balance entre subajuste causado por penalización extrema al efecto aleatorio, y sobreajuste causado por no penalizar lo suficiente. Creemos que basar la comparación de modelos únicamente en el WAIC puede no ser adecuado, porque el WAIC podría favorecer a los modelos que sobreajustan. Además, sospechamos que la regla $n/p_{WAIC} < 20$ podría ser demasiado estricta para modelos con efectos espaciales. Por lo tanto, comparamos los modelos considerando las medidas $RMSE$ y MAE calculadas con validación cruzada espacial además del WAIC y del cociente datos-parámetros. Estas medidas por validación cruzada espacial nos permitieron evaluar el ajuste del modelo en datos de testeo y no solo en datos de entrenamiento. La Tabla 2 muestra las medidas comparativas utilizadas.

Los resultados de la Tabla 2 mostraron que que los modelos con menor WAIC también fueron los modelos más complejos, con mayor cantidad de parámetros efectivos y menor cociente datos-parámetros, pero que no siempre se correspondieron con los modelos con menor error de predicción calculado con conjuntos de testeo. Por ejemplo, en los modelos BYM2, guiarse únicamente por el WAIC implicaría elegir modelo 7, con menor WAIC, pero este modelo no fue el modelo con menor $RMSE$ o MAE; en los modelos *iid*, el modelo 9 tuvo casi el triple número de parámetros efectivos que el modelo 8, y sustancialmente menor WAIC, aunque mostró exactamente los mismos $RMSE$ Y MAE de entrenamiento y testeo que el modelo 8.

En el caso del modelo *iid* las métricas por validación cruzada no mostraron que una distribución

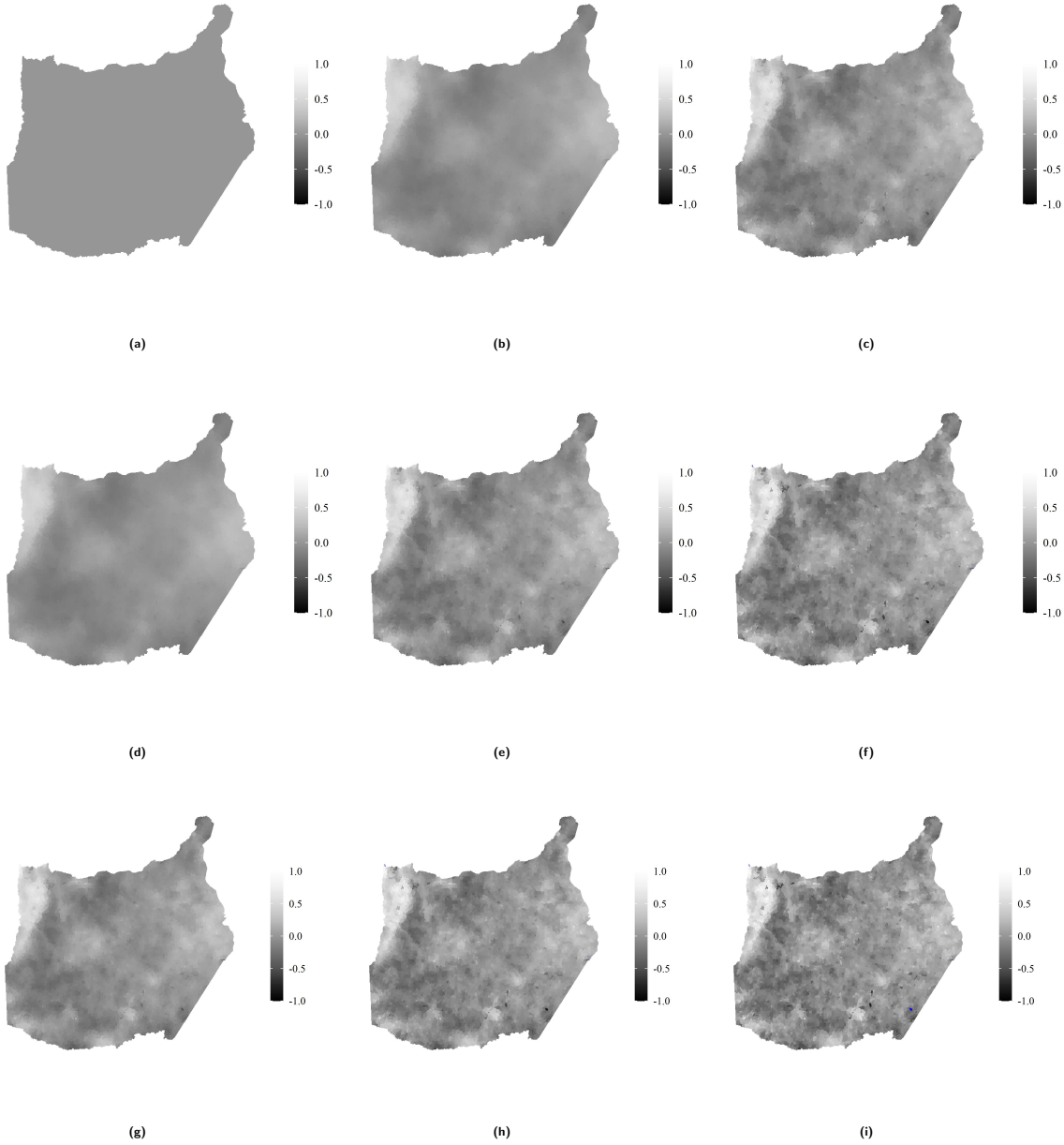

Figura 17: Media a posteriori del efecto aleatorio BYM2. *Primera fila*: $U1 = 0.0001$ con (a) $\alpha_1 = 0.45$, (b) $\alpha_1 = 0.55$ y (c) $\alpha_1 = 0.75$. *Segunda fila*: $U1 = 0.0005$ con (d) $\alpha_1 = 0.05$, (e) $\alpha_1 = 0.25$) y (f) $\alpha_1 = 0.55$). *Tercera fila*: $U1 = 0.001$ con (g) $\alpha_1 = 0.05$, (h) $\alpha_1 = 0.25$ y (i) $\alpha_1 = 0.55$.

Tabla 2: Medidas de comparación de los modelos ajustados sobre el Conjunto 1.

Modelo	Tipo de Modelo	U	α	WAIC	n/p_{WAIC}	RMSE_E	RMSE_T	MAE_E	MAE_T
1	*iCAR*	1e-04	0.65	-2045	16.12	0.207	0.232	0.158	0.175
2	iCAR	1e-04	0.75	-2806	5.66	0.155	0.221	0.119	0.163
3	*iCAR*	**5e-04**	**0.15**	-2380	9.72	0.187	0.227	0.142	0.169
4	**bym2**	**0.001**	**0.1**	-3394	3.54	0.146	0.203	0.11	0.154
5	bym2	5e-04	0.25	-1300	69.11	0.11	0.22	0.084	0.161
6	bym2	5e-04	0.4	-3685	3.14	0.095	0.22	0.073	0.161
7	bym2	5e-04	0.5	-4080	2.79	0.136	0.218	0.102	0.16
8	**iid**	**0.01**	**0.75**	-1331	10.92	0.186	0.231	0.14	0.174
9	iid	0.05	0.25	-1591	3.6	0.186	0.231	0.14	0.174
10	mlr	NA	NA	-1299	69.24	0.207	0.231	0.158	0.174

Los subíndices $_\text{E}$ y $_\text{T}$ corresponden a valores calculados sobre las muestras de entrenamiento y testeo respectivamente. Las distribuciones a priori en **negrita** resultaron elegidas.

a priori generase mejores resultados que la otra. Además, ambas tuvieron performance similar al modelo lineal base sin efectos aleatorios. Por lo tanto se eligió la distribución a priori ($U = 0.01,\ \alpha = 0.75$), porque generó el modelo más sencillo.

En el caso del modelo $iCAR$ se eligió el modelo 3 con distribución a priori ($U = 5E-04,\ \alpha = 0.15$) porque si bien obtuvo los segundos menores MAE_T y $RMSE_T$ dentro de las opciones $iCAR$, las diferencias al modelo 2 con menor MAE y $RMSE$ no fueron grandes ni en testeo ni entrenamiento, por lo tanto elegimos el modelo 3, menos complejo que el 2.

En el modelo BYM2 se eligió la distribución ($U = 0.001,\ \alpha = 0.1$) porque obtuvo los menores MAE_T y $RMSE_T$ dentro de las opciones BYM2.

4.2. Análisis del impacto de la multicolinealidad

Para determinar los efectos de la multicolinealidad en el ajuste entrenamos modelos con cinco conjuntos alternativos de datos, que fueron generados con la aplicación de PCA y MULTISPATI a las variables y a los grupos de variables colineales detectados con clústering jerárquico.

La Sección 4.2.1 describe la resultados de la aplicacion de PCA, MULTISPATI y clustering jerárquico a las variables del conjunto de datos de la Cuenca del Carcarañá y presenta los resultados de la comparación entre PCA y MULTISPATI en términos de pérdida de inercia, ganacia de información y cargas de las covariables.

La Sección 4.2.2 describe el impacto de la multicolinealidad en los coeficientes de regresión y en la capacidad predictiva de los modelos MLR, *iid*, $iCAR$ y BYM2 entrenados con cinco conjuntos alternativos de covariables.

4.2.1. Comparación de PCA y MULTISPATI pre-modelado

En el conjunto "PCA80" que se generó aplicando PCA a las 36 covariables de la Tabla fue posible explicar el 80 % de la variabilidad de las 36 covariables con las primeras 12 componentes principales. En el conjunto "MULTISPATI80" se generó aplicando MULTISPATI a las mismas 36 covariables se necesitaron las primeras 14 componentes principales para explicar el 80 % de la variabilidad.

La Tabla 10 en el Anexo B muestra las cargas de las covariables en las primeras cinco componentes principales de los conjuntos "PCA80" y "MULTISPATI80". Las primeras componentes PCA y MULTISPATI tuvieron sentido opuesto, pero representaron la misma variable latente: valores altos de la primera componente indicaron altos valores de Arena y Actividades Estivales, y bajos valores de CIC, CC, Arcilla, Nt, CEextsat, Mn, Cu, Materia Orgánica, K, Mg y P. La segunda componente también representó el mismo factor latente en ambos casos, con algunas diferencias muy pequeñas en las cargas de las componentes. Valores altos de esta segunda componente indicaron suelos con altos valores de Actividades Anuales de Doble Ciclo, Fe, P, K, con baja pendiente y elevación, bajo contenido de Pastizal Natural, Arbustales y Matorrales, y Materia Orgánica. En la tercera componente se apreciaron cambios considerables en las cargas entre los métodos. En general, se pudo observar que valores altos de la tercera componente correspondieron a suelos con alto CE, Na, Fe, Pasturas Naturales Manejadas y Zonas Anegables, y bajo porcentaje de Actividades Estivales y bajo Valor Unitario de la Tierra.

La Tabla 3 muestra la varianza, el porcentaje explicado de varianza y el índice de correlación de Moran de cada componente en PCA y en MULTISPATI, la pérdida de inercia y la ganancia en información espacial (ver las fórmulas para el cálculo de la pérdida de inercia y la ganancia espacial en la Sección 3.2.2). No se observaron grandes diferencias entre PCA y MULTISPATI para las primeras dos componentes a nivel de pérdida de inercia o de ganancia en información espacial, pero sí a partir de la tercera componente. En la mayoría de las componentes, el índice de Moran fue mayor en MULTISPATI que en la PCA. Estos resultados sugirieron que MULTISPATI

Tabla 3: Resumen de PCA y MULTISPATI aplicado al Conjunto 1 con 36 variables.

	Varianza		Variabilidad acumulada		I de Moran		P.I.	G.I.E
Componente	PCA	MS-PCA	PCA	MS-PCA	PCA	MS-PCA		
1	8.33	8.33	0.25	0.23	0.99	0.99	0.00	0.00
2	5.35	5.34	0.40	0.38	0.97	0.97	0.19	0.00
3	2.83	2.56	0.48	0.45	0.65	0.90	9.54	38.46
4	2.30	2.04	0.54	0.51	0.41	0.83	11.30	102.44
5	2.25	1.34	0.60	0.55	0.72	0.87	40.44	20.83
6	1.45	1.77	0.64	0.59	0.72	0.55	-22.07	-23.61
7	1.29	1.38	0.67	0.63	0.66	0.63	-6.98	-4.55
8	1.13	1.16	0.70	0.66	0.61	0.61	-2.65	0.00
9	1.08	1.13	0.73	0.70	0.47	0.56	-4.63	19.15
10	1.00	0.80	0.76	0.72	0.47	0.70	20.00	48.94
11	0.97	0.81	0.79	0.74	0.65	0.65	16.49	0.00
12	0.90	0.65	0.81	0.76	0.20	0.70	27.78	250.00
13	0.80	0.81	0.83	0.78	0.30	0.49	-1.25	63.33
14	0.79	0.74	0.85	0.80	0.40	0.43	6.33	7.50
15	0.76	0.50	0.87	0.82	0.55	0.50	34.21	-9.09

P.I. = Pérdida de Inercia. G.I.E. = Ganancia de Información Espacial. Solo se muestran dos decimales para los valores de varianza, variabilidad acumulada e índice de Moran, pero los valores de P.I. y de G.I.E se calcularon con todas las cifras. Ganancias cercanas a cero implican que no hay diferencias en la captación de la información espacial entre PCA y MULTISPATI. Pérdidas cercanas a cero implican que no hay diferencias en la captación de la inercia entre PCA y MULTISPATI.

resumió los datos de distinta manera que PCA y que podría aprovechar mejor la información espacial.

