PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Regresión espacial cuantílica para variables acotadas entre (0,1)

TESIS PARA OPTAR POR EL GRADO DE MAGÍSTER EN ESTADÍSTICA

Presentado por:

Carlos Jeffer García Céspedes

Asesora: Dra. Zaida Jesús Quiroz Cornejo

Miembros del jurado:
Dr. Cristian Luis Bayes Rodriguez
Dra. Rocío Paola Maehara Aliaga

Lima, Octubre 2020

Dedicatoria

El presente trabajo es dedicado a Dios, a mi familia, a mis amigos y a mis profesores de la Maestría quienes me han apoyado y motivado en todo momento.

Agradecimentos

En primer lugar, a mi asesora, la Dra. Zaida Quiroz Cornejo a quien le expreso mi completa gratitud por todo su apoyo. Por haberme guiado, por los conocimientos compartidos y por la paciencia que ha tenido conmigo durante el desarrollo de este trabajo.

A todos los profesores de la Maestría de Estadística de la PUCP, por sus enseñanzas, por el respeto y la motivación que me han brindado.

A la PUCP, mi centro de estudios y de trabajo, por impulsarme a alcanzar mis metas y brindarme un ambiente e infraestructura ideal durante toda la Maestría.

A mis amigos y compañeros de la Maestría por siempre brindarme una sonrisa y apoyo cuando lo necesitaba.

A mi familia por estar siempre conmigo.

Resumen

El Perú es un país emergente donde el desarrollo se centra en algunas ciudades y distritos específicos. Esto conlleva a mucha desigualdad económica por ello resulta importante dar seguimiento a la incidencia de pobreza en el país. De acuerdo al nivel de precariedad, la pobreza puede considerarse extrema o no extrema. En este contexto, estudiamos la incidencia de pobreza no extrema a través de un modelo de regresión cuantílica espacial a nivel distrital en la provincia de Lima utilizando la distribución de Kumaraswamy combinada con un efecto espacial intrínseco condicional autorregresivo (ICAR). Para tratar y evaluar la posible confusión espacial entre los efectos espaciales y las covariables de efectos fijos, se considera, también, el enfoque SPOCK (Spatial Orthogonal Centroid "K"orrection). Nuestros modelos pertenecen a la clase de modelos jerárquicos, para los cuales la inferencia se puede realizar utilizando el metodo de Monte Carlo Hamiltoniano. Por lo tanto, el modelo es computacionalmente factible para grandes conjuntos de datos, puede describir puntos extremos de la distribución de la incidencia de pobreza no extrema e identificar qué factores son importantes en las colas de la distribución de los datos.

Índice general

Lista de Abreviaturas

SPOCK	Spatial Orthogonal Centroid Korrection
MCMC	Cadenas de Markov de Monte Carlo.
ADL	Distribución asimétrica de Laplace.
CAR	Modelo condicional autoregresivo.
GMRF	Campo aleatorio de Markov gaussiano.
WAIC	Criterio de información Watanable-Akaike
LPML	Logaritmo de la verosimilitud pseudo marginal
RMSE	Raiz cuadrada del error cuadrático medio
INEI	Instituto Nacional de Estadística e Informática

Lista de Símbolos

$f_X(.)$	Función de densidad de probabilidad de X.
$F_X(.)$	Función de probabilidad acumulada de X.
$F^{-1}(.)$	Función del cuantil.
$E(X)$	Esperanza de la variable aleatoria X.
μ_X	Media de la variable aleatoria X.
σ_X	Desviación estándar de la variable aleatoria X.
$\boldsymbol{Q}$	Matriz de precisión.
$\boldsymbol{W}$	Matriz de vecindad.

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Consideraciones preliminares

En muchas ocasiones se conoce el espacio o región geográfica en donde ocurre una variable bajo estudio. Intuitivamente se puede presumir que la variable se comporta de manera parecida en lugares o regiones cercanas, lo que conlleva a una autocorrelación espacial entre variables medidas en distintos espacios geográficos. Surge entonces la necesidad de diseñar modelos que permitan trabajar con la dependencia espacial de los datos para tomar en cuenta la influencia de las áreas cercanas o vecinas.

Por otro lado, el comportamiento de la variable perteneciente a un espacio geográfico puede ser muy asimétrico lo que conlleva a que los modelos gaussianos subestimen las probabilidades en los cuantiles, pese a que, en muchas ocasiones, es muy importante estimar correctamente la probabilidad en las colas de la distribución. Además, la relación entre la variable respuesta y covariables puede ser no lineal y los efectos de las covariables podrían no estar solamente restringidos a la media de la variable en estudio. Surge, entonces, la necesidad de utilizar un modelo estadístico que prediga, no solamente la media de la variable en análisis, sino, también los cuantiles, a este tipo de modelo se le conoce como regresión cuantílica.

En particular en esta tesis se propone un modelo de regresión cuantílica para datos acotados en el intervalo (0,1) y definidos en áreas geográficas. La relevancia de este tema es debido a que la literatura es limitada para el análisis de datos acotados en (0,1) con dependencia espacial.

La variable que se desea estudiar hace referencia a un indicador económico en un espacio geográfico por lo que su valor se encuentra en el intervalo (0,1). Algunas distribuciones que modelan variables continuas en este intervalo son la beta, logística y Kumaraswamy. Para el presente proyecto se utilizará la distribución de Kumaraswamy la cual es similar a la distribución beta pero con algunas propiedades adecuadas para este tipo de modelos.

Aunque se puede ajustar modelos espaciales desde el punto de vista de inferencia clásica, los modelos bayesianos permiten cuantificar la incertidumbre de las estimaciones realizadas y brindan una mayor flexibilidad para trabajar con otras distribuciones no gaussianas y con datos faltantes, por ejemplo, para áreas en donde no ha podido realizarse una medición. Para la estimación del presente modelo se utilizará el paquete STAN el cual se basa en el método de las Cadenas de Markov de Monte Carlo (MCMC) que es generalmente usado para realizar inferencia bayesiana.

1.2. Objetivos

El objetivo general de la tesis es estudiar propiedades, estimar y aplicar a un conjunto de datos reales un modelo de regresión cuantílica para variables acotadas entre (0,1) en datos de áreas geográficas. De manera específica:

- Revisar la literatura acerca de las diferentes propuestas de modelos de regresión cuantílica.
- Proponer, estudiar las propiedades, e implementar la estimación de un modelo bayesiano de regresión cuantílica, basado en la distribución de Kumaraswamy, para datos de áreas geográficas.
- Realizar estudios de simulación considerando computación intensiva sobre diferentes escenarios. Esto nos permitirá comprender como afectan los distintos parámetros al ajuste del modelo. Y, además, debido a la complejidad del modelo espacial, este código servirá de modelo para futuras reproducciones del mismo.
- Aplicar el modelo a un conjunto de datos referidos a la incidencia de pobreza no extrema en la provincia de Lima a nivel distrital.

1.3. Organización del trabajo

El presente documento está organizado de la siguiente manera:

- En el Capítulo 2, presentamos conceptos previos al desarrollo del modelo que se describe en el presente trabajo. Aquí se describen las propiedades y la reparametrización de la distribución de Kumaraswamy, el modelo espacial condicional autorregresivo para datos de áreas geográficas que se utilizará, y como se realizará la implementación de la regresión cuantílica utilizando el paquete STAN.
- En el Capítulo 3, se describe la estructura del modelo de regresión cuantílica basado en la distribución de Kumaraswamy para datos de áreas, específicamente, se presenta el modelo jerárquico de tres niveles que se utilizará, en cuyo primer nivel se presenta la función de verosimilitud, en el segundo nivel se presenta el modelo condicional autoregresivo utilizado y en el tercer nivel, los hiperparámetros; además, se presenta las distribuciónes a priori, funciones de enlace utilizadas, distribuciones conjuntas a posteriori y sus distribuciones marginales respectivas.
- En el Capítulo 4, se presenta los resultados de un estudio de simulación del modelo en estudio bajo tres escenarios correspondientes a los cuantiles $q = 0.1$, $q = 0.5$ y $q = 0.9$. Dentro de cada escenario se presentan cuatro subescenarios que permiten analizar el modelo ante distintos valores en sus parámetros.
- En el Capítulo 5, se presenta los resultados de una aplicación para datos a nivel distrital del índice de pobreza no extrema en la provincia de Lima.
- Finalmente, en el Capítulo 6 se presentan las conclusiones del presente trabajo.

Capítulo 2

Conceptos preliminares

En este capítulo se revisan conceptos importantes para el modelo propuesto en el siguiente capítulo.

2.1. Regresión cuantílica

La definición usual del cuantil κ_q de una variable aleatoria y es aquel valor hasta el cual una proporción q de valores de la población son menores o iguales a éste. Formalmente el cuantil κ_q es definido por:

$$P(y \leq \kappa_q) \geq q, \tag{2.1}$$

donde $q \in (0, 1)$.
En el caso de variables continuas el cuantil se puede definir como la inversa de la función acumulada evaluada en q, es decir,

$$\kappa_q = F^{-1}(q)$$
$$F(\kappa_q) = q,$$

donde $Q(q) = F^{-1}(q) = \kappa_q$ es llamada la función del cuantil de y.
En los modelos de regresión usualmente se modela la media de la variable respuesta en base a las covariables; no obstante, muchas veces puede ser útil, modelar la varianza (para el análisis de heterocedasticidad). Por otro lado, puede ser de mucha utilidad estimar el efecto de las covariables en los cuantiles de la variable respuesta. Las principales ventajas de modelar los cuantiles de una variable son las siguientes:

- En base a la estimación de los cuantiles de una variable se puede obtener virtualmente la distribución completa de la variable respuesta. De esta manera se puede analizar la asimetría de la distribución, lo cual no es posible mediante la regresión de la media o la varianza.
- No requiere las restricciones de homocesdasticidad o un tipo específico de distribución para la variable respuesta o equivalentemente para los errores.
- Se puede analizar valores extremos de la distribución.

La definición para los cuantiles en la ecuación (2.1) se puede reformular como

$$\kappa_q = \underset{\kappa}{\operatorname{argmin}} \, \mathrm{E}(w_q(y, \kappa)|y - \kappa|),$$

donde

$$w_q(y, \kappa) = \begin{cases} 1 - q, & y < \kappa \\ 0, & y = \kappa \\ q, & y > \kappa \end{cases} \tag{2.2}$$

fue propuesto por (Fahrmeir et al., 2013).
Por lo que una estimación apropiada para el cuantil dada una muestra aleatoria $\boldsymbol{y} = y_1, \ldots, y_n$ vendría dada por

$$\widehat{\kappa_q} = \underset{\kappa}{\operatorname{argmin}} \sum_{i=1}^{n} w_q(y_i, \kappa)|y_i - \kappa|.$$

2.1.1. Regresión cuantílica clásica

La regresión cuantílica clásica propuesta por Koenker y Bassett (1978) se basa en una equivalencia en las técnicas de regresión de la media pero para el caso del cuantil de y. De esta manera si para la regresión lineal se tenía que

$$y = \boldsymbol{x}^\intercal \boldsymbol{\beta} + \epsilon,$$

donde $\boldsymbol{x^t}$ representa a las covariables, $\boldsymbol{\beta}$ los coeficientes y ϵ el error aleatorio, se tiene que $\mathrm{E}(\epsilon) = 0$ y, entonces $\mathrm{E}(y) = \boldsymbol{x}^\intercal \boldsymbol{\beta}$. Para el caso de la regresión cuantílica se asume un modelo parecido

$$y = \boldsymbol{x}^\intercal \boldsymbol{\beta_q} + \epsilon_q,$$

donde se asume que el cuantil κ_q de ϵ es 0, es decir, $q = F_{\epsilon_q}(0) = P(\epsilon_q \leq 0) = P(\boldsymbol{x}^\intercal \boldsymbol{\beta_q} + \epsilon_q \leq \boldsymbol{x}^\intercal \boldsymbol{\beta_q}) = P(y \leq \boldsymbol{x}^\intercal \boldsymbol{\beta_q}) = F_y(\boldsymbol{x}^\intercal \boldsymbol{\beta_q})$, por lo tanto κ_q de y es estimado por el predictor $\boldsymbol{x}^\intercal \boldsymbol{\beta_q}$. Entonces el coeficiente de regresión $\boldsymbol{\beta_q}$ es estimado por

$$\widehat{\boldsymbol{\beta_q}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} w_q(y_i, \eta_{iq})|y_i - \eta_{iq}|, \tag{2.3}$$

donde $\eta_{iq} = \boldsymbol{x}^\intercal \boldsymbol{\beta_q}$ y $w_q(.,.)$ está definida en la ecuación (2.2).