El clustering jerárquico de covariables generó el dendrograma de la Figura 29 en el Anexo. Se trazó una línea vertical a lo largo del dendrograma de manera tal que las 36 variables originales se agruparon en 23 clústeres. De estos 23 clústeres, 17 fueron simples, formados por una única variable, mientras que los seis restantes estuvieron formados por múltiples variables colineales entre sí. Los clústeres de variables colineales se muestran en la Tabla 4.

Al Clúster 1 lo asociamos con la textura del suelo, al Clúster 2 con el contenido de materia orgánica en el suelo y al Clúster 3 con el contenido de fósforo y potasio. Los clústeres 4-6 estuvieron más bien relacionados con características geográficas o socioeconómicas y no con características del suelo. Por ejemplo, el Clúster 4 agrupó variables que tomaron valores altos en la zona de las sierras en el Noroeste del territorio (Pendiente, Elevación y Pastizal natural con rocas), el Clúster 5 agrupó variables relacionadas con la presencia de agua, y el Clúster 6 estuvo

Tabla 4: Clústeres de variables colineales.

Clúster	Covariable	Cargas de las covariables	
		PCA	MS-PCA
1	Arcilla	-0.45	0.45
	Arena	0.45	-0.46
	CC	-0.45	0.45
	CIC	-0.45	0.45
	Mn	-0.43	0.43
2	Materia Orgánica	-0.47	-0.48
	Nt	-0.51	-0.51
	Cu	-0.48	-0.48
	CEextsat	-0.53	-0.53
3	P	-0.71	0.70
	K	-0.71	0.71
4	% Pastizal natural con rocas o suelo desnudo	0.55	-0.52
	Pendiente	0.59	-0.60
	DEM	0.59	-0.61
5	Recurrencia	-0.71	-0.74
	% Cuerpos de agua	-0.71	-0.67
6	% Zona urbana consolidada	0.62	-0.38
	% Zona urbana en proceso de consolidacion	0.47	-0.55
	VUT	0.63	-0.74

Las cargas de las covariables corresponden a la primera componente.

asociado a la presencia de población humana.

La Tabla 5 muestra los resultados de aplicar PCA y MULTISPATI a cada clúster. Todos los clústeres pudieron ser resumidos con su primera componente principal, tanto en PCA como en MULTISPATI. La variabilidad explicada por la primera componente principal de PCA estuvo entre el 77 % para el Clúster 3 (P y K) y el 93 % para el Clúster 5 (presencia de agua). La variabilidad explicada por la primera componente principal de MULTISPATI estuvo entre el 74 % para el Clúster 6 (presencia de población) y el 94 % para el Clúster 5.

En los Clústers 1 a 5 no se observaron diferencias entre PCA y MULTISPATI a nivel cargas de las covariables, pérdida de inercia ni ganancia en información espacial. Solo en el Clúster 6 (presencia de población) se observaron algunas diferencias entre PCA y MULTISPATI. En este clúster se vio una pérdida de inercia del 5.5 %, acompañada de una ganancia en información espacial del 13.3 % a favor de MULTISPATI. Además, las cargas de las covariables difirieron en magnitud: en PCA la Zona Urbana Consolidada y el VUT tuvieron la misma carga (0.62) y fueron las más importantes, mientras que en MULTISPATI la más importante fue VUT con

Tabla 5: Resumen de la primera componente obtenida con PCA y MULTISPATI aplicado a cada clúster.

Clúster	Varianza		Variabilidad explicada [%]		I de Moran		P.I.	G.I.E
	PCA	MS-PCA	PCA	MS-PCA	PCA	MS-PCA		
1	4.59	4.59	0.92	0.92	1	1	0	0
2	3.2	3.2	0.8	0.81	0.99	0.99	0	0
3	1.55	1.55	0.77	0.78	0.99	0.99	0	0
4	2.38	2.38	0.81	0.79	0.96	0.96	0	0.21
5	1.89	1.88	0.93	0.94	0.33	0.33	0.53	0.31
6	2.36	2.23	0.79	0.74	0.33	0.37	5.51	13.33

Todos los valores corresponden a la primera componente principal.

0.64, mientras que la Zona Urbana Consolidada se redujo a 0.38.

Concluimos que, con excepción del Clúster 6, donde MULTISPATI aparentó captar mejor la información espacial, los clústeres de variables colineales podrían haber sido resumidos con PCA o MULTISPATI con resultados equivalentes.

4.2.2. Comparación de PCA y MULTISPATI post-modelado

Primero se analizaron los coeficientes de regresión para los conjuntos de datos "Control", "PCA-hclust" y "MULTISPATI-hclust" dentro del mismo tipo de modelo. Solo se consideraron las 17 covariables que estuvieron presentes en los tres conjuntos, es decir, aquellas que no habían sido incluidas en ningún grupo de variables colineales. La Figura 18 muestra en sus filas estas 17 variables, junto su correlación lineal de a pares con el resto de las variables, medida con el coeficiente de correlación de Pearson.

La Figura 19 muestra los coeficientes de regresión para el modelo MLR entrenado con conjuntos "Control", "PCA-hclust" y "MULTISPATI-hclust". En MLR se observó que el ancho del intervalo de credibilidad se redujo para tanto en los conjuntos "PCA-hclust" y "MULTISPATI-hclust" respecto al conjunto "Control", y que MULTISPATI logró intervalos de credibilidad incluso más angostos que PCA.

También se observaron cambios en la magnitud y el sentido de la media a posteriori de los coeficientes. En algunos casos la magnitud del efecto aumentó cuando se redujo la multicolinealidad en los datos, en otros disminuyó, y en algunas pocas no hubo cambios apreciables. A nivel general, MULTISPATI mostró mayores diferencias con la media estimada con los datos originales que

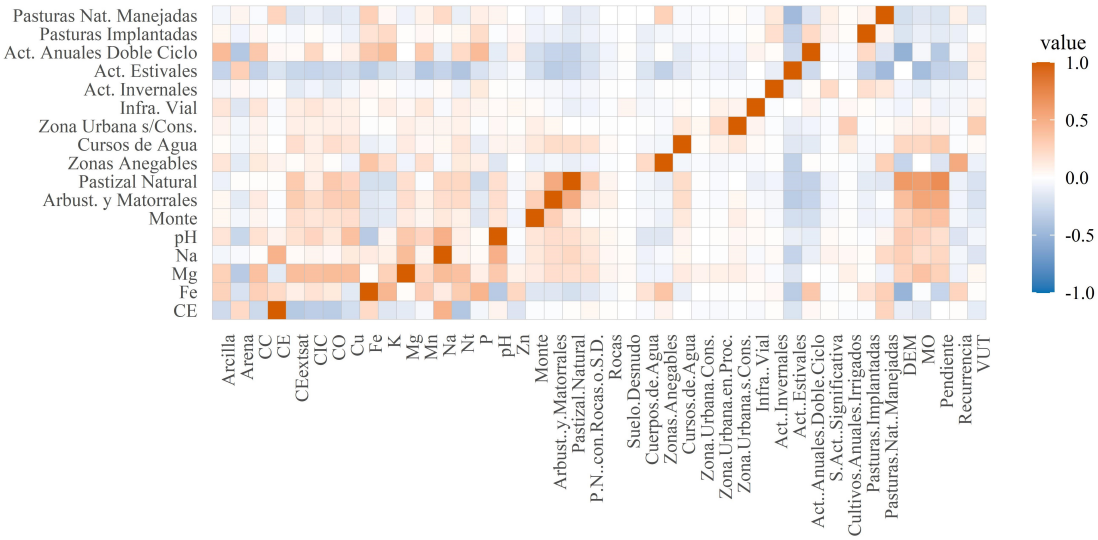

Figura 18: Correlación lineal de a pares para las 17 variables que no participaron de los grupos colineales.

PCA, aunque no podemos determinar si esto fue una ventaja o una desventaja de MULTISPATI porque no conocemos el valor verdadero de los coeficientes de regresión.

En algunas covariables se observaron cambios en el sentido del efecto. En la mayoría de las covariables el sentido de cambio fue el mismo para PCA y MULTISPATI: si un coeficiente resultó más positivo (o negativo) al eliminar la multicolinealidad, resultó más positivo (o negativo) tanto con MULTISPATI como con PCA.

Los resultados observados en el MLR concuerdan con la bibliografía sobre los efectos de la multicolinealidad en los modelos de regresión lineal frecuentista (ver, por ejemplo, Belsley et al., 1980), que se trasladaron a nuestro modelo de regresión lineal bayesiano porque fue estimado con distribuciones a priori vagas. En el modelo con efectos aleatorios *iid* (ver Figura 27 en el Anexo A) se observaron prácticamente los mismos coeficientes que en el MLR y por lo tanto los mismos efectos ante la multicolinealidad.

En el modelo *iCAR* (Figura 20) se observaron intervalos de credibilidad más anchos con el conjunto "Control", siendo "MULTISPATI-hclust" el conjunto que generó el intervalo de menor amplitud en la mayoría de las covariables. También se vieron algunos cambios en la magnitud y el sentido de la media de los coeficientes al variar el tratamiento de la multicolinealidad, aunque en muchas covariables las medias de los conjuntos "Control" y "MULTISPATI-hclust" fueron

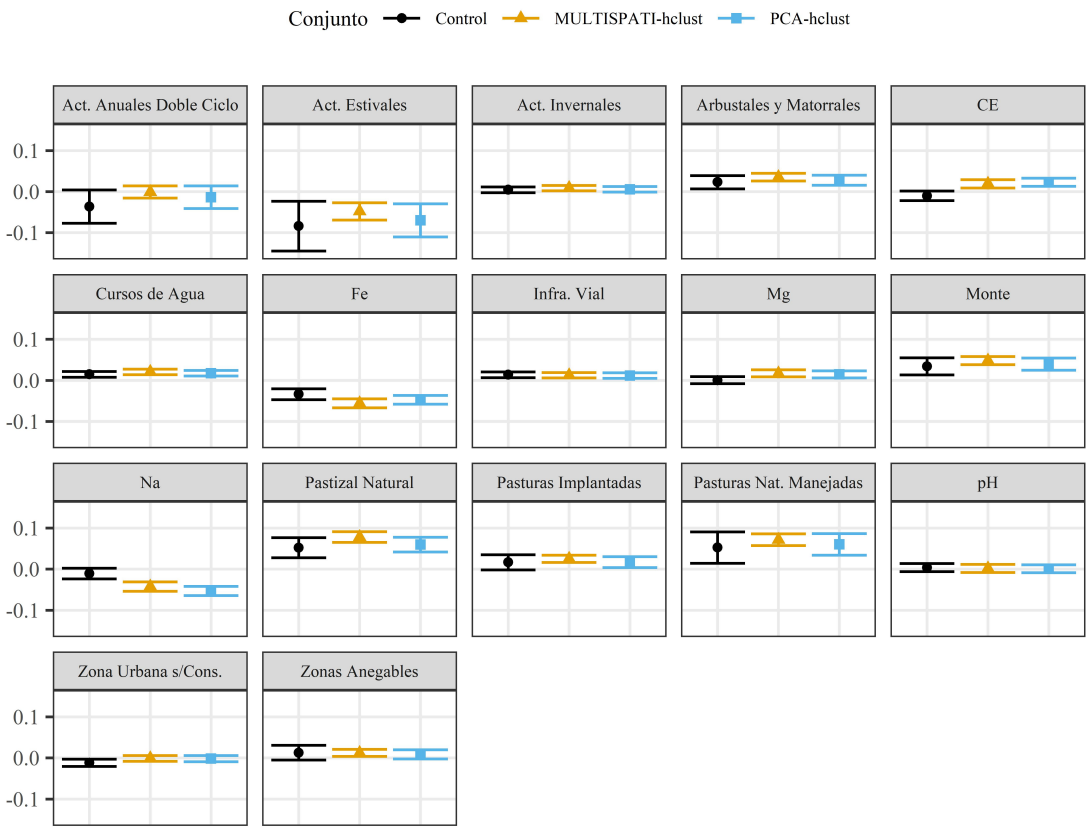

Figura 19: Intervalos de credibilidad del 95 % para los coeficientes de regresión del modelo MLR ajustado con los conjuntos "Control", "PCA-hclust" y "MULTISPATI-hclust". El punto central de cada intervalo representa la media a posteriori del coeficiente.

similares. Esto difirió del comportamiento observado en los modelos *iid* y MLR, donde las medias con "MULTISPATI-hclust" fueron las más diferentes a las medias con el conjunto "Control". La media a posteriori en "PCA-hclust" resultó en muchos casos inferior a la media a posteriori estimada con "MULTISPATI-hclust", pero sin conocer el verdadero valor de los coeficientes no nos animamos a concluir sobre el significado de este comportamiento. Para el modelo BYM2 (ver Figura 28 en el Anexo) se observaron coeficientes y comportamientos similares a los obtenidos con el modelo iCAR.

En términos de los coeficientes de regresión, pudimos concluir que en todos los modelos ajustados, espaciales o no, se observaron los mismos efectos negativos atribuidos a la presencia de multicolinealidad, i. e., menor precisión en los coeficientes y variaciones en su media a posteriori.

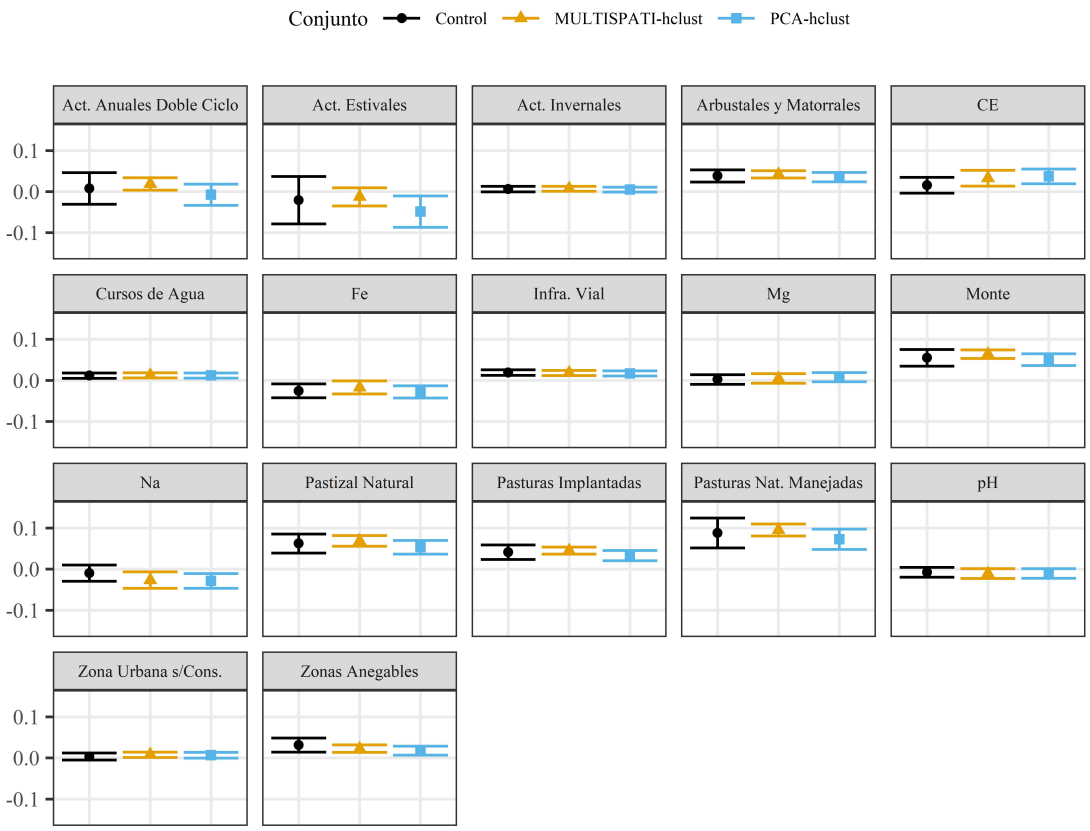

Figura 20: Intervalos de credibilidad del 95 % para los coeficientes de regresión del modelo $iCAR$ ajustado con los conjuntos "Control", "PCA-hclust" y "MULTISPATI-hclust". El punto central de cada intervalo representa la media a posteriori del coeficiente.

Además, en la mayoría de las covariables fue el conjunto "MULTISPATI-hclust" el que generó el intervalo de menor amplitud, aunque no pudimos determinar si los intervalos estuvieron centrados alrededor del verdadero valor de los coeficientes porque no conocemos el verdadero valor de los mismos. Una opción para resolver esta duda podría ser repetir este estudio con datos simulados.

El efecto de aplicar PCA o MULTISPATI en los coeficientes de regresión de los modelos espaciales fue distinto a su efecto observado en los modelos no espaciales. Por ejemplo, los conjuntos "Control" y "MULTISPATI-hclust" generaron coeficientes con medias a posteriori similares en los modelos espaciales, pero diferentes en los modelos no espaciales. Además, en los modelos con efectos espaciales la media a posteriori estimada con "PCA-hclust" fue sistemáticamente menor

que la media estimada con los conjuntos "Control" y "MULTISPATI-hclust". Si bien suponemos que estas diferencias entre PCA y MULTISPATI tuvieron origen en la combinación entre el tratamiento explícito de la espacialidad que hacen los modelos $iCAR$ y BYM2 y el tratamiento explícito de la espacialidad que hace MULTISPATI, no podemos concluir con certeza porque no conocemos el verdadero valor de los coeficientes de regresión.

Para comprender si la capacidad predictiva varió al entrenar los modelos con los distintos conjuntos de datos, se comparó la performance de los modelos a través de las medidas presentadas en la Tabla 6.

Tabla 6: Medidas de comparación de la capacidad predictiva de los 20 modelos finales.

Tipo de Modelo	Multicolinealidad	DIC	WAIC	LPML	RMSE_E	RMSE_T	MAE_E	MAE_T
iCAR	multispati−hclust	-2569	-2442	1195	0.172	0.222	0.129	0.163
iCAR	pca−hclust	-2483	-2378	1173	0.177	0.219	0.134	0.163
iCAR	multispati80	-1837	-1730	850	0.188	0.233	0.139	0.171
iCAR	pca80	-1932	-1812	884	0.183	0.234	0.135	0.172
iCAR	control	-2483	-2380	1175	0.187	0.227	0.142	0.169
bym2	multispati−hclust	-3559	-3437	1471	0.118	0.203	0.09	0.154
bym2	pca−hclust	-3484	-3351	1468	0.124	0.201	0.094	0.152
bym2	multispati80	-2738	-2589	1119	0.137	0.213	0.102	0.161
bym2	pca80	-2846	-2716	1139	0.132	0.214	0.099	0.162
bym2	control	-3536	-3394	1488	0.136	0.218	0.102	0.16
iid	multispati−hclust	-1413	-1384	445	0.144	0.242	0.107	0.181
iid	pca−hclust	-1536	-1510	496	0.15	0.238	0.112	0.181
iid	multispati80	-222	-228	101	0.227	0.253	0.168	0.19
iid	pca80	-21	-24	2	0.238	0.256	0.18	0.194
iid	control	-1363	-1331	649	0.186	0.231	0.14	0.174
mlr	multispati−hclust	-908	-890	445	0.218	0.242	0.164	0.181
mlr	pca−hclust	-1009	-992	496	0.215	0.238	0.164	0.181
mlr	multispati80	-206	-201	100	0.235	0.253	0.176	0.19
mlr	pca80	-9	-3	1	0.241	0.256	0.181	0.194
mlr	control	-1328	-1299	648	0.207	0.231	0.158	0.174

En la regresión lineal múltiple el menor error de predicción de testeo se obtuvo entrenando al modelo con el conjunto de datos "Control", es decir, aquel con las 36 covariables y multicolinealidad originales. Los conjuntos "PCA-hclust" y "MULTISPATI-hclust" generaron $RMSE_T$ y MAE_T similares, pero ligeramente peores que los datos originales. Los conjuntos "PCA80" y "MULTISPATI80" generaron peores resultados que cualquier otro método de reducción de

multicolinealidad. Suponemos que esta diferencia en contra de "PCA80" y "MULTISPATI80" se debió a que si bien la multicolinealidad desapareció por completo en estos conjuntos, también se perdió parte de la información, porque no se incluyó el 20 % de la información que quedó en las últimas componentes, o porque no se incorporaron términos polinómicos de las componentes como en los otros conjuntos, ya que no se detectaron tendencias no lineales entre las componentes y la respuesta.

En el modelo *iid* se observó la misma situación que en el MLR: el modelo con menores errores MAE_T y $RMSE_T$ fue el modelo generado con las variables originales, los segundos mejores modelos en términos de MAE_T y $RMSE_T$ fueron los entrenados con "PCA-hclust" y "MULTISPATI-hclust", y los modelos generados con "PCA80" y "MULTISPATI80" tuvieron los mayores errores de testeo.

En el modelo $iCAR$ fueron los conjuntos "MULTISPATI-hclust" y "PCA-hclust" los que mostraron los menores MAE_T y $RMSE_T$, aunque la diferencia en capacidad predictiva con el modelo entrenado con los datos originales no fue grande. Los conjuntos "PCA80" y "MULTISPATI80" generaron los modelos con mayor error de predicción de testeo. En el modelo BYM2 ocurrió lo mismo que en el modelo iCAR.

A modo de conclusión de esta sección, podríamos decir que en los modelos no espaciales MLR e *iid* no observamos mejoras en la capacidad predictiva al entrenarlos con datos donde se había reducido la multicolinealidad, respecto a la capacidad predictiva de los modelos entrenados con el conjunto "Control", pero que en los modelos espaciales BYM2 e $iCAR$ sí observamos que el error de predicción de testeo fue menor al entrenar los modelos con los conjuntos "MULTISPATI-hclust" y "PCA-hclust", respecto a entrenarlos con los datos originales.

De las técnicas analizadas para tratar la multicolinealidad, las técnicas que se basaron en identificar grupos colineales de variables y reducirlos luego mediante MULTISPATI o PCA a una única covariable sintética lograron reducir el error de predicción en los modelos espaciales, logrando PCA ligeramente menor error que MULTISPATI. Por el contrario, entrenar los modelos con los conjuntos de datos obtenidos reemplazando las covariables originales por componentes principales que representan el 80 % de la información original aumentó los errores de predicción $RMSE_T$ y MAE_T respecto a en todos los modelos. Cabe aclarar que cuando se habla de reducción o aumento del error de predicción, se habla de reducciones o aumentos relativos al error obtenido en el mismo modelo pero entrenado con los datos "Control".

4.3. Elección del mejor modelo predictivo

A partir del análisis de impacto de la multicolinealidad se decidió que el mejor modelo predictivo debe elegirse entre las opciones generadas con los conjuntos "MULTISPATI-hclust" y "PCA-hclust". A partir de este punto, se asume que todos los modelos que se mencionan han sido generados con estos conjuntos, a menos que se indique lo contrario.

Los criterios de información DIC y WAIC en la Tabla 6 indicaron que los modelos espaciales tuvieron mejor ajuste que los modelos no espaciales y que, entre ellos, el BYM2 mostró menores DIC y WAIC que el $iCAR$. A pesar de la corrección que incorporan por ser estimados con datos de entrenamiento (ver Sección 2.6.2), DIC y WAIC suelen favorecer al modelo que más sobreajusta. Este hecho fue expresado en la bibliografía (ver por ejemplo Beguin et al., 2012) y también fue verificado en este trabajo. Por este motivo no basamos la elección del mejor modelo predictivo únicamente en DIC o WAIC.

La medida de la verosimilitud LPML obtenida con validación cruzada no espacial también favoreció a los modelos espaciales y en especial al BYM2. LPML se consideró un poco más confiable que DIC y WAIC para evaluar el ajuste (ver Sección 2.6.2). Por ejemplo, el modelo iid entrenado con "MULTISPATI-hclust" obtuvo el mismo LPML que el MLR entrenado con el conjunto "Control.ª pesar de haber obtenido menores WAIC y DIC, indicando que no hubo mejoras en el ajuste al usar efectos aleatorios iid. Esta conclusión concordó con los valores de $RMSE$ y MAE de testeo obtenidos.

Los $RMSE_T$ y MAE_T obtenidos con validación cruzada espacial también indicaron que los mejores modelos fueron $iCAR$ y BYM2, y que incluso BYM2 logró menor error que $iCAR$. La mayor disminución del error promedio de predicción calculado con datos de testeo la logró el modelo BYM2 con "PCA-hclust", que obtuvo reducciones $\Delta_{RMSE_T} = 0.03$ y $\Delta_{MAE_T} = 0.02$ respecto al modelo base MLR entrenado con los datos originales. El $iCAR$ entrenado con "PCA-hclust" logró $\Delta_{RMSE_T} = 0.012$ y $\Delta_{MAE_T} = 0.011$ respectivamente.

La Tabla 7 muestra la correlación r entre predichos y observaciones calculada por validación cruzada espacial para datos de entrenamiento y testeo. Los modelos espaciales mostraron un aumento aproximado del 5 % en r_T respecto a los modelos no espaciales. Los modelos BYM2 mostraron las mayores correlaciones tanto en entrenamiento como en testeo, pero la diferencia

Tabla 7: Algunas medidas resumen de validación de los 20 modelos finales.

Tipo de Modelo	Multicolinealidad	r_E	r_T	n/p_{WAIC}	p-value$_{Moran}$
$iCAR$	multispati_hclust	0.806	0.617	7.58	0.001
$iCAR$	pca_hclust	0.791	0.624	9.26	0.001
$iCAR$	multispati80	0.762	0.57	9.44	0.001
$iCAR$	pca80	0.781	0.561	8.06	0.001
$iCAR$	ss	0.761	0.595	9.72	0.001
bym2	multispati_hclust	0.92	0.63	3.36	0.999
bym2	pca_hclust	0.911	0.636	3.55	0.998
bym2	multispati80	0.892	0.582	3.74	0.999
bym2	pca80	0.902	0.573	3.48	0.999
bym2	ss	0.885	0.632	3.54	0.999
iid	multispati_hclust	0.878	0.554	3.03	0.001
iid	pca_hclust	0.861	0.565	3.03	0.001
iid	multispati80	0.616	0.494	21.48	0.001
iid	pca80	0.556	0.456	23.61	0.001
iid	ss	0.765	0.576	10.92	0.001
mlr	multispati_hclust	0.652	0.554	103.09	0.001
mlr	pca_hclust	0.661	0.565	104.54	0.001
mlr	multispati80	0.572	0.494	214.71	0.001
mlr	pca80	0.545	0.456	223.69	0.001
mlr	ss	0.691	0.576	69.24	0.001

entre r_E y r_T fue cercana al 28 %, lo que podría indicar sobreajuste.