2.1.2. Regresión cuantílica bayesiana

La regresion cuantílica ha podido ser utilizada en la inferencia bayesiana gracias al aporte de Yu y Moyeed (2001). Para ello utilizaron una equivalencia entre la maximización de la función a posteriori de β con el criterio de optimización de la regresión cuantilica clásica definida en la ecuación (2.3). Para ello propusieron asumir la distribución asimétrica de Laplace (ALD) para la variable respuesta, $y \sim \mathrm{ALD}(\mu, \sigma^2, \mathrm{q})$, donde μ es el parámetro de localización, σ^2 es el parámetro de precisión y q un parámetro de simetría, cuya función de

densidad es dada por

$$f(y|\mu,\sigma^2,q) = \frac{q(1-q)}{\sigma^2}\exp\{-w_q(y,\mu)\frac{|y-\mu|}{\sigma^2}\}.$$

En su propuesta consideran el modelo $y_i = \boldsymbol{x}^\intercal\boldsymbol{\beta_q}+\epsilon_{iq}, i = 1,\ldots,n$ con $\epsilon_{iq}|\sigma^2 \overset{iid}{\sim} \text{ALD}(0,\sigma^2,q)$, con ϵ_{iq} variables aleatorias independientes e idénticamente distribuidas, esto induce a que las respuestas y_i tengan una distribución

$$y_i|\boldsymbol{\beta_q},\sigma^2 \overset{iid}{\sim} \text{ALD}(\boldsymbol{x}^\intercal\boldsymbol{\beta_q},q).$$

De esta manera, considerando una distribución a priori no informativa, $P(\boldsymbol{\beta_q})$, con σ^2 y q fijos, tenemos aquí que la distribución a posteriori será dada por

$$\begin{aligned} p(\boldsymbol{\beta_q}|y,\sigma^2) &\propto \prod_{i=1}^{n} p(y_i|\boldsymbol{\beta_q},\sigma^2) \\ &\propto \exp\left(-\sum_{i=1}^{n} w_q(y_i,\boldsymbol{x}^\intercal\boldsymbol{\beta_q})\frac{|y_i-\boldsymbol{x}^\intercal\boldsymbol{\beta_q}|}{\sigma^2}\right), \end{aligned} \tag{2.4}$$

en esta última expresión podemos notar, bajo estos supuestos, que maximizar el estimador a posteriori es equivalente al estimador obtenido en la ecuación (2.3).

2.1.3. Regresión cuantílica para datos limitados

Así como la distribución ALD, existen otras distribuciones que pueden ser reparametrizadas para asumir al cuantil como un parámetro de localización. Entre algunas de ellas podemos encontrar la distribución de Poisson, la distribución generalizada de Pareto, la distribución log-logística y la distribución de Kumaraswamy. Esta última, dadas sus propiedades, permiten utilizarla para estudios cuantílicos en los cuales la variable respuesta se encuentra limitada en el intervalo (0,1). Por ejemplo, ha sido utilizada por Carrasco et al. (2010), Mitnik y Baek (2011), Ali et al. (2015) y Taillardat (2016)

La distribución beta, a diferencia de la distribución Kumaraswamy, no es usada para modelos de regresión cuantílicos, ya que su función de distribución acumulada no posee una forma cerrada [Bayes et al. (2017) y Mitnik y Baek (2011)].

Otra opción para datos cuyo rango es limitado a un intervalo es la distribución log-logística la cual tiene como soporte a los reales positivos. Algunas aplicaciones de la distribución log-logística son presentadas en Shoukri et al. (1988) y Mkhandi et al. (1996).

2.2. Distribución Kumaraswamy

Una variable aleatoria Y sigue una distribución de Kumaraswamy con función de densidad de probabilidad (fdp) dada por

$$f_Y(y \mid \alpha,\beta) = \alpha\beta(y)^{\alpha-1}(1-y^\alpha)^{\beta-1}, 0<y<1, \alpha,\beta>0. \tag{2.5}$$

Esta distribución presentada en Kumaraswamy (1980), tiene características similares a la distribución beta. Entre las similitudes, se tiene que ambas presentan el mismo soporte (0,1),

y se encuentran caracterizadas por dos parámetros de forma. Por otro lado, la función de distribución acumulada de la distribución Kumaraswamy tiene una forma invertible cerrada lo que permite que sea más adecuada para la generación de variables aleatorias y ser usada en modelos de regresión cuantílica. La función de distribución acumulada de la variable aleatoria Y con distribución Kumaraswamy es

$$F(y) = 1 - (1 - y^{\alpha})^{\beta}. \tag{2.6}$$

Su media y varianza son expresadas respectivamente por

$$E(y) = \beta B(1 + \frac{1}{\alpha}, \beta),$$

$$V(y) = \beta B(1 + \frac{2}{\alpha}, \beta) - \beta^2 B^2(1 + \frac{1}{\alpha}, \beta),$$

donde $B(.,.)$ denota a la función beta.

2.2.1. Reparametrización de la distribución Kumaraswamy

En Mitnik y Baek (2011) y Bayes et al. (2017) se presenta una parametrización alternativa de la distribución Kumaraswamy en la que se toma al cuantil como parámetro de localización, y un parámetro de precisión. En el Apéndice A.1 se presenta con mayor detalle los cálculos realizados para llegar a la reparametrización de esta distribución. Esta reparametrización se logra debido a que la inversa de la función de distribución acumulada tiene forma cerrada. Si consideramos $\kappa(q)$ como el cuantil de nivel q este será dado por

$$\kappa(q) = F^{-1}(q) = [1 - (1 - q)^{1/\beta}]^{1/\alpha}. \tag{2.7}$$

Además Mitnik y Baek (2011) demostraron que

$$\phi = -\log(1 - (1 - q)^{1/\beta}), \tag{2.8}$$

puede ser considerado como un parámetro de precisión.
Las expresiones dadas en las ecuaciones (2.7) y (2.8) definen la reparametrización del modelo Kumaraswamy a ser utilizado,bajo el cual, las funciones de densidad de probabilidad y la acumulada son dadas por

$$f_Y(y|\kappa,\phi) = -\frac{\log(1-q)\phi}{\log(1-e^{-\phi})\log(\kappa)} y^{-\frac{\phi}{\log(\kappa)}-1}\{1 - y^{-\frac{\phi}{\log(\kappa)}}\}^{\frac{\log(1-q)}{\log(1-e^{-\phi})}-1} \tag{2.9}$$

$$F_Y(y|\kappa,\phi) = 1 - \{1 - y^{-\frac{\phi}{\log(\kappa)}}\}^{\frac{\log(1-q)}{\log(1-e^{-\phi})}}, \tag{2.10}$$

su media y varianza son expresadas respectivamente por

$$E(y) = \frac{\log(1-q)}{\log(1-e^{-\phi})} B\left(1 - \frac{\log(\kappa)}{\phi}, \frac{\log(1-q)}{\log(1-e^{-\phi})}\right).$$

$$V(y) = \frac{\log(1-q)}{\log(1-e^{-\phi})} B\left(1 - \frac{2\log(\kappa)}{\phi}, \frac{\log(1-q)}{\log(1-e^{-\phi})}\right) - \left(\frac{\log(1-q)}{\log(1-e^{-\phi})}\right)^2 B^2\left(1 - \frac{\log(\kappa)}{\phi}, \frac{\log(1-q)}{\log(1-e^{-\phi})}\right).$$

donde $B(.,.)$ denota a la función beta, $0 < \kappa < 1$, $\phi > 0$ y $0 < q < 1$.

Para el presente trabajo se utilizará la notación $Y \sim \text{Kumar}(\kappa, \phi, q)$ para definir a una variable aleatoria con distribución Kumaraswamy de parámetros κ, ϕ y q.

La Figura 2.1 muestra la fdp de la variable Kumaraswamy reparametrizada tomando como parámetro para el nivel del cuantil a 0.5, por lo que κ representa la mediana. En la parte superior se observan 3 casos en los que se ha fijado el parámetro ϕ y se comparan distintos valores de la mediana κ. Se observa que para cada valor de κ, la fdp tiende a centrarse en ese punto. Se observa también que conforme ϕ aumenta la distribución se va volviendo concava, ello se aprecia con mayor claridad en el caso del centro. En la parte inferior se observan 3 casos en los que que se ha fijado la mediana κ y se comparan distintos valores de la precisión. En cada gráfico, a medida que la precisión disminuye se tiene mayor concavidad; además, conforme se cambia la mediana la distribución tiende hacia ese punto.

Por otra parte, la Figura 2.2 muestra la fdp de la variable Kumaraswamy reparametrizada tomando como nivel del cuantil a 0.9 y se han tomado los mismos valores para κ y ϕ usados en la Figura 2.1 en donde ya se ha mostrado que κ es un parámetro de localización y ϕ es de precisión. En este caso, para $q = 0.9$, además, se puede observar que al utilizar un valor de κ bajo como 0.1 o incluso 0.5 la distribución tiene una asimetría positiva, mientras que para valores de κ altos la distribución tiene una asimetría negativa más marcada.

2.3. Estadística espacial para datos de áreas

La estadística espacial proporciona métodos para el análisis de variables aleatorias medidas en el espacio geográfico. Formalmente en este trabajo se asume al vector aleatorio $\boldsymbol{Y(s)} = (Y(s_1), Y(s_2), \ldots, Y(s_n))^\intercal$, donde $s_1, s_2, \ldots, s_n$ están contenidos en el espacio euclidiano bidimensional

La estadística espacial presenta tres perspectivas distintas de acuerdo al tipo de espacio euclidiano D.

En caso este espacio sea discreto se le conoce como estadística para datos de áreas (usualmente ocurre en agregaciones de alguna medida por ciudad o por departamento): los datos de áreas modelan un conjunto finito de variables aleatorias en el espacio. Este tipo de modelamiento asume la definición de áreas vecinas.