El cociente datos-parámetros n/p_{WAIC} resultó cercano a 100 en el MLR, fluctuó entre 3.03 y 23.61 para el iid, fue cercano a 10 en todos los modelos $iCAR$ y resultó menor a 4 en todos los BYM2. Si decidimos que los modelos BYM2 e $iCAR$ ajustados son válidos, entonces debemos concluir que la regla práctica $n/p_{WAIC} \geq 20$ utilizada en el análisis de sensibilidad quizá es demasiado estricta en el contexto de modelos jerárquicos bayesianos con efectos aleatorios espaciales, o al menos lo es cuando el número de parámetros se calcula con el WAIC.

Con base en las medidas de comparación y validación analizadas previamente decidimos continuar solo con los modelos entrenados con el conjunto "PCA-hclust". Esta decisión se tomó porque no pudimos verificar que los coeficientes de regresión estimados con los modelos entrenados con "MULTISPATI-hclust" hayan estado más cerca de la verdadera distribución del coeficiente de regresión que aquellos entrenados con "PCA-hclust", ya que no trabajamos con datos simulados, pero sí pudimos observar que los modelos entrenados con "PCA-hclust" obtuvieron en todos los casos igual o menor error de predicción de testeo que los modelos entrenados con

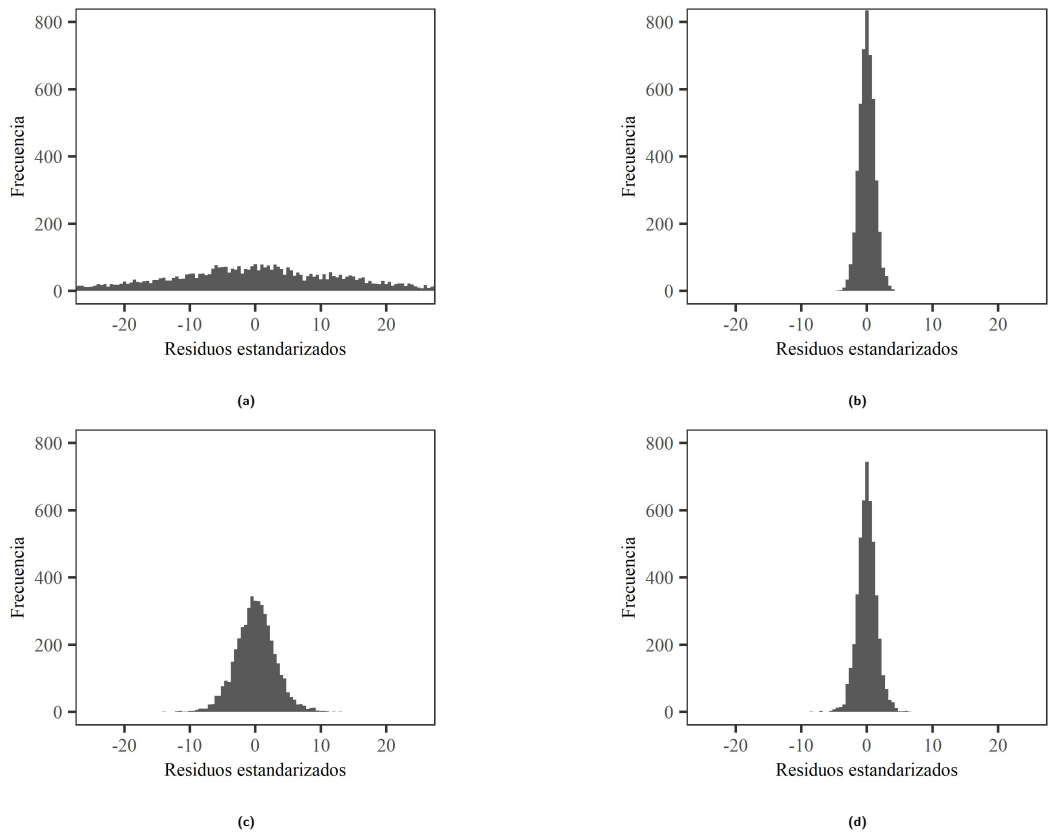

Figura 21: Histogramas de residuos. (a) MLR. (b) *iid*. (c) iCAR. (d) BYM2.

"MULTISPATI-hclust". En resumen, concluimos que en este caso la complejidad diferencial que agregó MULTISPATI respecto a PCA en el pre-procesamiento de los datos no se vio justificada por una mejora en la eficacia estadística del modelo.

Habiendo decidido únicamente buscar el mejor modelo entre los cuatro modelos finales entrenados con "PCA-hclust", procedimos a realizar un chequeo gráfico de ellos.

No se observaron desviaciones graves de la normalidad en la distribución de los residuos puntuales estandarizados al menos los modelos espaciales (Figura 21). Los residuos puntuales estandarizados se calcularon como la diferencia entre la moda a posteriori del predicho y la observación inicial del ESPI, dividido por el desvío estándar de la distribución a posteriori del predicho.

La Figura 22 muestra mapas de la moda a posteriori de los residuos para los cuatro modelos obtenidos con "PCA-hclust". En los modelos MLR e *iid* la autocorrelación espacial residual fue

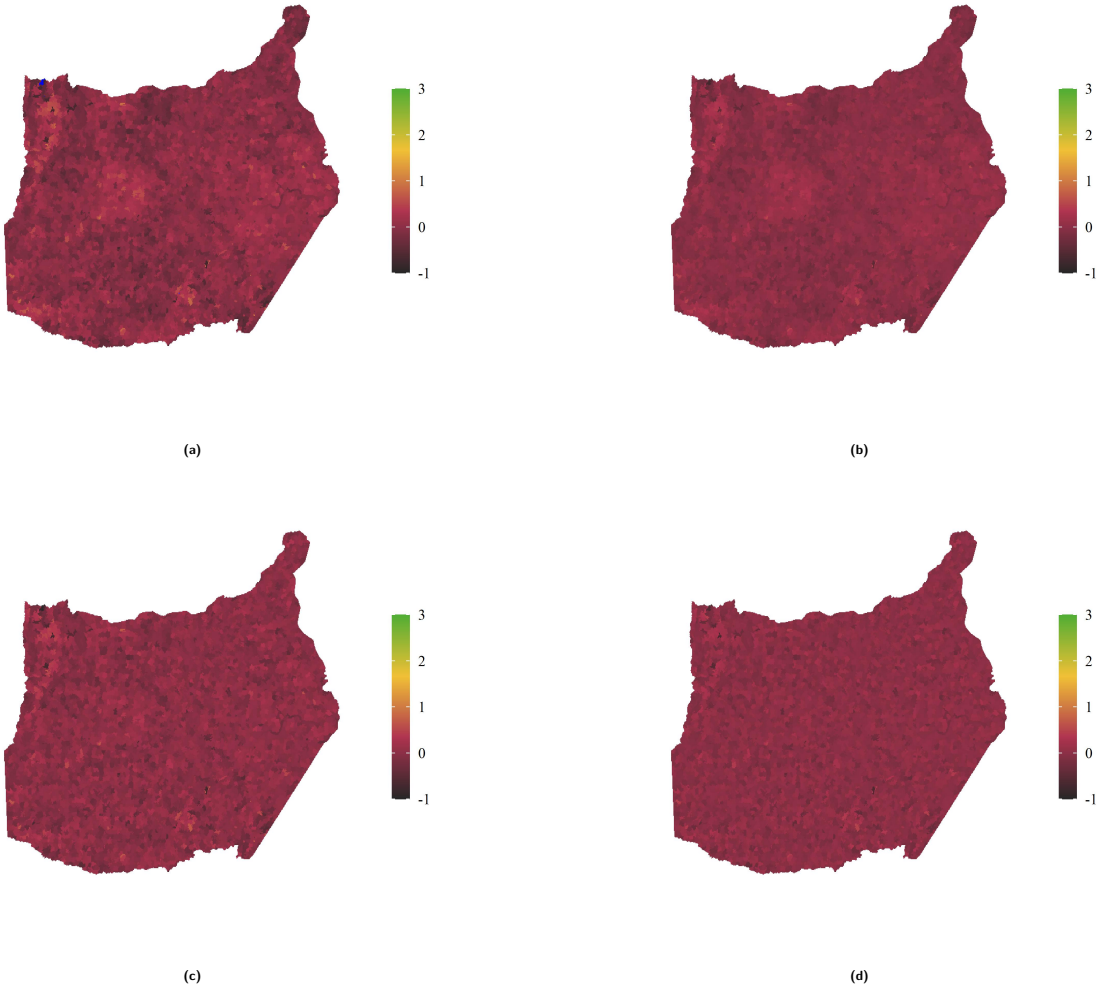

Figura 22: Mapas de la moda a posteriori de los residuos para los modelos entrenados con "PCA-hclust". (a) MLR. (b) *iid*. (c) iCAR. (d) BYM2. Las colores representan el valor de la moda del residuo. Se puede apreciar que la autocorrelación residual disminuyó respecto al modelo MLR (a) cuando se agregaron efectos aleatorios no estructurados (b), que prácticamente desapareció cuando se agregaron efectos espaciales $iCAR$ (c) y que desapareció totalmente cuando se agregaron efectos espaciales BYM (d).

muy visible, especialmente en las sierras de Córdoba y en el centro del territorio. En el modelo espacial $iCAR$ casi no se apreció autocorrelación. En el modelo BYM2 no se observó visualmente autocorrelación, apreciación que comprobamos estadísticamente con un test de Moran. El modelo BYM2 fue el único modelo donde el test de Moran no rechazó la hipótesis nula de ausencia de autocorrelación espacial (Ver 7).

La Figura 23 muestra mapas donde cada region i se ha coloreado de acuerdo al CPO_i obtenido. Las regiones pintadas en negro son regiones que obtuvieron CPO_i cercanos a cero, donde los CPO_s cercanos a cero indican que es baja la verosimilitud de obtener los valores de ESPI observados si el modelo fuese cierto. Los modelos no espaciales mostraron gran cantidad de regiones con bajo CPO. Los modelos espaciales mostraron menor cantidad de bajos CPOs, indicando que estos modelos fueron menos discrepantes con los datos que los modelos espaciales. La excepción fueron las sierras del Noroeste del territorio, que en todos los casos - espaciales y no espaciales - mostraron valores de CPO muy bajos.

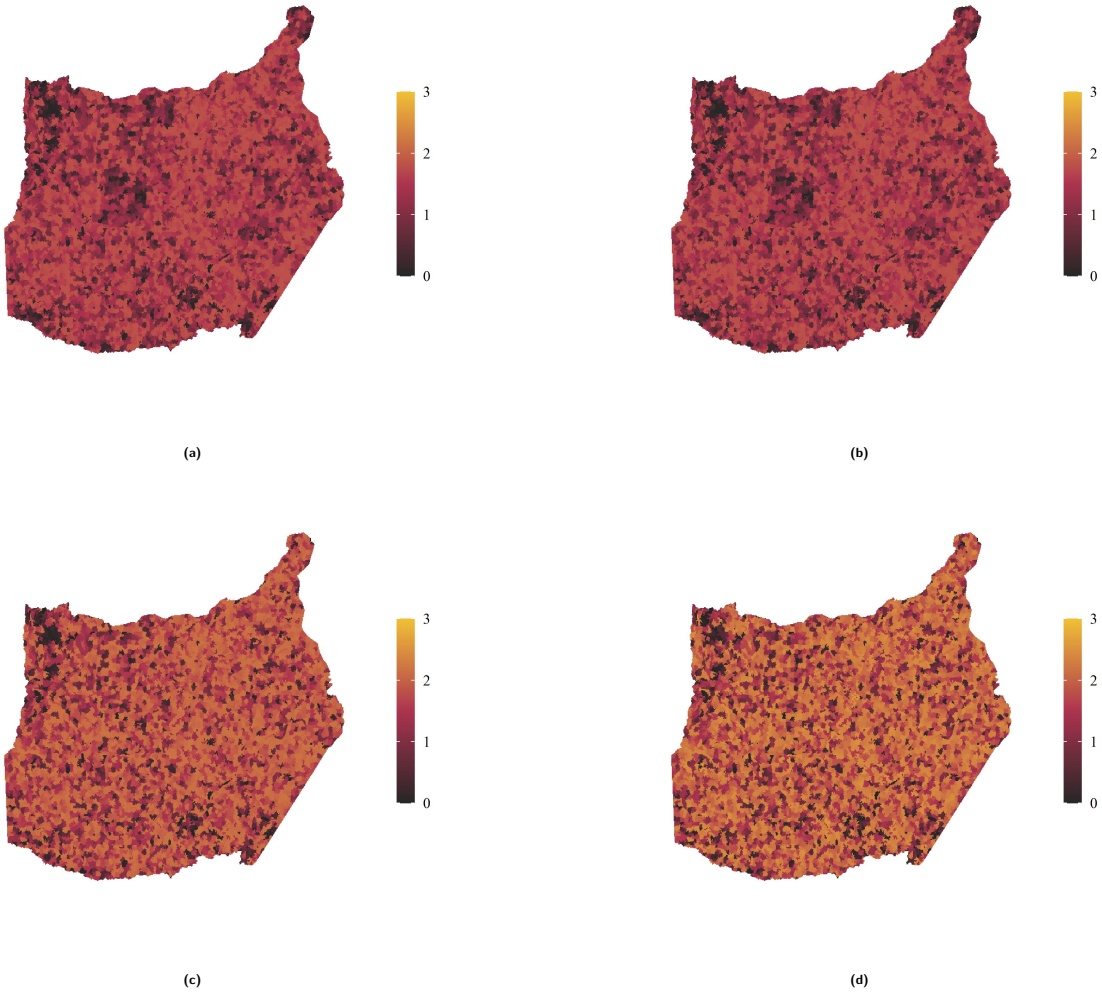

Figura 23: Mapas de CPO. (a) MLR. (b) *iid*. (c) iCAR. (d) BYM2. Los modelos espaciales (c) y (d) mostraron menor cantidad de regiones con bajo CPO, indicando mayor verosimilitud de obtener los datos dados bajo esos modelos.