Basados en modelos no parámetricos, los vecinos son usados para modelar dependencia espacial a través de un grafo predefinido sobre el cual se aplica inferencia usando técnicas de verosimilitud [Besag (1974) y Kunsch (1987)]. Para definir a los vecinos de cada área se usa la matriz de vecindad $\boldsymbol{W}$:

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix}. \tag{2.11}$$

Se asume que el valor para cada elemento w_{ii} de la diagonal principal es 0. Los elementos de $\boldsymbol{W}$ pueden ser vistos como pesos en donde el peso será mayor mientras exista mayor proximidad entre las áreas. Existen muchas maneras de definir los pesos w_{ij}:

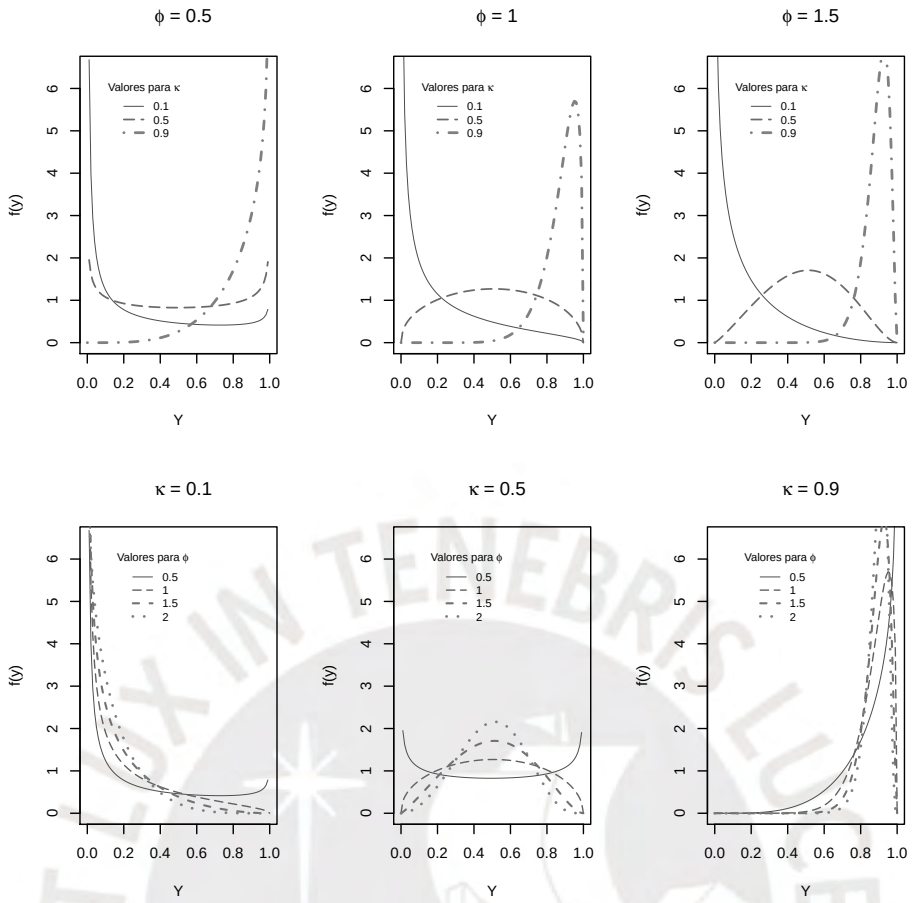


Figura 2.1: Función de densidad de probabilidad de la distribución Kumaraswamy para la mediana

- Si se toma como referencia los centroides de cada área se tienen, la distancia radial, la distancia exponencial o la distancia de potencia.
- Si se toma como referencia los límites entre áreas se tiene los pesos por contiguidad o los pesos por límite compartido.

2.3.1. Medidas de correlación espacial

Para poder aplicar los métodos de la estadística espacial en datos de áreas se debe verificar previamente que las variables Y medidas en las áreas $s_1, s_2 \ldots, s_n$ presentan dependencia o autocorrelación espacial. Dos índices utilizados para medir la presencia de correlación son los de Moran y Geary, los cuales se definen de la siguiente manera: Dadas las variables aleatorias $\{Y(s_1), Y(s_2), \ldots, Y(s_n)\}$ correspondientes a las n áreas y dada la matriz de vecindad $\boldsymbol{W}$, el índice de Moran se define como

$$I_{Moran} = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{S_0 \sum_{i=1}^{n}(y_i - \overline{y})^2} \quad \text{con} \quad S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij},$$

donde $-1 \leq I_{Moran} \leq 1$. Si:

- $I_{Moran} \approx 0$ indica que no hay autocorrelación espacial o hay aleatoriedad espacial entre las áreas vecinas,

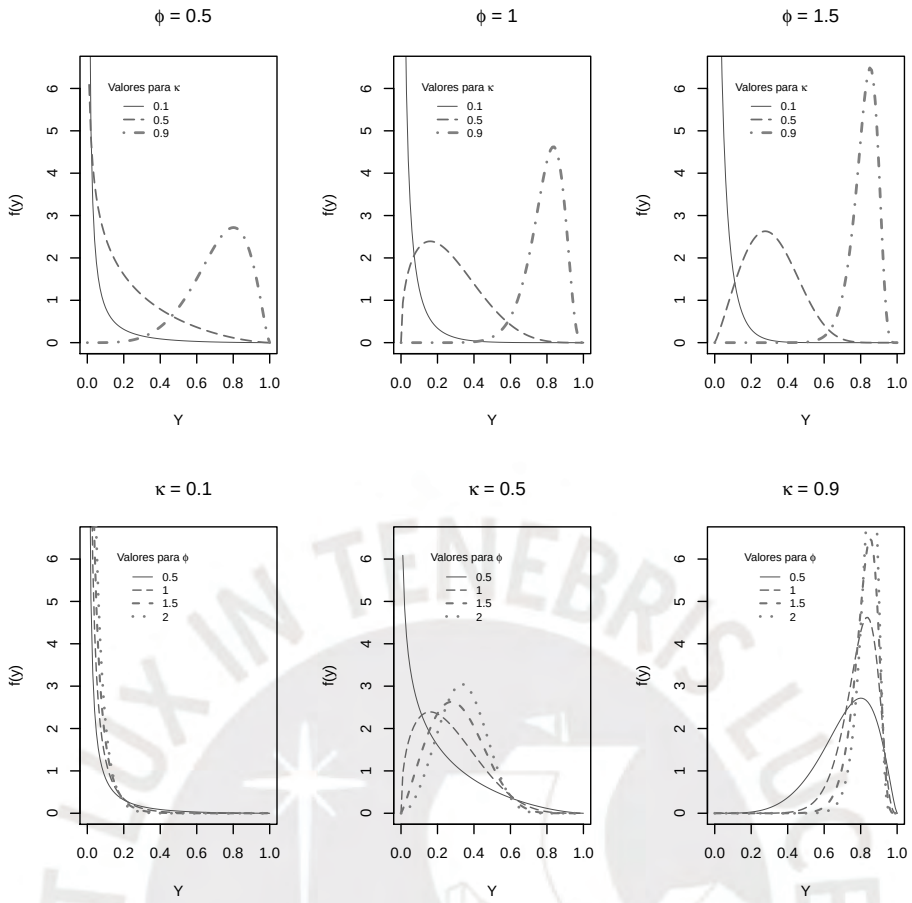


Figura 2.2: Función de densidad de probabilidad de la distribución Kumaraswamy para el cuantil 0.9.

- $I_{Moran} \approx 1$ indica que hay autocorrelación espacial entre las áreas vecinas,

- $I_{Moran} \approx -1$ indica que hay dispersión perfecta entre las áreas vecinas.

El índice de Geary es dado por

$$C_{Geary} = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{2S_0 \sum_{i=1}^{n}(y_i - \overline{y})^2} \quad \text{con} \quad S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij},$$

donde $C_{Geary} > 0$ y valores entre 0 y 1 indican la presencia de autocorrelación espacial.

2.3.2. Campos aleatorios de Markov gaussianos (GMRF)

Una definición importante para la elaboración de los modelos para datos de áreas son los campos aleatorios de Markov Gaussianos (Gaussian Markov random fields - GMRF), los cuales identifican a un conjunto de variables aleatorias, $\boldsymbol{X}^\intercal = (X_1, \ldots, X_n)$, con distribución normal multivariada, $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ que tienen las siguientes características, como indica Gelfand et al. (2010):

X_i y X_j son condicionalmente independientes si se cumple que $Q_{ij} = 0$, donde $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$. Esta independencia se representa usualmente con un grafo espacial. La definición de los GMRF está relacionada con el lema de Brook: Sea $p(\boldsymbol{x})$ la densidad de $\boldsymbol{X} \in \mathbb{R}^n$ y $\Omega = \{\boldsymbol{X} \in \mathbb{R}^n : p(\boldsymbol{X}) > 0\}$. Sea el vector fijo $\boldsymbol{X'} \in \Omega$

$$\begin{aligned}\frac{p(\boldsymbol{X})}{p(\boldsymbol{X'})} &= \prod_{i=1}^{n} \frac{p(X_i|X_1, \ldots, X_{i-1}, X'_{i+1}, \ldots, X'_n)}{p(X'_i|X_1, \ldots, X_{i-1}, X'_{i+1}, \ldots, X'_n)}, \\ &= \prod_{i=1}^{n} \frac{p(X_i|X'_1, \ldots, X'_{i-1}, X_{i+1}, \ldots, X_n)}{p(X'_i|X'_1, \ldots, X'_{i-1}, X_{i+1}, \ldots, X_n)}.\end{aligned}$$

El lema de Brook indica que sea el campo aleatorio de Markov gaussiano $\boldsymbol{X}$ y $\boldsymbol{X'}$ un vector fijo, se puede obtener la distribución conjunta de $\boldsymbol{X}$ a partir de las condicionales completas, ya que del lado derecho de la ecuación se observa que $p(\boldsymbol{X})$ es proporcional a la multiplicación de las condicionales completas (Rue y Held, 2005).
Como se menciona en Besag (1974), a partir de las distribuciones condicionales completas, se puede obtener una distribución conjunta, su media μ y su matriz de precisión $\boldsymbol{Q}$, la cual es definida a partir de la matriz de vecindad dada en (2.11) y la matriz diagonal $n \times n$ formada por las desviaciones estándar de cada X_i de $\boldsymbol{X}$.

2.3.3. El modelo condicional autoregresivo (CAR)

Los modelos CAR se realizan sobre campos aleatorios de Markov definidos en la sección anterior. Sea u_i el efecto de la i-esima área con estructura espacial. Dado un campo aleatorio de Markov gaussiano $\boldsymbol{u}^\intercal = \{u_1, u_2, \ldots, u_n\}$, en este caso correspondiente a las n áreas, cada una con un grupo de vecinos, se asume que:

$$u_i|\boldsymbol{u_{-i}} \sim \mathrm{N}\left(\frac{\sum_{j\sim i} w_{ij}(u_j)}{d + w_{i+}}, \frac{\sigma^2}{d + N_i}\right), \tag{2.12}$$

donde $\boldsymbol{u}^\intercal_{\boldsymbol{-i}} = (u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n)$, $j \sim i$ indica si j es vecino de i, w_{ij} son los componentes de la matriz de vecindad $\boldsymbol{W}$ dada en (2.11), w_{i+} es la suma de valores en la fila i de $\boldsymbol{W}$, N_i es el número de vecinos del área i y d es un parámetro que controlará si la matriz de precisión que se forme a partir de la matriz de vecindad es invertible. Si los componentes de la matriz de vecindad son definidos por

$$w_{ij} = \begin{cases} 1 & \text{, si } i \sim j (i \text{ es vecino de } j), \\ 0 & \text{, si } i \text{ no es vecino de } j \end{cases}$$

entonces

$$u_i|\boldsymbol{u_{-i}} \sim \mathrm{N}\left(\frac{\sum_{j\sim i} w_{i,j}(u_j)}{d + N_i}, \frac{1}{\tau(d + N_i)}\right), \tag{2.13}$$

donde $\tau = 1/\sigma^2$. Es decir, la esperanza de la distribución condicional completa de u_i es dada por la media aritmética de los efectos de sus vecinos y su varianza es proporcional al número de vecinos.
Luego la distribución conjunta de $\boldsymbol{u}$, por el Lema de Brook (Ecuación 2.13), puede ser deri-

vada a partir de (2.13), tal que:

$$\boldsymbol{u} \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{Q}^{-\mathbf{1}}), \tag{2.14}$$

donde $\boldsymbol{Q}^{-\mathbf{1}} = \tau(\boldsymbol{D} - \boldsymbol{W})$ y $\boldsymbol{D}$ es la matriz diagonal $(d + N_i)$. Si $d = 0$ se tiene que $\boldsymbol{D} - \boldsymbol{W}$ es una matriz singular, por lo tanto, la distribución conjunta es impropia y por ello a este modelo se le conoce como CAR impropio o ICAR. A pesar de ello, esta distribución puede ser usada como una a priori para un efecto aleatorio espacial ya que su distribución a posteriori usualmente es válida. Si $d > 0$ la matriz es postiva definida y al modelo se le conoce como CAR propio.
Una definición alternativa para el modelo CAR tiene la siguiente expresión

$$\boldsymbol{u} \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{Q}^{-\mathbf{1}}), \tag{2.15}$$

donde $\boldsymbol{Q}^{-\mathbf{1}} = D_\tau(I - \alpha B))$, $D_\tau = \tau D$, $\boldsymbol{D}$ es la matriz diagonal (N_i), donde, como ya se mencionó, N_i es el número de vecinos del área i-ésima. I es la matriz identidad, α es un parámetro que controla la dependencia espacial ($\alpha = 0$ indica independencia y $\alpha = 1$ indica un modelo ICAR) y $B = D^{-1}W$ es la matriz de vecindad escalada.
En el contexto bayesiano, actualmente los modelos CAR se usan en modelos jerárquicos como es el caso del modelo del presente proyecto.

2.3.4. Matriz de vecindad SPOCK

Dentro del contexto de la estadística espacial puede ocurrir que en los modelos en estudio se tenga una confusión entre los efectos fijos y efectos espaciales debido a la alta correlación espacial que puede existir en las variables del modelo.
Una de las propuestas para corregir este problema consiste en realizar una transformación en el espacio de manera que se pueda asegurar independencia entre los efectos fijos y espaciales. A este enfoque se le ha denominado SPOCK (Spatial Orthogonal Centroid Korrection) (Prates et al., 2019).
De acuerdo a la definición realizada en Prates et al. (2019), este enfoque asume una nueva matriz de vecindad a la que se denomina $\boldsymbol{W^*}$ con la que se alivia una posible confusión entre las variables independientes y el efecto espacial. A partir de ella, se define un nuevo efecto aleatorio espacial $\boldsymbol{u}^\perp$ para garantizar que sea ortogonal a los efectos fijos,

$$\boldsymbol{u}^\perp \sim N_n(\mathbf{0}, \boldsymbol{Q}^{*-\mathbf{1}}), \tag{2.16}$$

donde se define $\boldsymbol{Q}^{*-\mathbf{1}}$ como en (2.14) o (2.15) tal que $\boldsymbol{W}$ es reemplazada por $\boldsymbol{W^*}$. Se crea una matriz de vecindad relacionada con una mapa construido a partir del mapa original de la siguiente manera:
Se tiene una matriz $n \times 2$ denominada $\boldsymbol{s} = [\boldsymbol{s_1}, \boldsymbol{s_2}] = [s_{i1}, s_{i2}]$, $\forall i = 1, \ldots, n$ donde s_{i1}, s_{i2} es la coordenada del centroide del área i-ésima. Luego se definirá $\boldsymbol{s^*} = \boldsymbol{P}^\perp \boldsymbol{s}$ donde $\boldsymbol{P}^\perp = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal$ es la matriz proyectada en el espacio ortogonal a $\boldsymbol{X}$. Considerando $\boldsymbol{s^*}$ se construirá la matriz $\boldsymbol{W^*}$ tomando como vecinos a los centroides mas cercanos con respecto a su distancia Euclidiana.

Para entender el razonamiento usado por el método SPOCK, se puede asumir que el nuevo efecto espacial u_i es un punto $\Psi(s_1, s_2)$ sobre una superficie suave definida para una posición arbitraria (s_1, s_2) que representa a alguno de los centroides de las áreas del mapa y sea $\Lambda\Psi = (\gamma_1, \gamma_2)^\intercal$ la gradiente de Ψ evaluada en el punto (s_{01}, s_{02}) utilizada para realizar la siguiente expansión de Taylor

$$\begin{aligned}\Psi(s_1, s_2) &= \Psi(s_0 1, s_0 2) + (s_1 - s_{01}, s_2 - s_{02})\Lambda\Psi + R(s_1, s_2, s_{10}, s_{20}) \\ &= \gamma_0 + \gamma_1(s_1 - s_{01}) + \gamma_2(s_2 - s_{02}) + R(s_1, s_2, s_{10}, s_{20}),\end{aligned} \tag{2.17}$$

donde el resto $R(s_1, s_2, s_{10}, s_{20})$ tiene forma cuadrática dada por $\boldsymbol{hH(r)h}$, $\boldsymbol{h} = (s_1 - s_{01}, s_2 - s_{02})$, $\boldsymbol{H(r)}$ es la matriz hessiana de Ψ evaluada en el punto $\boldsymbol{r}$, que está en algún punto entre (s_1, s_2) y (s_{01}, s_{02}).
Luego evaluando la ecuación (2.17) en cada uno de los centroides $\boldsymbol{s} = [\boldsymbol{s_1}, \boldsymbol{s_2}] = [s_{i1}, s_{i2}]$, $\forall i = 1, \ldots, n$, se tiene el vector

$$\begin{aligned}\boldsymbol{u} = \Psi(\boldsymbol{s}) &= (\lambda_0 - s_{01} - s_{02})\boldsymbol{1} + \lambda_1 \boldsymbol{s_1} + \lambda_2 \boldsymbol{s_2} + R(\boldsymbol{s_1}, \boldsymbol{s_2}, s_{01}, s_{02}) \\ &= [\boldsymbol{1}, \boldsymbol{s_1}, \boldsymbol{s_2}]\boldsymbol{\gamma} + \boldsymbol{R}\end{aligned}$$

donde $\boldsymbol{\gamma}$ es la gradiente de $\boldsymbol{\Psi}$ evaluada en cada uno de los puntos de referencia del mapa.
Luego el predictor lineal quedará definido por

$$\begin{aligned}\boldsymbol{X\beta} + \boldsymbol{u} &= \boldsymbol{X\beta} + [1, s_1, s_2]\boldsymbol{\gamma} + \boldsymbol{R}, \\ &= \boldsymbol{X\beta} + \boldsymbol{P}[\boldsymbol{1}, \boldsymbol{s_1}, \boldsymbol{s_2}]\boldsymbol{\gamma} + \boldsymbol{P}^\perp[\boldsymbol{1}, \boldsymbol{s_1}, \boldsymbol{s_2}]\boldsymbol{\gamma} + \boldsymbol{R},\end{aligned}$$

y el componente $\boldsymbol{P}[\boldsymbol{1}, \boldsymbol{s_1}, \boldsymbol{s_2}]\boldsymbol{\gamma}$ se removerá en el modelo SPOCK.

2.4. Inferencia bayesiana

De acuerdo a la definición realizada en Hoff (2010) la inferencia bayesiana hace referencia al proceso inductivo de aprendizaje a través del teorema de Bayes. El teorema de Bayes se define de la siguiente manera: Sean dos eventos A y B:

$$p(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

A modo general, se podría decir que si se tuviera alguna creencia previa sobre un evento A, el teorema de Bayes provee un método racional para actualizar esta probabilidad a la luz de nueva información.
Formalmente, como se menciona en Gelman et al. (2004a), el método bayesiano se puede definir en los siguientes tres pasos:

- Definir un modelo probabilístico para todas las cantidades observables (que se obtendrán en la muestra a realizarse) y no observables (parámetros o variables latentes) del problema. Ello implica que, a diferencia de la inferencia clásica, los parámetros también tendrán una distribución de probabilidad. La notación usual que se asigna al vector de parámetros de interés es θ y, por lo tanto, su distribución a priori se denota

$p(\theta)$.

- Sea $\boldsymbol{y}^\intercal = (y_1, y_2, \ldots, y_n)$ una muestra observada de la población en estudio y en base al teorema de Bayes, calcular la distribución a posteriori de las cantidades no observables condicionadas a la muestra observada, $p(\theta|\boldsymbol{y})$.
- Evaluar el ajuste del modelo y, si fuera necesario, alterar o expandir el modelo, y repetir los 3 pasos nuevamente.

El teorema de Bayes en combinación con los 3 pasos antes mencionados generan la siguiente expresión para la distribución a posteriori

$$p(\theta|\boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{\int p(\boldsymbol{y}|\theta)p(\theta)d\theta}.$$

Dado que usualmente θ es una variable continua se ha reemplazado la sumatoria del teorema de Bayes, definido, por la integral. Además, dado que en el denominador, el vector de observaciones $\boldsymbol{y}$ es fijo, no depende de θ, se podría considerar constante y, por lo tanto, la expresión para la distribución a posteriori quedaría expresada de la siguiente manera:

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

2.4.1. El método de las cadenas de Markov de Monte Carlo (MCMC)

Muchas veces la distribución a posteriori puede presentar una forma no conocida, por lo que suele ser complicado obtener estimaciones exactas sobre el vector de parámetros $\boldsymbol{\theta}$ en estudio; sin embargo, gracias al método de simulaciones de Monte Carlo, se puede estimar estas cantidades de interés. También ocurre que podría ser complicado realizar las simulaciones de Monte Carlo por lo que el método se combina con las denominadas cadenas de Markov.
Simular cadenas de Markov implica que en lugar de simular de valores independientes de la distribución a posteriori se obtendrá una muestra de una cadena de Markov que es una colección de valores dependientes del valor inmediato anterior cuya distribución estacionaria es la distribución a posteriori. Esta propiedad de independencia se representa con la siguiente expresión.

$$p(\theta^{(t)}|\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t-1)}) = p(\theta^{(t)}|\theta^{(t-1)}), \tag{2.18}$$

donde $\theta^{(1)}, \ldots, \theta^{(n)}$ son variables aleatorias que ocurren en un tiempo $t = 1, 2, \ldots, n$. Las propiedades que deben satisfacer la cadena de Markov para que la distribución alcance la estacionariedad pueden revisarse en Robert y Casella (2004)
Los algoritmos basados en MCMC más usados son el muestreo de Gibbs y el algoritmo Metropolis-Hastings;sin embargo, para el presente trabajo se utiliza el método de Monte Carlo Hamiltoniano (HMC) que es utilizado por el software STAN.