El hecho de que los modelos espaciales y en especial el BYM2, que pareció ajustarse adecuadamente a los datos, no haya sido capaz de generar verosimilitudes altas para las regiones de las sierras de Córdoba, nos hizo considerar que puede haber diferencias estructurales entre la zona Noroeste de la cuenca y el resto del territorio, ya que se trata de un territorio muy amplio a escala de paisaje que contiene elementos muy distintos, como valles, ríos, montañas y poblaciones humanas. En futuros trabajos podría evaluarse la opción de ajustar un modelo para la zona serrana y otro para el valle, o también la inclusión de nuevas covariables que le proporcionen al modelo información que le falta.

Los histogramas de los PIT_i no llegaron a ser totalmente uniformes en ningún modelo (Figura 24). En todos los modelos entrenados con "PCA-hclust"se observó que los PIT_i cercanos a 0 y a 1 tuvieron menor frecuencia, lo que podría indicar algún problema de ajuste en las colas de la respuesta y coincide con lo observado en los mapas de CPO. En el modelo BYM2 los PIT se vieron algo mas uniformes que en el resto de los modelos, aunque aqui nos parecio adecuado recordar que la uniformidad del histograma no valida automáticamente un modelo (Gelman et al., 2014).

La Figura 25 muestra los gráficos de residuos puntuales versus predichos puntuales, donde ambos se calcularon utilizando la moda a posteriori de la distribución de los predichos. Coloreando la nube de puntos de acuerdo al valor de la respuesta ESPI, donde verde señaliza ESPI observado altos, y rojo simboliza ESPI observado bajo, se vio que en el MLR fue donde más probable resultó encontrar residuos positivos para ESPI altos y residuos negativos para ESPI bajos. En otras palabras, el MLR predijo de menos para los ESPI altos, y predijo de más en las regiones con ESPI bajos. Este mal comportamiento en las colas mejoró al agregar efectos aleatorios al modelo, estructurados o no. En el modelo espacial BYM2 se observó que los residuos decrecieron en magnitud y se ubicaron más cerca del cero.

Finalmente, la Figura 26 compara los mapas de las observaciones del ESPI con los mapas de la moda a posteriori de las predicciones del ESPI generadas con los modelos ajustados con "PCA-hclust". El MLR falló en predecir los ESPI bajos y moderadamente bajos, simbolizados con los colores rojos y naranjas en la Figura 26a, y también falló en predecir los ESPIs más altos de las sierras de Córdoba, simbolizados con color verde intenso. El modelo iid se observó demasiado poco suavizado, especialmente en los ESPIs moderadamente bajos. El modelo $iCAR$ mostró suavizado espacial aceptable. El modelo BYM2 logró captar gran detalle de la respuesta original,

incluso las diferencias en la zona de las sierras.

Con base en lo observado elegimos el modelo BYM2 entrenado con "PCA-hclust" como el mejor modelo predictivo, ya además de mostrar el menor error de predicción de testeo y mayor LPML, fue el único modelo capaz de eliminar la autocorrelación espacial, inclusive en la zona de las sierras. El único gráfico de validación que señaló algún problema de este modelo fue el mapa de CPO, que mostró CPO_is bajos en la zona de las sierras, aunque no consideramos que estos CPOs ameriten descartar el modelo.

Los coeficientes de regresión del modelo BYM2 se pueden ver en la Tabla 8, en orden descendiente de magnitud. Los coeficientes que se consideraron significativos se marcaron con un asterisco. Las covariables sin asterisco tuvieron intervalos de credibilidad del 95 % que abarcaron el 0, o la magnitud de su efecto se consideró demasiado baja comparada respecto a las otras covariables. Los coeficientes se pueden comparar entre covariables porque todas las variables fueron estandarizadas antes de ingresar al modelo.

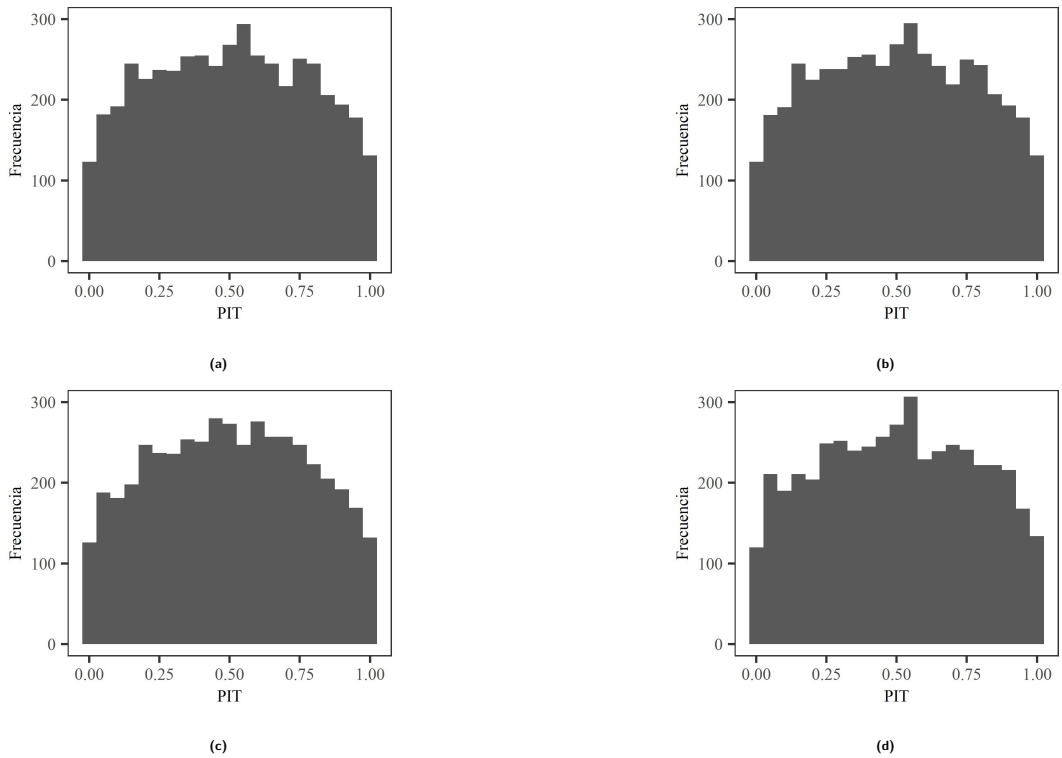

Figura 24: Histogramas PIT. (a) MLR. (b) *iid*. (c) iCAR. (d) BYM2.

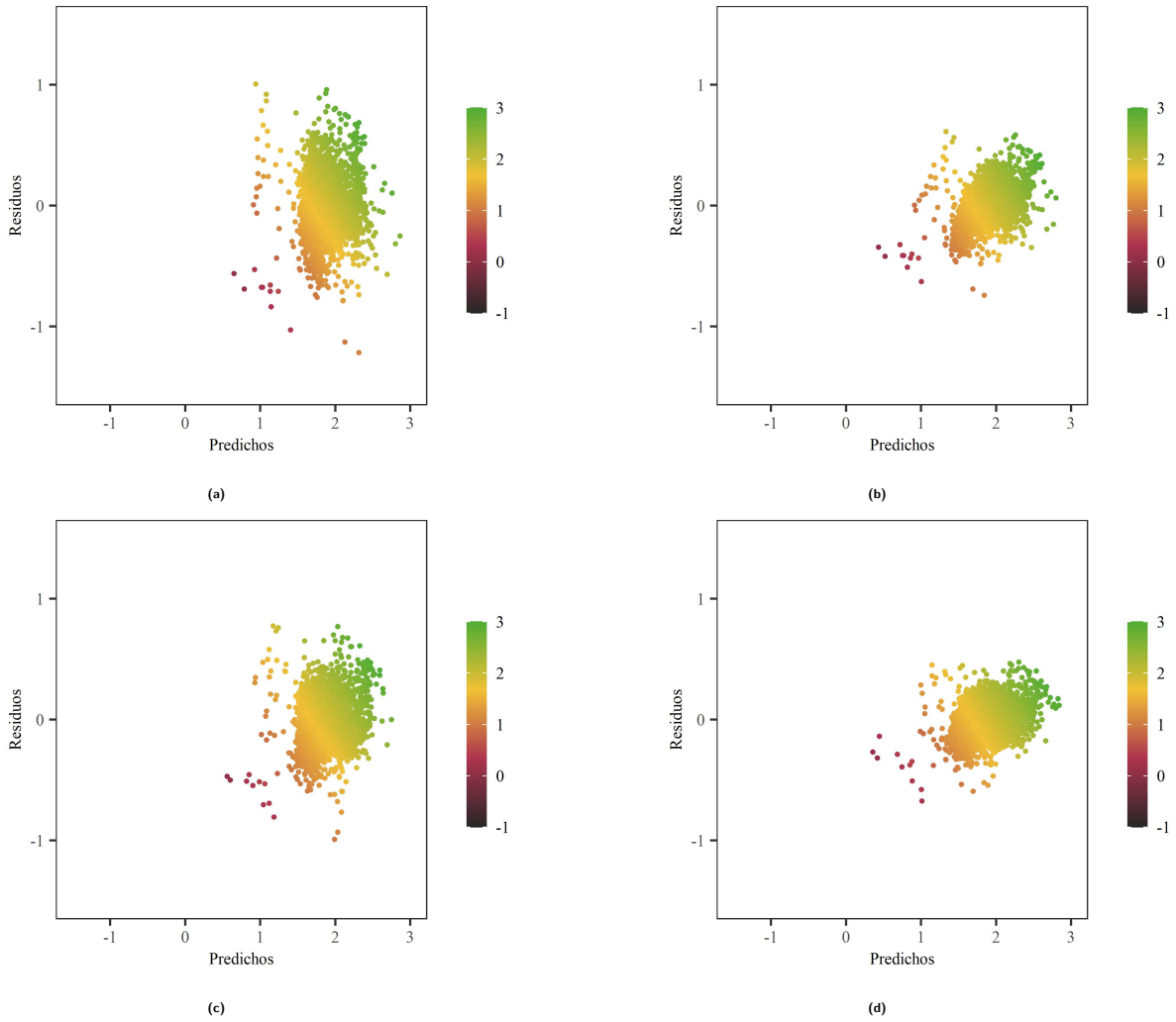

Figura 25: Residuos puntuales versus moda a posteriori de los predichos para los modelos ajustados con "PCA-hclust". (a) MLR. (b) *iid*. (c) iCAR. (d) BYM2.