2.4.2. El muestreo de Gibbs

De acuerdo a Gelman et al. (2004b) este algoritmo supone que una vez definida la distribución a posteriori del vector de parámetros en estudio $\boldsymbol{\theta}$, este se dividirá en sus componentes.

$$\boldsymbol{\theta}^{\intercal} = (\theta_1, \theta_2, \ldots, \theta_m),$$

donde m será el número de componentes de $\boldsymbol{\theta}$. Basándose en el MCMC definido en la Sección 2.4.1, el algoritmo, en cada iteración t, genera un valor de cada uno de los componentes usando la distribución de dicho componente condicionado a los valores de todos los demás. Luego en cada iteración habrán m pasos correspondientes a cada uno de los componentes. El orden para generar un valor de cada componente en cada iteracion se escoge convenientemente y tomando la probabilidad

$$p(\theta_j|\boldsymbol{\theta}_{-j}^{(t-1)}, \boldsymbol{y}), \tag{2.19}$$

donde $\boldsymbol{\theta}_{-j}^{t-1}$ representa todos los componentes de $\boldsymbol{\theta}$ excepto por θ_j en sus valores actualizados dados por

$$\boldsymbol{\theta}_{-j}^{(t-1)} = (\theta_1^{(t)}, \ldots, \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, \ldots, \theta_m^{(t-1)}).$$

Como ya se ha mencionado, de cada componente θ_j se genera un valor de la distribución condicional completa denotada $p(\theta_j|\boldsymbol{\theta}_{-j}, \boldsymbol{y})$ donde $\boldsymbol{\theta}_{-j}$ presenta los valores actuales de acuerdo a la iteración t y al paso j del muestreo. Cuando las distribuciones condicionales completas son distribuciones conocidas por lo que es sencillo simular de ellas.

2.4.3. El algoritmo Metropolis y Metropolis-Hastings

De acuerdo a Gelman et al. (2004b) el algoritmo de Metropolis basado también en el MCMC utiliza una regla de aceptación y rechazo para que se pueda converger a la distribución objetivo. Para este algoritmo se parte de un valor inicial digamos $\theta^{(0)}$ a partir de una distribución inicial tal que $p(\theta^{(0)}|\boldsymbol{y}) > 0$. Luego para cada iteración t se realiza lo siguiente:

- Se genera un valor de una distribución propuesta $J_t(\theta^*|\theta^{t-1})$.
- Se calcula la razón

$$r = \frac{p(\theta^*|\boldsymbol{y})}{p(\theta^{(t-1)}|\boldsymbol{y})}$$

- Se actualiza $\theta(t)$ tal que:

$$\theta^{(t)} = \begin{cases} \theta^* & \text{con probabilidad min(r,1)} \\ \theta^{(t-1)} & \text{de otro modo.} \end{cases},$$

Por lo tanto, en cada iteración si el valor generado por la distribución de probabilidad propuesta aumenta la densidad a posteriori de $\theta^{(t-1)}$ entonces $\theta^{(t)} = \theta^*$ y en caso

contrario $\theta^{(t)} = \theta^{(t-1)}$.

El algoritmo Metropolis-Hastings es una generalización del Algoritmo de Metropolis en la que la distribución propuesta no requiere ser simétrica y en la que la razón r se calcula de la siguiente manera:

$$r = \frac{p(\theta^*|\boldsymbol{y})/J_t(\theta^*|\theta^{(t-1)})}{p(\theta^{(t-1)}|\boldsymbol{y})/J_t(\theta^{(t-1)}|\theta^*)}.$$

2.4.4. Método de Monte Carlo Hamiltoniano

El método llamado Monte Carlo Hamiltoniano (MHC) recibe su nombre porque está basado en la definición de la función Hamiltoneana la cual describe la energía total de un sistema cerrado a partir de la energía potencial $U(\theta)$ y la energía cinética $C(\phi)$

$$H(\theta, \phi) = U(\theta) + C(\phi),$$

por lo tanto la dinámica Hamiltoniana es definida por el sistema de ecuaciones diferenciales

$$\frac{d\theta}{dt} = \frac{dH}{d\phi} = \Delta_\phi C(\phi)$$
$$\frac{d\phi}{dt} = -\frac{dH}{d\theta} = \Delta_\theta U(\theta),$$

las cuales suelen obtenerse mediante métodos numéricos (Neal, 2012).
El método MHC trata de sobrellevar el hecho de las ineficiencias que pueden existir debido a que en el algoritmo de Metropolis la distribución propuesta (usualmente una normal multivariada) se centra en la posición actual, la cual podría estar en las colas de la distribución a posteriori o solo se ocupe de una parte de la posteriori y no de otra. El MHC presenta mayor flexibilidad pues cambia dependiendo de la posición actual para lo cual utiliza el gradiente de la distribución a posteriori de manera que la distribución propuesta se basa en este gradiente. Para el MHC la probabilidad de aceptación no solo toma en cuenta la distribución a posteriori relativa también el momento entre las posiciones actual y la propuesta (https://tereom.github.io/est-computacional-2019/hmc-y-stan.html).

2.4.5. Modelos gaussianos latentes

De acuerdo a la definición mencionada en Blangiardo y Cameletti (2015), sea Y una variable aleatoria en análisis y sean los valores observados $\boldsymbol{y}^\intercal = (y_1, y_2, \ldots, y_n)$, la distribución de Y_i presenta el parámetro κ_i , el cual podría ser la media o algún cuantil. Se define la siguiente estructura aditiva para el predictor lineal $g(\kappa_i) = \eta_i$,

$$\eta_i = \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{l=1}^{L} f_l(z_{li}),$$

donde $\boldsymbol{\beta}^\intercal = (\beta_1, \ldots, \beta_M)$ son los coeficientes de regresión (o efectos fijos) de las covariables $\boldsymbol{x}^\intercal = (1, x_2, \ldots, x_M)$ y $\boldsymbol{f}^\intercal = \{f_1(\cdot), \ldots, f_L(\cdot)\}$ es una colección de funciones definidas en términos de un conjunto de covariables $\boldsymbol{z}^\intercal = (z_1, \ldots, z_L)$. Los términos de $f_l(\cdot)$ pueden

asumir diferentes formas suaves, efectos no lineales, efectos temporales o espaciales. Se tiene entonces el conjunto de parámetros $\boldsymbol{\theta}^\intercal = \{\boldsymbol{\beta}, \boldsymbol{f}\}$, los cuales son campos aleatorios de Markov gaussianos definidos en la Sección 2.3.2 y presentan a su vez una distribución definida por el conjunto de hiperparámetros $\boldsymbol{\psi} = \{\psi_1, \ldots, \psi_k\}$.
Por lo tanto el campo gaussiano $\boldsymbol{\theta}$ tendrá la siguiente distribución:

$$\boldsymbol{\theta} \ \sim \ N(\mathbf{0}, \boldsymbol{Q}^{-1}(\boldsymbol{\Psi})),$$

donde $\boldsymbol{Q}(\boldsymbol{\Psi})$ es una matriz dispersa (con gran cantidad de valores 0 en sus elementos) lo que conlleva a una mayor eficiencia computacional.

2.5. Criterios de evaluación de modelos

Para evaluar la bondad de un modelo en el enfoque bayesiano entre los criterios a utilzar se tiene el WAIC (Criterio de información Watanable-Akaike), el LPML (Logaritmo de la verosimilitud pseudo marginal) y el RMSE (Raiz cuadrada del error cuadrático medio).
El WAIC presentado por Watanabe (2010) es definido como sigue:

$$\text{WAIC} = -2\left(\sum_{i=1}^{n} \log E_{\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y}}\Big(p(y_i|\boldsymbol{\theta},\boldsymbol{\psi})\Big) - Pw \right),$$

donde Pw es el término que penaliza y que está dado por

$$Pw = -2\sum_{i=1}^{n}\left(E_{\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y}}\Big(\log p(y_i|\boldsymbol{\theta},\boldsymbol{\psi})\Big) - \sum_{i=1}^{n} \log E_{\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y}}\Big(p(y_i|\boldsymbol{\theta},\boldsymbol{\psi})\Big)\right),$$

$$E_{\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y}}\Big(p(y_i|\boldsymbol{\theta},\boldsymbol{\psi})\Big) \approx \frac{1}{M}\sum_{j=1}^{M} p(y_i|\boldsymbol{\theta}^{(j)},\boldsymbol{\psi}^{(j)}),$$

y

$$E_{\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y}}\Big(\log p(y_i|\boldsymbol{\theta},\boldsymbol{\psi})\Big) \approx \frac{1}{M}\sum_{j=1}^{M} \log p(y_i|\boldsymbol{\theta}^{(j)},\boldsymbol{\psi}^{(j)}),$$

donde $\boldsymbol{\theta}^{(j)}$ y $\boldsymbol{\psi}^{(j)}$ son las simulaciones a posteriori de $p(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y})$.
El LPML es definido como la suma de los logaritmos de la ordenada condicional predictiva (CPO_i) (Geisser y Eddy, 1979) que está definida como

$$CPO_i = p(y_i|y_{-i}) = \Big(\int\int \frac{p(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y})}{p(y_i|\boldsymbol{\theta},\boldsymbol{\psi})}\delta\boldsymbol{\theta}\delta\boldsymbol{\psi}\Big)^{-1} = E_{\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y}}\Big(\frac{1}{p(y_i|\boldsymbol{\theta},\boldsymbol{\psi})}\Big),$$

donde n es el número de observaciones e y_i es el valor de la i-esima observación. Luego la estimación del CPO_i (Dey et al., 1997) queda definida por

$$\widehat{CPO_i} = \frac{1}{M}\sum_{j=1}^{M} \frac{1}{p(y_i|\boldsymbol{\theta}^{(j)},\boldsymbol{\psi}^{(j)})},$$

donde $\boldsymbol{\theta}^{(j)}$ y $\boldsymbol{\psi}^{(j)}$ son las simulaciones a posteriori de $p(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y})$. La estimación del LPML es dada por $\widehat{\text{LPML}} = \sum_{i=1}^{n} \log \widehat{\text{CPO}_i}$.

El RMSE es definido como

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

donde $i = 1, \ldots, n$, y_i es el valor observado e $\hat{y}_i$ es la estimación para la i-ésima observación.

Capítulo 3

Modelo de regresión cuantílica para datos de áreas

En este capítulo se presenta la definición del modelo de regresión cuantílica para proporciones en datos de áreas, la función de verosimilitud, la distribución a posteriori, los supuestos asumidos para las distribuciones a priori y los criterios de comparación del modelo.

3.1. Definición del modelo

Sea $\boldsymbol{Y}^\intercal = (Y_1, ..., Y_n)$ el vector de variables aleatorias independientes las cuales siguen una distribución Kumaraswamy, con parámetros κ_i, ϕ_i y q,

$$Y_i \sim Kumar(\kappa_i, \phi_i, q), i = 1, \ldots, n,$$

la cual, como se ha mencionado en la Sección 2.2.1, es una reparametrización de la representación usual de la variable aleatoria Y_i con distribución Kumaraswamy de parámetros (α, β) en la que κ_i, es el cuantil q de Y_i, ϕ_i es el parámetro de precisión, la fdp queda definida por

$$f_{Y_i}(y_i|\kappa_i, \phi_i) = -\frac{\log(1-q)\phi_i}{\log(1-e^{-\phi_i})\log(\kappa_i)} y_i^{-\frac{\phi_i}{\log(\kappa_i)}-1}\{1-y_i^{-\frac{\phi_i}{\log(\kappa_i)}}\}^{\frac{\log(1-q)}{\log(1-e^{-\phi_i})}-1},$$

y

$$E(Y_i) = \frac{\log(1-q)}{\log(1-e^{-\phi_i})} B\left(1-\frac{\log(\kappa_i)}{\phi_i}, \frac{\log(1-q)}{\log(1-e^{-\phi_i})}\right)$$

donde $B(.,.)$ denota a la función beta, $0 < \kappa_i < 1$, $0 < \phi$, $Y_i \in (0, 1)$ representa el valor de la variable aleatoria en el área s_i para $i = 1, \ldots, n$ para n áreas en estudio.
Los parámetros κ_i y ϕ_i son asociados a las covariables mediante las siguientes funciones de enlace:

$$g_1(\kappa_i) = \boldsymbol{X_i}^\intercal \boldsymbol{\beta} + u_i,$$

$$g_2(\phi_i) = \boldsymbol{W_i}^\intercal \boldsymbol{\delta},$$

donde $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_P)^\intercal$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_L)^\intercal$ son los vectores de los coeficientes de regresión asociados a κ_i y ϕ_i respectivamente; $\boldsymbol{u} = (u_1, \ldots, u_n)^\intercal$ son los efectos aleatorios con estructura espacial donde cada u_i está asociado a cada κ_i; $X_i = (x_{i,1}, \ldots, x_{i,P})^\intercal$ y $W_i = (w_{i,1}, \ldots, w_{i,L})^\intercal$ son los vectores de covariables; $g_1(\cdot)$ y $g_2(\cdot)$ son funciones de enlace que para el presente trabajo serán consideradas como $g_1(\cdot) = \text{logit}(\cdot)$, la función de enlace logística, y $g_2(\cdot) = \log(\cdot)$, la función de enlace logarítmica.

Con respecto a los efectos aleatorios espaciales $u_1, \ldots, u_n$ se asumirá

$$\boldsymbol{u} \sim N_n(\mathbf{0}, \boldsymbol{Q^{-1}}), \tag{3.1}$$

donde $\boldsymbol{Q^{-1}} = D_\tau(\boldsymbol{I} - \alpha \boldsymbol{B}))$, $\boldsymbol{D_\tau} = \tau \boldsymbol{D}$, $\boldsymbol{D}$ es la matriz diagonal (N_i), donde como ya se mencionó N_i es el número de vecinos del área i-ésima, $\boldsymbol{I}$ es la matriz identidad, α es un parámetro que controla la dependencia espacial ($\alpha = 0$ indica independencia y $\alpha = 1$ indica un modelo ICAR), $\boldsymbol{B} = \boldsymbol{D}^{-1}\boldsymbol{W}$ es la matriz de vecindad escalada, y $\boldsymbol{W}$ es la matriz de vecindad cuyos componentes son definidos por

$$w_{ij} = \begin{cases} 1 & \text{, si } i \sim j (i \text{ es vecino de } j), \\ 0 & \text{, si } i \text{ no es vecino de } j \end{cases}$$

Por lo tanto, dadas las distribuciones gaussianas independientes para $\boldsymbol{u}$, $\boldsymbol{\beta}$ y $\boldsymbol{\delta}$, entonces $\boldsymbol{\theta}^\intercal = (\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\delta})$ también sigue una distribución gaussiana multivariada definida por

$$\boldsymbol{\theta} \ \sim \ N_k(\mathbf{0}, \boldsymbol{Q^{-1}(\psi)}), \tag{3.2}$$

donde $\boldsymbol{\psi}^\intercal = (\phi, \tau, \alpha)$ es el vector de hiperparámetros y $\boldsymbol{Q}(\boldsymbol{\psi})$ es una matriz dispersa (con gran cantidad de valores 0 en sus componentes) lo que conlleva a una mayor eficiencia computacional. Definiendo $\boldsymbol{\theta}^\intercal = (\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\delta})$, $\boldsymbol{\psi}^\intercal = (\tau, \alpha)$ e $\boldsymbol{Y}^\intercal = (Y_1, ..., Y_n)$, la función de verosimilitud para el modelo puede ser escrita como sigue:

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) &= p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= f_{Y_1}(y_1|\boldsymbol{\theta_1}, \boldsymbol{\psi_1}) \times f_{Y_2}(y_2|\boldsymbol{\theta_2}, \boldsymbol{\psi_2}) \times \ldots \times f_{Y_n}(y_n|\boldsymbol{\theta_n}, \boldsymbol{\psi_n}) \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i|\boldsymbol{\theta_i}, \boldsymbol{\psi_i}) \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i|\kappa_i, \phi_i), \end{aligned} \tag{3.3}$$

donde $\boldsymbol{\kappa}^\intercal = \kappa_1, \ldots, \kappa_n$, y $\phi = \phi_1, \ldots, \phi_n$ son vectores de parámetros definidos por

$$\kappa_i = \frac{1}{1 + \exp(\boldsymbol{X_i^\intercal}\boldsymbol{\beta} + u_i)}, \tag{3.4}$$

$$\phi_i = \exp(\boldsymbol{w_i^\intercal}\boldsymbol{\delta}), \tag{3.5}$$

y $f_{Y_i}(y_i|\kappa_i, \phi_i)$ es la función de densidad de probabilidad de una variable aleatoria con distribución Kumaraswamy.

3.2. Inferencia bayesiana

La fdp a posteriori para $\boldsymbol{\theta}^\intercal = (\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\delta})$ y $\boldsymbol{\psi}^\intercal = (\tau, \alpha)$ denotada como $p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y})$ se define como sigue:

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\psi}) \times p(\boldsymbol{\theta}|\boldsymbol{\psi}) \times p(\boldsymbol{\psi})}{p(\boldsymbol{Y})},$$

donde $p(\boldsymbol{Y})$ no depende de $\boldsymbol{\theta}$ por lo tanto

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

y puede también ser expresada como

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) \propto L(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) \times p(\boldsymbol{\theta}|\boldsymbol{\psi}) \times p(\boldsymbol{\psi}), \tag{3.6}$$

donde $p(\boldsymbol{\theta}|\boldsymbol{\psi})$ es la distribución condicional de $\boldsymbol{\theta}|\boldsymbol{\psi}$ y $p(\boldsymbol{\psi})$ es la distribución a priori de $\boldsymbol{\psi}$. Para el presente documento se asume independencia entre $\boldsymbol{u}$, $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ y entre τ y α por lo que se puede tener la distribución a priori

$$p(\boldsymbol{\theta}|\boldsymbol{\psi}) \times p(\boldsymbol{\psi}) = p(\boldsymbol{u}|\tau, \alpha) \times p(\boldsymbol{\beta}) \times p(\boldsymbol{\delta}) \times p(\tau) \times p(\alpha), \tag{3.7}$$

donde $p(\boldsymbol{\beta}) = p(\beta_1)p(\beta_2)\dots p(\beta_p) = \prod_{j=1}^{p} p(\beta_j)$, y $p(\boldsymbol{\delta}) = p(\delta_1)p(\delta_2)\dots p(\delta_l) = \prod_{k=1}^{l} p(\delta_k)$. Para los coeficientes $\boldsymbol{\beta}$ se asume que $\beta_j \sim N(0, 10^2), j = 1, \dots, p$, donde p es el número de covariables del cuantil κ. Para los coeficientes δ_k de la precision ϕ se asume una distribución $\delta_k \sim N(0, 10^2), k = 1, \dots, l$, donde l es es número de covariables de ϕ. Por lo tanto, $p(\beta_j) = 1/(\sqrt{200\pi})e^{-\frac{1}{200}\beta_j^2}$ para $j = 1, \dots, p$, y $p(\delta_k) = 1/(\sqrt{200\pi})e^{-\frac{1}{200}\delta_k^2}$ para $k = 1, \dots, p$.
Para $\boldsymbol{u}$, como ya se ha definido en la ecuación (3.1), se asume la distribución

$$\boldsymbol{u} \sim N_n(\boldsymbol{0}, \boldsymbol{Q_u^{-1}}). \tag{3.8}$$

Tomando en cuenta que se tienen los campos aleatorios de Markov gaussianos (GMRF) independientes $\boldsymbol{u}$, $\boldsymbol{\beta}$ y $\boldsymbol{\delta}$ entonces $\boldsymbol{\theta}^\intercal = (\boldsymbol{u}, \boldsymbol{\beta}$ y $\boldsymbol{\delta})$ dado el conjunto de hiperparámetros es una familia GMRF definida de la siguiente manera

$$\boldsymbol{\theta}|\boldsymbol{\psi} \sim \ N_k(\boldsymbol{0}, \boldsymbol{Q^{-1}(\psi)}), \tag{3.9}$$

donde $\boldsymbol{Q(\psi)}$ es una matriz dispersa definida por:

$$\boldsymbol{Q}(\psi) = \begin{bmatrix} \boldsymbol{Q_\beta} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Q_\delta} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Q_u} \end{bmatrix},$$

con

$$\boldsymbol{Q_\beta} = \begin{bmatrix} \tau_{\beta_1} & 0 & \cdots & 0 & 0 \\ 0 & \tau_{\beta_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & \tau_{\beta_{p-1}} & 0 \\ 0 & 0 & \cdots & 0 & \tau_{\beta_p} \end{bmatrix},$$

$$\boldsymbol{Q_\delta} = \begin{bmatrix} \tau_{\delta_1} & 0 & \cdots & 0 & 0 \\ 0 & \tau_{\delta_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & \tau_{\delta_{l-1}} & 0 \\ 0 & 0 & \cdots & 0 & \tau_{\delta_l} \end{bmatrix},$$

donde $\tau_{(\cdot)} = 1/\sigma^2$ es la precisión de cada parámetro. También $\boldsymbol{Q_u}$ como ya se ha mencionado presenta gran cantidad de ceros en sus elementos dada su definición en función de la matriz de vecindad. Para los hiperparámetros de $\boldsymbol{\psi}^\intercal = (\tau \text{ y } \alpha)$ se asumirán a priori las siguientes distribuciones

$$\tau \sim \text{gamma}(4, 0.5),$$

$$\alpha \sim \text{Uniforme}(0, 1).$$

Considerando la función de verosimilitud definida en la ecuación (3.3) y las distribuciones a priori definidas anteriormente la fdp a posteriori definida en la ecuación (3.6) puede ser expresada como:

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) &\propto \prod_{i=1}^{n} f_{Y_i}(y_i|\kappa_i, \phi_i) \times p(\boldsymbol{u}|\tau, d) \times p(\boldsymbol{\beta}) \times p(\boldsymbol{\delta}) \times p(\tau) \times p(\alpha) \\ &\propto \prod_{i=1}^{n} f_{Y_i}(y_i|\kappa_i, \phi_i) \times p(\boldsymbol{\theta}|\boldsymbol{\psi}) \times p(\boldsymbol{\psi}), \end{aligned}$$

donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ con $K = (n+p+l)$ indica el tamaño del campo aleatorio gaussiano $\boldsymbol{\theta}$ formado por $\boldsymbol{u}$, $\boldsymbol{\beta}$ y $\boldsymbol{\delta}$; y cuya distribución, dados los hiperparámetros $\boldsymbol{\psi}$, se definió en (3.9). Por lo tanto

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) \propto \prod_{i=1}^{n} f_{Y_i}(y_i|\kappa_i, \phi_i) \times \frac{|\boldsymbol{Q}(\boldsymbol{\psi})|^{\mathbf{1/2}}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\theta^t \boldsymbol{Q}(\boldsymbol{\psi})\theta \right) \times p(\boldsymbol{\psi}),$$

donde $|\boldsymbol{Q}(\boldsymbol{\psi})|$ es el determinante de $\boldsymbol{Q}(\boldsymbol{\psi})$. Además, para los hiperparámetros $p(\boldsymbol{\psi})$, tomando en cuenta las distribuciones a priori definidas se tiene que

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) \propto \prod_{i=1}^{n} f_{Y_i}(y_i|\kappa_i, \phi_i) \times \frac{|\boldsymbol{Q}(\boldsymbol{\psi})|^{\mathbf{1/2}}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\theta^t \boldsymbol{Q}(\boldsymbol{\psi})\theta \right) \times \tau_u^3 \exp\left(-0.5\tau_u\right).$$

Para estimar el modelo propuesto se utilizará la librería RSTAN (https://mc-stan.org/users/interfaces/stan) que se encuentra incluida en el entorno de programación R. La estimación es realizada mediante los algoritmos basados en MCMC y en el algoritmo HMC descrito en la Sección 2.4.4.