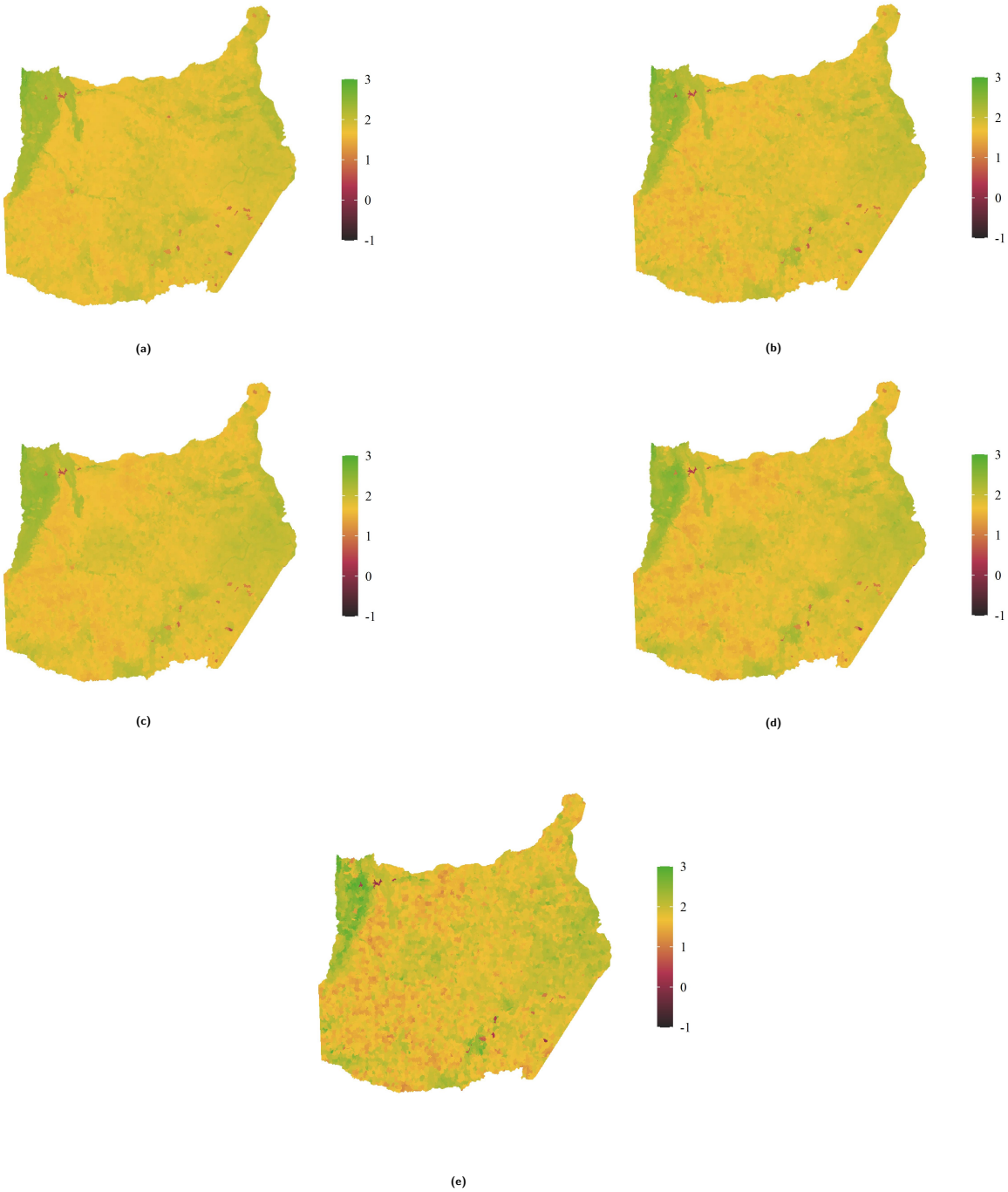

Figura 26: Valores predichos del ESPI para los modelos ajustados con "PCA-hclust" y su comparación con el ESPI observado. (a) MLR. (b) *iid*. (c) iCAR. (d) BYM2. (e) Observaciones originales el ESPI.

Tabla 8: Coeficientes de regresión del modelo BYM2, elegido como el mejor modelo predictivo.

covariable	$\hat{\beta}_j(IC_{95})$	covariable	$\hat{\beta}_j(IC_{95})$
Intercepto*	1.841 (1.837, 1.845)	Cluster5 (Ausencia de Agua)	-0.019 (-0.04, 0.003)
Cluster5^2 (Ausencia de Agua)*	-0.083 (-0.098, -0.068)	% Infraestructura vial*	0.017 (0.011, 0.023)
% Pasturas Naturales Manejadas*	0.079 (0.054, 0.104)	Media Na	-0.016 (-0.039, 0.008)
% Monte*	0.057 (0.043, 0.072)	Cluster2 (Materia Orgánica)	-0.015 (-0.043, 0.014)
% Pastizal Natural*	0.054 (0.037, 0.071)	Media pH	-0.012 (-0.026, 0.003)
% Pasturas Implantadas*	0.039 (0.026, 0.051)	% Cursos de Agua	0.010 (0.004, 0.015)
Media CE*	0.038 (0.016, 0.059)	% Zona Urbana sin Consolidar	0.010 (0.003, 0.017)
% Arbustales y Matorrales*	0.038 (0.027, 0.050)	Cluster3 (P-K)	-0.010 (-0.031, 0.011)
Cluster4^2 (Terreno elevado)*	0.037 (0.000, 0.074)	Cluster1 (Textura del Suelo)	-0.007 (-0.040, 0.026)
% Actividades Estivales*	-0.034 (-0.073, 0.005)	% Actividades Invernales	0.006 (0.000, 0.012)
Cluster4 (Terreno elevado)	0.027 (-0.028, 0.083)	Media Mg	0.005 (-0.010, 0.019)
Media Fe*	-0.023 (-0.042, -0.005)	Cluster6 (Población Humana)	0.004 (-0.014, 0.023)
Cluster6^2 (Población Humana)*	-0.020 (-0.035, -0.004)	% Act. Anuales de Doble Ciclo	0.000 (-0.026, 0.026)
% Zonas Anegables*	0.019 (0.007, 0.030)		

El modelo encontró que el ESPI promedio en el territorio de la cuenca fue de 1.841, con un intervalo de credibilidad muy angosto de entre aproximadamente 1.84 y 1.85.

La Ausencia de Agua fue la covariable más importante según este modelo BYM, con una media estimada del efecto de -0.083. La presencia de agua se asoció con altos valores de provisionamiento de servicios ecosistémicos de forma cuadrática. La segunda covariable con mayor efecto fue el porcentaje de Pasturas Naturales Manejadas en la región, que obtuvo una magnitud de 0.079, similar a la de la Ausencia de Agua. Este efecto puede interpretarse de la siguiente manera: al aumentar el porcentaje de Pasturas Naturales Manejadas en la región, se espera un aumento en el ESPI promedio en la región, descontando el efecto del resto de las covariables.

Las terceras y cuartas covariables más importantes fueron el porcentaje de Monte y de Pastizal Natural; ambas mostraron asociación positiva con el ESPI y sus efectos tuvieron ambos una media a posteriori cercana a 0.55, con similares intervalos de confianza.

El porcentaje de Pasturas Implantadas (alfalfa y pasto), de Arbustales y Matorrales, el Terreno Elevado y la Conductividad Eléctrica tuvieron relaciones positivas con el ESPI de similar magnitud, con una media aproximada de 0.38. El porcentaje de Actividades Estivales en la región, donde las actividades estivales consisten en cultivos anuales o de grano para cosecha o a cultivos anuales para forraje, estuvo relacionado negativamente con el ESPI con una media a posteriori de -0.34, aunque el extremo derecho del intervalo de credibilidad abarcó el cero y por lo tanto la

probabilidad de que el verdadero efecto sea negativo es ligeramente inferior a 0.95.

El contenido de hierro también mostró una relación negativa con el ESPI pero débil, con una media de -0.023. Efectos de similar magnitud pero positivos mostraron las covariables Zonas Anegables y porcentaje de Infraestructura vial, donde las zonas anegables fueron zonas que en el pasado mostraron agua pero ya no, al momento de tomar los datos.

El resto de las covariables o bien tuvieron efectos de magnitud muy baja, o bien tuvieron intervalos de credibilidad cuya media fue muy cercana al cero, por lo que consideramos que estas covariables fueron de escasa importancia para predecir el ESPI.

Entre las covariables poco importantes contamos el porcentaje de Cursos de Agua, de Zona Urbana sin Consolidar y de Actividades invernales, con medias a posteriori cercanas o menores a 0.01.

Entre las covariables cuyo intervalo de credibilidad abarcó al cero contamos al sodio, el Clúster 2 que resumió la información del contenido de nitrógeno y materia orgánica, el pH del suelo, el contenido de fósforo, de potasio y de magnesio, la textura del suelo y el porcentaje de Actividades Anuales de Doble Ciclo.

5. Conclusión y comentarios finales

En esta tesis aplicamos modelos jerárquicos bayesianos para datos regionales con efectos aleatorios *iid*, $iCAR$ y BYM2 para modelar una respuesta continua georreferenciada en función de más de 30 covariables georreferenciadas. La naturaleza multicolineal de las covariables y la flexibilidad que los modelos jerárquicos bayesianos pueden introducir a través del efecto aleatorio hicieron necesario preguntarnos qué procedimientos seguir para realizar un correcto modelado de la respuesta.

En nuestra primera hipótesis, sustentada ampliamente en la literatura, supusimos que las distribuciones a priori de los hiperparámetros de los efectos aleatorios serían cruciales para determinar la tendencia al sobreajuste del modelo. Para evitar elegir distribuciones a priori que fuercen el sobreajuste analizamos la sensibilidad de cada modelo a distintas distribuciones a priori para los hiperparámetros. Además, utilizamos distribuciones a priori penalizadas por complejidad, pensadas especialmente para facilitar al usuario el control del sobreajuste en los modelos jerárquicos bayesianos a través de la modificación de sus parámetros U y α.

Para reconocer los valores de U y α que generasen distribuciones a priori que pudiesen causar sobreajuste examinamos el WAIC, el cociente datos-parámetros y la escala del efecto aleatorio en función de U y α. Para la distribución a priori de la precisión en los modelos *iid*, $iCAR$ y BYM2 encontramos que aumentos de U y α se tradujeron en menores precisiones a posteriori y, por lo tanto, en mayor variabilidad del efecto aleatorio. Para la distribución a priori del parámetro de mezcla del modelo BYM2 encontramos que los valores de α, fijando $U = 0.5$, no tuvieron demasiada influencia en la precisión a posteriori. Por lo tanto, concluimos que en la reparametrización BYM2 del modelo BYM, la distribución a priori de la precisión puede resultar más importante que la distribución del parámetro de mezcla ϕ en la determinación de la flexibilidad del modelo.

Con el modelo $iCAR$ la relación entre la escala del efecto aleatorio y la elección de U y α mostró una tendencia. En el modelo BYM2, por el contrario, cambios pequeños en U o en α parecieron generar cambios abruptos en la escala del efecto aleatorio. Como también observamos este efecto en el *iid*, suponemos que este comportamiento de menor suavizado en el BYM2 puede deberse a la incorporación del efecto aleatorio *iid*, pero también nos preguntamos si puede estar relacionado con la parametrización del BYM2. A futuro, sería interesante replicar el análisis de sensibilidad

para el modelo BYM con su parametrización original y verificar si también la escala del efecto aleatorio en esta parametrización resulta altamente sensible a cambios en las distribuciones a priori de las precisiones.

Durante el análisis de sensibilidad también notamos que el umbral datos-parámetros en los modelos $iCAR$ y BYM2 quedó en la mayoría de los casos por debajo de 20. Esto sugiere que la regla práctica que aparece en la literatura podría ser demasiado estricta para estos modelos jerárquicos bayesianos complejos para datos espaciales, y que no necesariamente un cociente menor a 20 está relacionado con modelos que sobreajustan.

Consideramos que el procedimiento manual de selección de distribuciones a priori penalizadas por complejidad guiado por el WAIC, el cociente de datos-parámetros y la escala del efecto aleatorio nos permitió encontrar distribuciones a priori candidatas adecuadas pero requirió un profundo trabajo computacional y también mucho criterio personal. Además, encontramos que este procedimiento no fue lo suficientemente preciso como para discriminar entre dos distribuciones a priori similares. Por lo tanto, realizamos la elección final de las distribuciones a priori analizando el error de predicción de testeo medido con $RMSE$ y MAE calculados con validación cruzada espacial. No utilizamos WAIC ni DIC para elegir entre dos distribuciones a priori muy similares porque mostraron tendencia a seleccionar la distribución a priori que produjo el modelo más complejo.

En nuestra segunda hipótesis supusimos que la multicolinealidad en las covariables afectaría negativamente el ajuste de los modelos jerárquicos bayesianos. Para probar esta hipótesis entrenamos los modelos con cinco conjuntos alternativos de covariables donde la multicolinealidad se trató con técnicas distintas en cada caso, para luego analizar el efecto de la reducción de la multicolinealidad en los coeficientes de regresión y en la capacidad predictiva.

El primer conjunto de entrenamiento o conjunto "Control" estuvo formado por los datos originales. A los segundos y tercer conjuntos los creamos aplicando PCA y MULTISPATI a la matriz de covariables. No se observaron grandes diferencias entre PCA y MULTISPATI para las primeras dos componentes a nivel de pérdida de inercia o de ganancia en información espacial, pero sí se observaron diferencias a partir de la tercera componente. En la mayoría de las componentes, el índice de Moran fue mayor en MULTISPATI que en PCA. Estos resultados sugirieron que

MULTISPATI aprovechó mejor la información espacial de los datos.

A los terceros y cuartos conjuntos los creamos aplicando PCA y MULTISPATI por separado a cada uno de los seis grupos de variables colineales detectados previamente con un algoritmo de clustering jerárquico. Con excepción del Clúster 6, relacionado con la existencia de poblaciones humanas, donde las componentes MULTISPATI obtuvieron mayores índices de Moran y aparentaron captar mejor la información espacial, se encontró que los clústeres colineales podrían haber sido resumidos con PCA y con MULTISPATI con resultados equivalentes.

Posterior al ajuste, los coeficientes de regresión de todos los modelos, inclusive los de los modelos espaciales, mostraron que los intervalos de credibilidad se hicieron más precisos y que la media a posteriori sufrió cambios de magnitud y de sentido en muchas covariables cuando se eliminó la multicolinealidad de los datos. Estos resultados coincidieron con lo expresado en la bibliografía para modelos de regresión frecuentista, y bayesianos con distribuciones a priori no informativas. Como la bibliografía no es abundante al respecto de la multicolinealidad en modelos jerárquicos bayesianos, consideramos que este trabajo puede servir como evidencia a favor de que la multicolinealidad también produce efectos indeseados en los coeficientes de regresión de los modelos jerárquicos bayesianos para datos espaciales.