Capítulo 4

Estudio de simulación

En este capítulo se muestra un estudio de simulación del modelo de regresión cuantílica Kumaraswamy para datos de áreas presentado en el Capítulo 3.
El estudio consiste en la generación de los efectos espaciales, luego en la generación de los datos con distribución Kumaraswamy fijando los parámetros para tres escenarios que corresponden a los cuantiles $q = 0.1$, $q = 0.5$ y $q = 0.9$. Por último se presenta el análisis de la recuperación de los parámetros en cada escenario.

4.1. Generación de los efectos espaciales

En el contexto de datos de áreas en estadística espacial, se consideran $n = 56$ áreas. Para cada una se tiene una medida Y_i que será la variable respuesta en la i-ésima área.
Se genera la matriz de covariables $\boldsymbol{x}^\intercal = [\boldsymbol{x_0}, \boldsymbol{x_1}]$ donde $\boldsymbol{x_0} = (1, \ldots, 1)^\intercal$ es un vector de n elementos iguales a 1 y $\boldsymbol{x_1}$ es simulada de una variable aleatoria con distribución normal con media 0 y varianza 1.
Para la simulación de los efectos espaciales se ha utilizado el método descrito en Rue y Held (2005). Los efectos se construyen a partir de un grafo que representa a las áreas vecinas y consiste en un conjunto de aristas y vértices en donde las aristas representan a cada área y los vértices unen a las áreas vecinas.
Usando este grafo se construye la matriz de vecindad $\boldsymbol{W}$ representada gráficamente en la Figura 4.1. Esta representación gráfica consiste en una matriz donde cada fila representa a cada una de las 56 áreas y cada columna también representa a cada una de ellas. En esta matriz se colorean las celdas (i, j) en caso las áreas i-ésima y j-ésima sean vecinas, y también colorea la diagonal de la matriz. Para la simulación de este efecto espacial por área se define $\boldsymbol{u}^\intercal = (u_1, \ldots, u_n)$ donde u_i es el efecto espacial del área i-ésima, entonces la distribución conjunta para $\boldsymbol{u}$ tiene la distribución mencionada en (3.1) la cual es la siguiente:

$$\boldsymbol{u} \sim N_n(\boldsymbol{0}, \boldsymbol{Q_u^{-1}}), \tag{4.1}$$

donde $\boldsymbol{Q_u^{-1}} = \tau_{\boldsymbol{u}}(\boldsymbol{D} - \alpha \boldsymbol{W})$, $\boldsymbol{D}$ es la matriz diagonal (N_i), α es un parámetro que controla la dependencia espacial tal que $\alpha = 0$ implica independencia espacial mientras que $\alpha = 1$ conlleva a un modelo condicional autoregresivo intrínseco (ICAR) y $\boldsymbol{W}$ es la matriz de

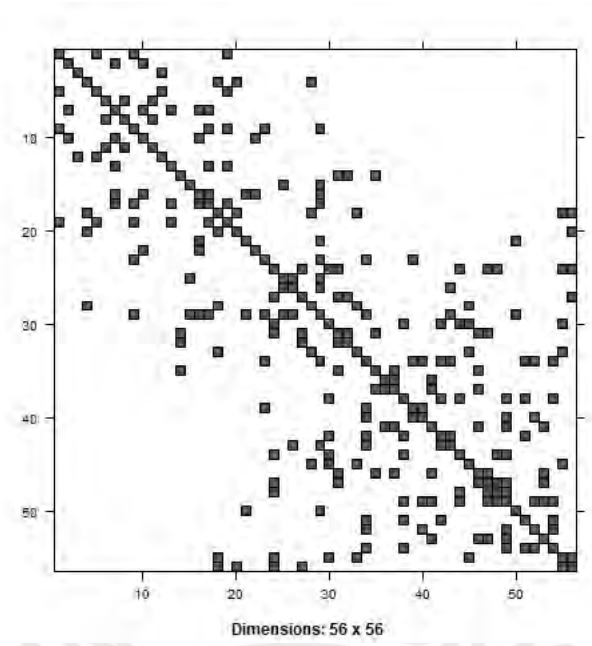


Figura 4.1: Matriz de vecindad a utilizarse para generar los efectos espaciales del modelo propuesto

vecindad, cuyos componentes son definidos por

$$w_{ij} = \begin{cases} 1 & \text{, si } i \sim j \text{ } (i \text{ es vecino de } j), \\ 0 & \text{, si } i \text{ no es vecino de } j \end{cases}$$

Para cada uno de los 3 escenarios se establecen 4 subescenarios para cada uno de los cuales se fijará el parámetro $\tau_{\boldsymbol{u}}$ en 2 y 7 respectivamente y α en 0.9.
Una vez obtenida la matriz de precisión $\boldsymbol{Q}$ para $\boldsymbol{u}$, dado que esta es dispersa (presenta gran cantidad de ceros), la matriz triangular L obtenida de la factorización de Cholesky $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^{\intercal}$, también es dispersa. Se realiza una simulación de 56 valores $\boldsymbol{z}^{\intercal} = (z_1, \ldots, z_n)$ de una variable aleatoria $Z \sim N(0, \boldsymbol{I})$ y se resuelve el sistema $\boldsymbol{L}^{\intercal}\boldsymbol{u} = \boldsymbol{z}$. Los valores de $\boldsymbol{u}$ contienen los 56 valores simulados para los efectos aleatorios con estructura espacial.

4.2. Generación de los datos

Se incorporan covariables solo al parámetro κ de la regresión cuantílica. Para cada uno de los cuatro subescenarios se fijaran los mismos valores para los coeficientes de los efectos fijos $\boldsymbol{\beta}^{\intercal} = (\beta_0, \beta_1)$ en $(0.5, -1)$. Para el parámetro de precisión, ϕ se asignan los valores 5, 10, 20 y 26 para cada subescenario respectivamente. Como se mencionó, los efectos espaciales se simulan en las áreas a partir de la matriz de vecindad, fijando el parámetro $\tau_{\boldsymbol{u}}$ en 0.1, 2 y 7 respectivamente en cada subescenario y α en 0.9. Este conjunto de parámetros de cada subescenario se repite para cada uno de los tres escenarios. Cada escenario, como ya se ha mencionado, tiene respectivamente el valor de $q = 0.1, 0.5, 0.9$ para el nivel del cuantil κ. Los subescenarios se presentan en el Cuadro 4.1

Escenarios: q=0.1, q=0.5 , q=0.9						
	Kumar			CAR		
Subesc.	β_0	β_1	ϕ	τ	α	Priori de ϕ
1	0.5	-1	5	2	0.9	$\phi \sim \text{Gamma}(2, 2)$
2	0.5	-1	10	2	0.9	$\phi \sim \text{Gamma}(4, 0.5)$
3	0.5	-1	20	7	0.9	$\phi \sim \text{Gamma}(4, 0.5)$
4	0.5	-1	25	7	0.9	$\phi \sim \text{Gamma}(4, 0.5)$

Cuadro 4.1: Parámetros establecidos para cada escenario-subescenario

4.3. Recuperación de parámetros

La estimación de los parámetros se realiza a través del algoritmo HMC usando RStan. En el Apéndice B.1, se puede observar los gráficos de cadenas de las simulaciones a posteriori de cada parámetro e hiperparámetro. Se han realizado 4 cadenas de 1000 iteraciones en cada escenario-subescenario con un burning de 500, las cuales han mostrado una convergencia aceptable.

A modo general, se ha recuperado los parámetros e hiperparámetros fijados con un tiempo de ejecución bajo y con estimaciones para la variable respuesta aproximadas a sus valores reales.

Los Cuadros 4.2, 4.3 y 4.4 muestran un resumen de los valores recuperados para los parámetros en cada uno de los escenarios-subescenarios. Se puede observar que en cada uno se ha recuperado los parámetros e hiperparámetros del modelo. La primera columna muestra el valor que se ha fijado para cada parámetro en cada escenario, las siguientes columnas muestran las estimaciones de la media y desviación estándar a posteriori que corresponden a la esperanza y la desviación estándar de la distribución a posteriori de cada parámetro respectivamente. Se puede observar que la media de los parámetros se aproxima a su valor real y la desviación estándar es baja en cada uno de ellos. El intervalo de credibilidad (IC) presenta los cuantiles 2.5 % y 97.5 % de las simulaciones realizadas a posteriori. Se observa que en todos los escenarios los IC de los parámetros contienen al valor real del parámetro.

En las Figuras 4.2, 4.3 y 4.4, se observa los histogramas de los subescenarios 1 y 2 de las simulaciones a posteriori de cada parámetro e hiperparámetro. Las líneas en azul corresponden a los intervalos de credibilidad (IC) al 95 % mientras que la línea roja indica el valor real. El resto de histogramas de todos los escenarios y subescenarios se encuentran en el apéndice B.2.

Con respecto a la recuperación de los efectos espaciales, en las Figuras 4.5, 4.6 y 4.7 se verifica que los valores de las estimaciones se encuentran cercanos a los valores reales simulados en los tres escenarios.

Además en las Figuras 4.8 hasta 4.10, se muestran los intervalos de credibilidad de los efectos espaciales mediante barras. Se muestran 20 de las áreas para no sobrecargar el gráfico. La barra azul gruesa en cada subescenario muestra el IC al 50 % de los efectos espaciales mientras las líneas delgadas corresponden al IC al 90 %. Se puede apreciar que muchos de los efectos espaciales estimados han resultado significativos en cada escenario-subescenario, como era de predecirse dado que los valores reales son distintos de 0.

Asimismo, también se ha realizado la estimación de las observaciones correspondientes

Subescenario	Parámetro	Real	Media	Desv. est.	95 % IC
1	β_0	0.50	0.56	0.16	0.24 ; 0.88
	β_1	-1.00	-0.89	0.11	-1.12 ; -0.68
	ϕ	5.00	5.17	0.69	4.05 ; 6.89
	τ	2.00	1.77	0.65	0.84 ; 3.34
	α	0.90	0.67	0.20	0.21 ; 0.95
2	β_0	0.50	0.55	0.15	0.26 ; 0.88
	β_1	-1.00	-0.92	0.07	-1.06 ; -0.79
	ϕ	10.00	11.75	3.13	7.05 ; 19.41
	τ	2.00	2.13	0.52	1.29 ; 3.27
	α	0.90	0.81	0.13	0.51 ; 0.98
3	β_0	0.50	0.48	0.09	0.31 ; 0.66
	β_1	-1.00	-0.98	0.04	-1.06 ; -0.89
	ϕ	20.00	16.93	3.87	11.08 ; 26.16
	τ	7.00	5.04	1.13	3.15 ; 7.45
	α	0.90	0.75	0.15	0.38 ; 0.96
4	β_0	0.50	0.45	0.08	0.29 ; 0.63
	β_1	-1.00	-0.98	0.04	-1.07 ; -0.90
	ϕ	25.00	17.09	3.74	11.42 ; 25.96
	τ	7.00	5.22	1.13	3.31 ;7.66
	α	0.90	0.76	0.15	0.40 ; 0.97

Cuadro 4.2: Valor real, media, desviación estándar e intervalo de credibilidad (IC) al 95 % de los parámetros para el primer escenario $q = 0.1$.