El comportamiento de PCA y MULTISPATI en los modelos con efectos espaciales $iCAR$ y BYM2 fue distinto a su comportamiento en los modelos sin efectos espaciales MLR e *iid*. Por ejemplo, los conjuntos "Control" y "MULTISPATI-hclust" generaron medias similares en los modelos espaciales pero diferentes en los modelos no espaciales. Además, en los modelos con efectos espaciales la media estimada con los datos "PCA-hclust" fue sistemáticamente menor que la media estimada con los conjuntos "Control" y "MULTISPATI-hclust". Si bien suponemos que estas diferencias entre PCA y MULTISPATI tuvieron origen en el tratamiento explícito de la espacialidad que hacen los modelos $iCAR$ y BYM2, en el tratamiento explícito de la espacialidad que hace MULTISPATI, o en ambos, no podemos concluir con certeza porque no conocemos el verdadero valor de los coeficientes de regresión. Para investigar más profundamente este cambio de comportamiento entre MULTISPATI y PCA que se observó en los modelos espaciales se podría realizar un nuevo estudio con datos simulados donde se conozca de antemano la distribución de los coeficientes de regresión y se pueda dilucidar cuál de las dos técnicas logró estimaciones más certeras.

A nivel de capacidad predictiva medida con los errores $RMSE$ y MAE de testeo calculados

con validación cruzada espacial encontramos que la reducción de la multicolinealidad en los datos pudo haber mejorado levemente la capacidad predictiva de los modelos con efecto aleatorio espacial $iCAR$ y BYM2. Además, "PCA-hclust" logró mayor reducción del error de testeo que "MULTISPATI-hclust" en todos los modelos entrenados, espaciales o no. Por el contrario, la aplicación de PCA y MULTISPATI a la matriz original de covariables y la posterior retención del componentes hasta sumar 80 % de variabilidad incrementó el error de predicción. Como la bibliografía para modelos de regresión frecuentista o bayesiana con distribuciones a priori no informativas suele asumir que la multicolinealidad no modifica la precisión predictiva, consideramos importante resaltar que (a) nuestros hallazgos se basaron en datos de ilustración, no en datos simulados, y que (b) podríamos haber incurrido en una subestimación de la capacidad predictiva al no contar con datos puramente de testeo y haber utilizado validación cruzada espacial.

Nuestra tercera y última hipótesis predijo que los modelos con efectos aleatorios estructurados espacialmente lograrían mayor eficacia estadística que los modelos con efectos aleatorios no estructurados en el modelado de datos espaciales. Nuestros resultados para el conjunto de datos de la Cuenca del Carcarañá mostraron que los modelos espaciales obtuvieron mejores métricas de capacidad predictiva, tanto en WAIC, DIC y LPML, como en $RMSE$ y MAE de entrenamiento y testeo calculados por validación cruzada espacial, y en criterios de validación como la correlación lineal entre las observaciones y los predichos. El modelo BYM2 para el conjunto de datos "PCA-hclust" fue el que generó menor error de predicción calculado tanto en los conjuntos de testeo como de entrenamiento. El modelo $iCAR$ con "PCA-hclust" mostró error levemente mayor al BYM2. Los mapas de predichos mostraron ajuste aceptable para el MLR, ajuste aceptable pero falta de suavidad espacial en el modelo iid, buen ajuste y suavidad para el modelo iCAR, y muy buen ajuste en el modelo BYM2.

En los modelos no espaciales la autocorrelación espacial en los residuos fue muy visible, mientras que en el modelo espacial $iCAR$ casi desapareció y en el modelo espacial BYM2 incluso el test de Moran no rechazó la hipótesis de autocorrelación espacial. En general, los modelos espaciales mostraron menor cantidad de CPOs cercanos a cero, con la excepción de la zona alta en el Noroeste de la cuenca, que correspondió a las sierras de Córdoba con valor de ESPI muy alto, y que obtuvo CPOs bajos en todos los modelos, demostrando falta de ajuste para esa región. Los histogramas de los PIT coinciden con lo observado en los mapas de CPO y sugieren que podría

haber un problema de ajuste en las colas. Por este motivo creemos que podría ser interesante modelar el territorio de la cuenca con modelos separados para las sierras y para el valle, ya que por la escala de paisaje a la que trabajamos, asumimos que podría haber diferencias estructurales en el territorio no captadas por las covariables.

Finalmente elegimos el modelo BYM2 entrenado con "PCA-hclust" como el mejor modelo predictivo. La elección se basó en que este modelo fue el que menor error de predicción de testeo, menores DIC y WAIC, y mayor LPML mostró, y además fue el único que generó residuos no correlacionados espacialmente. Los coeficientes estimados con el BYM2 indicaron que el ESPI en la región de la Cuenca del Carcarañá tuvo un valor promedio de 1.84 y que niveles altos de provisión de servicios ecosistémicos estuvieron principalmente relacionados con la presencia de agua, pasturas naturales manejadas, monte y pastizal natural. Con menor fuerza se observó una relación positiva entre el ESPI de la región y la elevación del terreno, la Conductividad Eléctrica, y el porcentaje de alfalfa, pasto, arbustales y matorrales. Con similar magnitud de observó una relación negativa entre el ESPI y el porcentaje de cultivos anuales para cosecha o forraje en la región

6. Referencias

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Albert, J. (2009). *Bayesian computation with R*. Springer.

Allenby, G. M., Rossi, P. E., McCulloch, R. E. (2005). *Hierarchical Bayes Models: A Practitioners Guide*. Ssrn Electronic Journal.

Banerjee, S., Carlin, B. P., Gelfand, A. E. (2015). *Hierarchical modeling and analysis for spatial data*. Chapman & Hall/CRC.

Beale C. M. , Lennon J. J., Yearsley J. M., Brewer M. J., Elston D. A. (2010) Regression analysis of spatial data. *Ecology Letters*, 2010 Feb;13(2):246-64. doi: 10.1111/j.1461-0248.2009.01422.x. PMID: 20102373.

Beguin, J., Martino, S., Rue, H., Cumming, S. G. (2012). Hierarchical analysis of spatially autocorrelated ecological data using integrated nested Laplace approximation. *Methods in Ecology and Evolution*, 3(5), 921-929. https://doi.org/10.1111/j.2041-210X.2012.00211.x

Belsley, D. A., Kuh, E., Welsch, R. E. (1980). *Regression diagnostics*. John Wiley & Sons.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society*. Series B (Methodological), 36(2), 192–236. http://www.jstor.org/stable/2984812

Besag, J., York, J. and Mollie, A. (1991) Bayesian Image Restoration with Two Applications in Spatial Statistics (with Discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1-59. https://doi.org/10.1007/BF00116466

Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V. (2008). *Applied spatial data analysis with R* (Vol. 747248717). Springer.

Blangiardo, M., Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Breslow, N. E., Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Mo-

dels. *Journal of the American Statistical Association*, 88(421), 9–25. https://doi.org/10.2307/2290687

Carlin, B. P., Louis, T. A., Carlin, B. P. (2009). *Bayesian methods for data analysis*. CRC Press.

Casella, G., Berger, R. L. (2002). *Statistical inference*. Duxbury.

Cressie, N. A. C. (2015). *Statistics for spatial data*. John Wiley & Sons.

Dray, S., Saïd, S., Débias, F. (2008). Spatial Ordination of Vegetation Data Using a Generalization of Wartenberg's Multivariate Spatial Correlation. *Journal of Vegetation Science*, 19(1), 45–56. http://www.jstor.org/stable/25173169

Gelman, A., Meng, X.-L., Stern, H. (1996). Posterior Predictive Asessment of Model Fitness via Realized Discrepancies. *Statistica Sinica*, 6(4), 733–760. http://www.jstor.org/stable/24306036

Gelman, A. (2004). *Bayesian data analysis*. Chapman & Hall/CRC.

Gómez-Rubio, Virgilio. (2020). *Bayesian Inference with INLA*. Chapman & Hall/CRC.

Gregoire, T. G., Brillinger, D. R., Diggle, P. J., Russek-Cohen, E., Warren, W. G., Wolfinger, R. D. (1997). *Modelling Longitudinal and Spatially Correlated Data*. Springer.

Harrell, J. F. E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer.

Hastie, T., Tibshirani, R., Friedman, J. (2009) *The Elements of Statistical Learning*. Springer.

Illian, J., Sørbye, S., Rue, H., Hendrichsen, D. (2010). Fitting a log Gaussian Cox process with temporally varying effects – a case study. NTNU.

Illian, J. B., Sørbye, S. H., Rue, H. (2012). A Toolbox for Fitting Complex Spatial Point Process Models Using Integrated Nested Laplace Approximation (INLA). *The Annals of Applied Statistics*, 6(4), 1499–1530. http://www.jstor.org/stable/41713484

Leamer, E. E. (1973). Multicollinearity: A Bayesian Interpretation. *The Review of Economics and Statistics*, 55(3), 371–380. https://doi.org/10.2307/1927962

Lindgren, F., Rue, H. (2015). Bayesian Spatial Modelling with R-INLA. *Journal of Statistical*

Software, 63(19), 1 - 25. doi:http://dx.doi.org/10.18637/jss.v063.i19

McCullagh, P., Nelder, J. (1989). *Generalized Linear Models*, Second Edition. Chapman & Hall.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan.* CRC Press/Taylor & Francis Group.

Montgomery, D. C., Peck, E. A., Vining, G. G. (2012). *Introduction to linear regression analysis.* John Wiley & Sons, Inc.

Nelder, J. A., Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society.* Series A (General), 135(3), 370–384. https://doi.org/10.2307/2344614

Pinheiro, J. C., Bates, D. M. (1995), "Approximations to the Log-likelihood Function in the Nonlinear Mixed Effects Model". *Journal of Computational and Graphical Statistics*, 4, 12–35.

Pinheiro, J. C., Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS.* Springer.

Riebler, A., Sørbye, S., Simpson, D.P., Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25, 1145 - 1165.

Rue, H., Martino, S., 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. http://www.jstor.org/stable/40247579

Rue, H., Held, L. (2005). *Gaussian Markov random fields: Theory and applications.* Chapman & Hall/CRC.

Schabenberger, O., Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis.* Chapman & Hall/CRC

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H. (2017). Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, 32(1), 1–28. http://www.jstor.org/stable/26408114

Tomlinson (1968). A Geographic Information System for Regional Planning.

Molenberghs, G., Verbeke, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer.

Wang, X., Yue, Y., Faraway, J. J. (2018). *Bayesian regression modeling with INLA*. CRC Press.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.*, 14, 867-897.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., Guillera-Arroita, G. blockCV: an R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *bioRxiv 357798*; doi: https://doi.org/10.1101/357798

7. Anexos

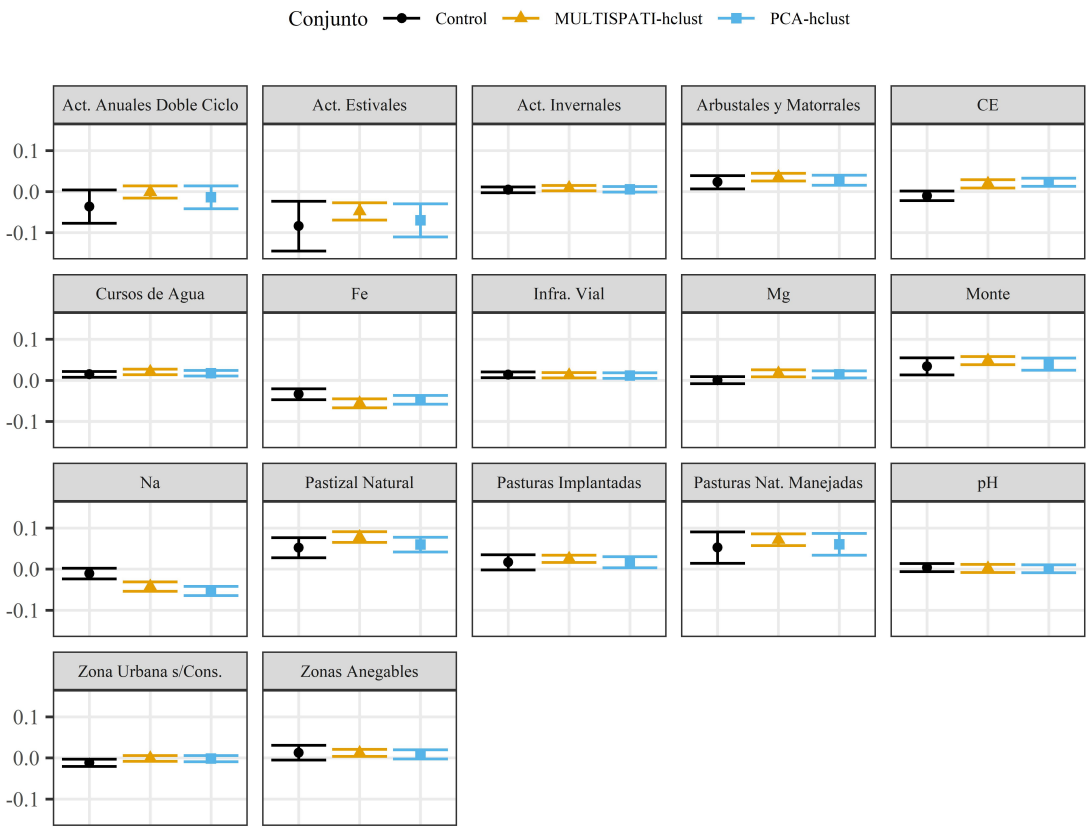

Figura 27: Intervalos de credibilidad del 95 % para los coeficientes de regresión del modelo *iid* ajustado con los conjuntos "Control", "PCA-hclust" y "MULTISPATI-hclust". El punto central de cada intervalo representa la media a posteriori del coeficiente.

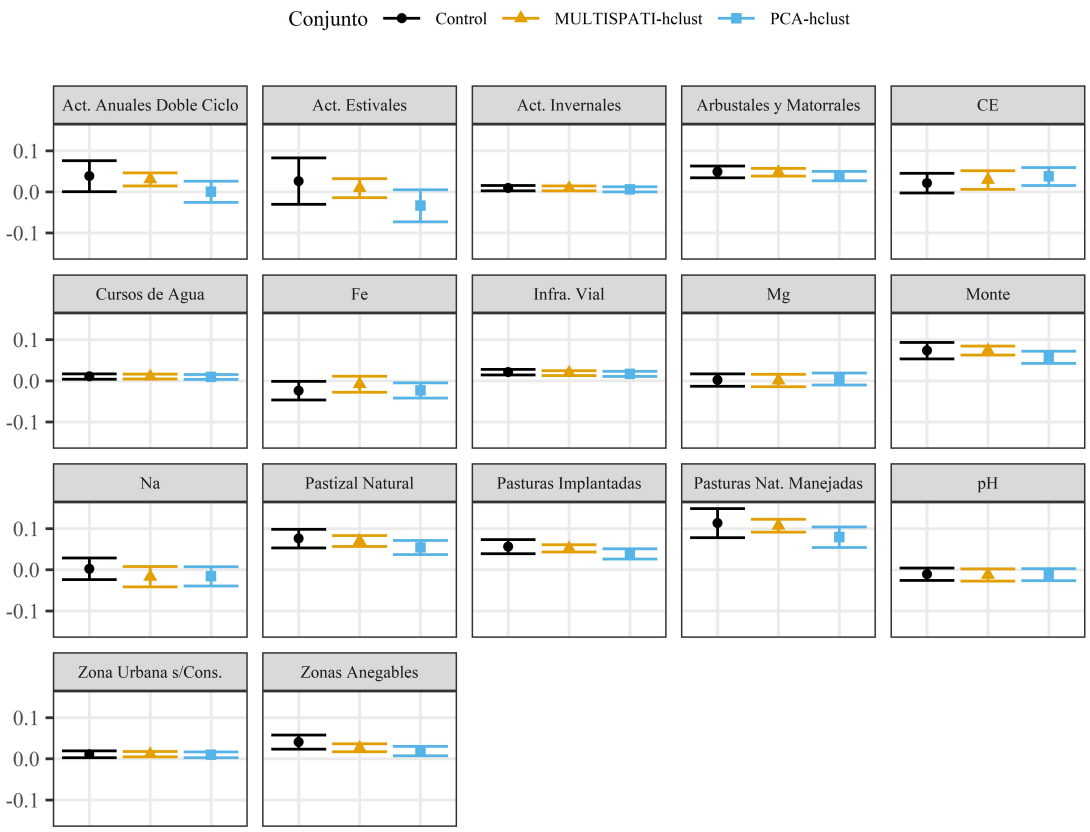

Figura 28: Intervalos de credibilidad del 95 % para los coeficientes de regresión del modelo BYM2 ajustado con los conjuntos "Control", "PCA-hclust" y "MULTISPATI-hclust". El punto central de cada intervalo representa la media a posteriori del coeficiente.

Tabla 9: Tabla con las 36 covariables originales en escala original, sus distribuciones y cantidad de observaciones faltantes.

Variable	Min.	Q1	Mediana	Media	Q3	Max.	Faltantes
Arcilla	9.2	13.8	16	17.5	21.8	29.7	406
Arena	7.3	24.6	48.2	44.1	61.4	79.7	406
CC	7.6	13.6	16.3	17.4	21.2	28.9	406
CE	-1.9	0.3	0.9	1	1.3	6.8	406
Ceextsat	6.4	11.2	12.9	13	15	21.1	406
CIC	7.3	11.6	14.1	14.3	16.6	25.1	406
Cu	0.4	0.9	1.3	1.5	2.1	3.9	406
Fe	57.7	102.7	137.4	139.4	172.7	257.8	406
K	1.3	1.7	1.9	2	2.2	3.2	406
Mg	0.7	2.3	2.7	2.8	3.3	5.6	406
Mn	39	91.8	122.3	123.8	151.6	220.7	406
Na	-0.2	0.6	0.9	0.9	1.2	3	406
Nt	-2.7	-2.3	-2.1	-2.1	-2	-1.7	406
P	13.7	28.1	40.6	42.4	53.9	97.2	406
pH	5.5	6.3	6.5	6.6	6.8	9.2	406
% Monte	0	0	0	2	0.3	100	NA
%. Arbustales y Matorrales	0	0	0	1.7	0	74.4	NA
% Pastizal Natural	0	0	0	2.5	0	100	NA
% Pastizal Natural con Rocas	0	0	0	1.1	0	100	NA
% Cuerpos de Agua	0	0	0	2.6	1.2	100	NA
% Zonas Anegables	0	0	0.6	3.2	3.5	100	NA
% Cursos de Agua	0	0	0	0.2	0	20	NA
% Zona Urbana Consolidada	0	0	0	0.2	0	98.4	NA
% Zona Urbana en Proceso de Cons.	0	0	0	0.4	0	100	NA
% Zona Urbana sin Consolidar	0	0	0	0.3	0	64.7	NA
% Infraestructura Vial	0	0	0.3	1	1.6	28.9	NA
% Actividades Invernales	0	0	0	0.4	0.4	45	NA
% Actividades Estivales	0	15.9	36.2	39.1	59.9	100	NA
% Actividades Anuales Doble Ciclo	0	9.1	23.7	24.9	37.9	100	NA
% Pasturas Implantadas	0	0.2	1.8	5	5.8	86.5	NA
% Pasturas Naturales Manejadas	0	3.1	8.5	14.8	19.6	100	NA
Elevacion	65.8	123.9	182.8	297.7	346.9	2521.6	NA
Materia Organica	0.5	1.8	2.1	2.2	2.5	7.3	26
Pendiente	0.1	2.3	2.8	3.8	3.6	44.4	NA
Recurrencia	0	0	0.9	3.8	3.6	95.2	NA
Valor Unitario de la Tierra (VUT)	0	244941	334127	339129	372265	9592857	178

Tabla 10: Tabla con las cargas de las covariables en las componentes principales de PCA y MULTISPATI para la aplicacion sobre el conjunto de las 36 covariables originales.

	Cargas de las covariables					
Variable	PC1	PC2	PC3	MS-PC1	MS-PC2	MS-PC3
Arena	0.32	-0.08	0.06	-0.32	-0.09	0.06
% Actividades Estivales	0.15	0.03	-0.43	-0.14	0.03	-0.37
CE	0.09	0.01	0.38	-0.09	0.01	0.46
% Pasturas Naturales Manejadas	0.02	0.08	0.37	-0.02	0.09	0.33
% Actividades Invernales	0.02	0.04	0.08	-0.02	0.04	0.05
DEM	0.01	-0.38	-0.04	-0.01	-0.41	-0.04
% Cuerpos de Agua	0.00	0.06	0.14	0.00	0.05	0.04
% Zona Urbana Consolidada	-0.01	0.01	-0.12	0.01	0.00	-0.03
% Pasturas Implantadas	-0.02	0.10	0.11	0.02	0.09	0.07
% Zona Urbana en Proc. De Consolidacion	-0.02	0.01	-0.11	0.02	0.01	-0.06
Recurrencia	-0.02	0.09	0.20	0.02	0.08	0.10
VUT	-0.03	0.09	-0.26	0.03	0.09	-0.18
% Zona Urbana sin Consolidar	-0.04	-0.03	-0.04	0.04	-0.03	-0.02
% Zonas Anegables	-0.03	0.11	0.21	0.04	0.12	0.18
% Cursos de Agua	-0.04	-0.13	0.01	0.04	-0.12	-0.01
Na	-0.04	-0.14	0.36	0.04	-0.13	0.46
% Infraestructura Vial	-0.04	0.03	-0.07	0.05	0.04	-0.06
% Pastizal Natural con Rocas	-0.04	-0.20	0.05	0.05	-0.24	0.08
% Monte	-0.05	-0.17	0.04	0.05	-0.15	0.02
Pendiente	-0.05	-0.36	0.08	0.06	-0.37	0.08
% Pastizal Natural	-0.06	-0.32	0.10	0.06	-0.32	0.10
Fe	-0.08	0.25	0.31	0.08	0.24	0.34
% Arbustales y Matorrales	-0.09	-0.27	0.12	0.09	-0.24	0.10
% Actividades Anuales de Doble Ciclo	-0.10	0.25	-0.01	0.10	0.24	0.02
pH	-0.10	-0.16	0.05	0.10	-0.15	0.10
P	-0.14	0.24	0.08	0.14	0.23	0.11
Mg	-0.16	-0.12	0.03	0.16	-0.12	0.08
K	-0.19	0.22	0.06	0.19	0.22	0.09
Materia Organica	-0.25	-0.23	0.05	0.26	-0.22	0.03
Cu	-0.28	-0.12	-0.11	0.28	-0.12	-0.11
Mn	-0.29	0.09	0.05	0.29	0.10	0.06
Ceextsat	-0.29	-0.11	-0.09	0.29	-0.11	-0.11
Nt	-0.31	-0.02	0.00	0.31	-0.02	-0.02
Arcilla	-0.31	0.13	-0.07	0.31	0.13	-0.07
CC	-0.33	0.08	-0.05	0.33	0.09	-0.05
CIC	-0.33	0.02	-0.08	0.33	0.02	-0.08

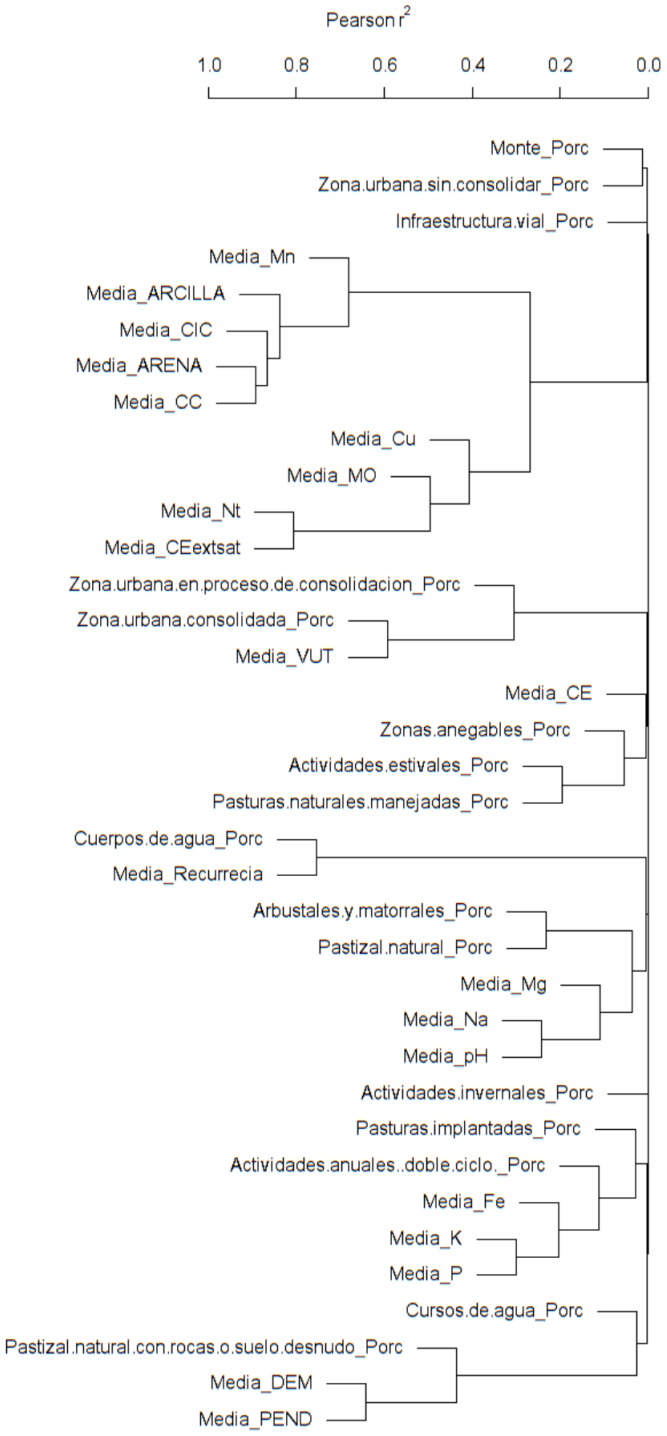

Figura 29: Dendrograma del clustering jerárquico para detectar los grupos de variables colinealies en las 36 covariables del Conjunto 1 de datos.