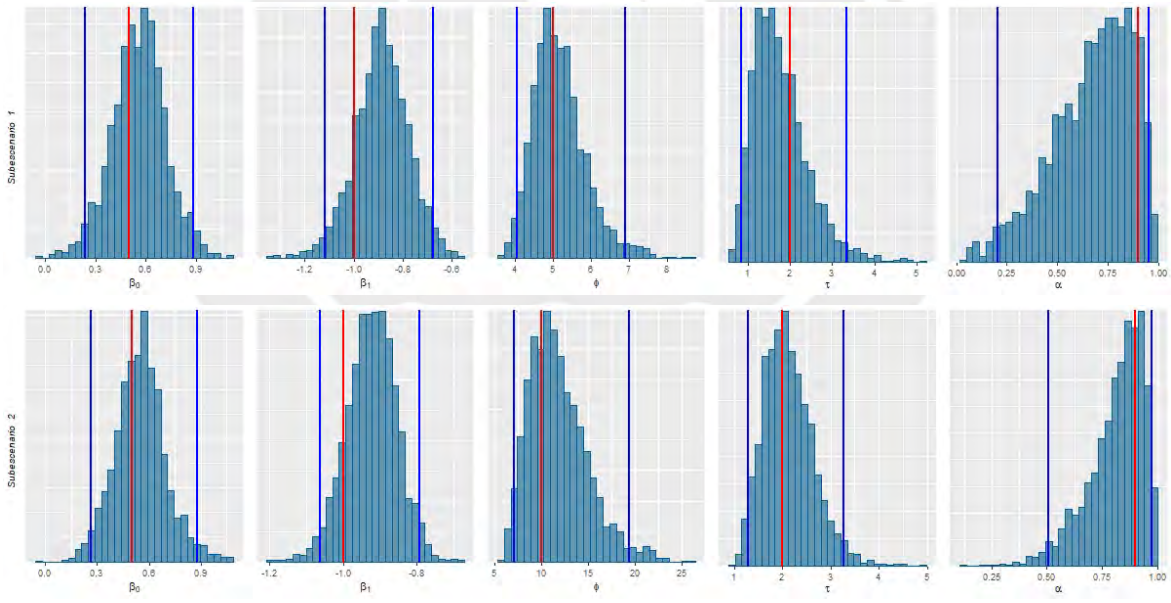


Figura 4.2: Histograma de las simulaciones a posteriori de los parámetros para el primer escenario $q = 0.1$. La línea en rojo representa el valor real mientras las líneas azules de los extremos representan el intervalo de credibilidad al 95 %.

a los datos simulados para la variable Y_i . En los Cuadros 4.5, 4.6 y 4.7 se muestra una comparación de los resultados de cada escenario tomando como criterios el RMSE definido en 2.5 además del tiempo de procesamiento de cada estimación.

Los valores obtenidos para el RMSE en cada escenario son bajos y ello indica un buen ajuste en cada uno de ellos. El tiempo de ejecución de los tres escenarios es rápido a pesar que se

Subescenario	Parámetro	Real	Media	Desv. est.	95 % IC
1	β_0	0.50	0.50	0.13	0.23 ; 0.74
	β_1	-1.00	-1.03	0.07	-1.16 ; -0.89
	ϕ	5.00	4.56	0.93	3.02 ; 6.59
	τ	2.00	2.12	0.66	1.14 ; 3.76
	α	0.90	0.77	0.16	0.36 ; 0.98
2	β_0	0.50	0.52	0.10	0.33 ; 0.76
	β_1	-1.00	-1.00	0.05	-1.11 ; -0.92
	ϕ	10.00	11.77	3.68	6.25 ; 19.83
	τ	2.00	2.24	0.53	1.38 ;3.41
	α	0.90	0.82	0.12	0.50 ; 0.97
3	β_0	0.50	0.51	0.07	0.37 ; 0.67
	β_1	-1.00	-1.01	0.03	-1.08 ; -0.94
	ϕ	20.00	16.37	3.98	10.11 ; 25.55
	τ	7.00	5.20	1.13	3.29 ; 7.60
	α	0.90	0.75	0.15	0.41 ; 0.97
4	β_0	0.50	0.50	0.08	0.33 ; 0.67
	β_1	-1.00	-1.01	0.03	-1.08 ; -0.94
	ϕ	25.00	16.34	4.02	9.76 ; 25.74
	τ	7.00	5.34	1.16	3.31 ; 7.90
	α	0.90	0.76	0.16	0.38 ; 0.97

Cuadro 4.3: Valor real, media, desviación estándar e intervalo de credibilidad (IC) al 95 % de los parámetros para el segundo escenario $q = 0.5$.

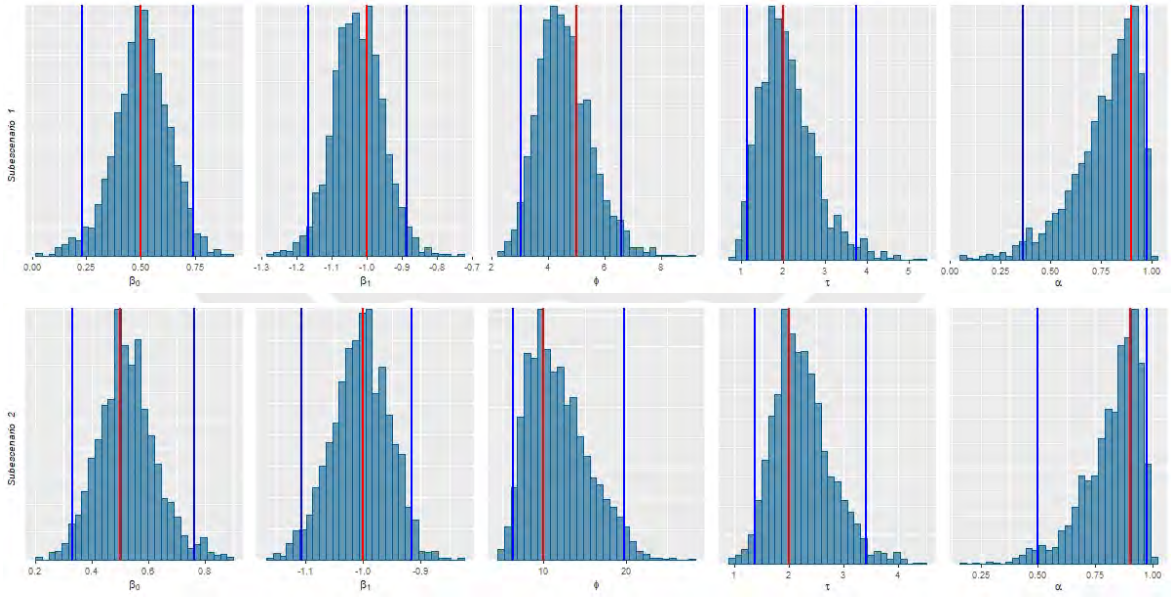


Figura 4.3: Histograma de las simulaciones a posteriori de los parámetros para el segundo escenario $q = 0.5$. La línea en rojo representa el valor real mientras las líneas azules de los extremos representan el intervalo de credibilidad al 95 %.

ha tenido una gran cantidad de párametros a simular.

Subescenario	Parámetro	Real	Media	Desv. est.	95 % IC
1	β_0	0.50	0.61	0.17	0.32 ; 1.02
	β_1	-1.00	-1.07	0.08	-1.23 ; -0.93
	ϕ	5.00	3.53	0.86	2.10 ; 5.36
	τ	2.00	2.32	0.70	1.22 ; 4.05
	α	0.90	0.79	0.15	0.40 ; 0.98
2	β_0	0.50	0.48	0.17	-0.01 ; 0.78
	β_1	-1.00	-1.07	0.06	-1.18 ; -0.95
	ϕ	10.00	10.46	3.60	5.06 ; 18.79
	τ	2.00	2.16	0.53	1.30 ; 3.31
	α	0.90	0.83	0.12	0.54 ; 0.98
3	β_0	0.50	0.51	0.08	0.36 ; 0.65
	β_1	-1.00	-1.04	0.04	-1.11 ; -0.96
	ϕ	20.00	15.55	4.17	9.14 ; 25.52
	τ	7.00	5.10	1.10	3.14 ; 7.54
	α	0.90	0.75	0.15	0.41 ; 0.96
4	β_0	0.50	0.53	0.09	0.38 ; 0.72
	β_1	-1.00	-1.03	0.04	-1.10 ; -0.95
	ϕ	25.00	16.06	4.59	8.64 ; 25.96
	τ	7.00	5.27	1.18	3.31 ; 8.00
	α	0.90	0.77	0.14	0.42 ; 0.97

Cuadro 4.4: Valor real, media, desviación estándar e intervalo de credibilidad (IC) al 95 % de los parámetros para el tercer escenario $q = 0.9$.

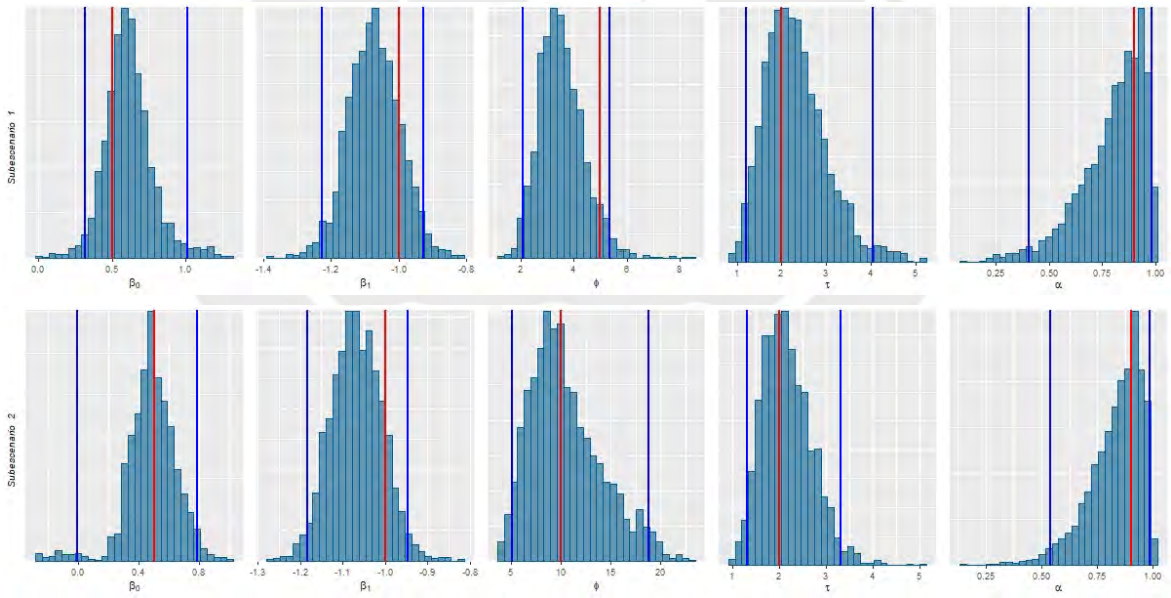


Figura 4.4: Histograma de las simulaciones a posteriori de los parámetros para el tercer escenario $q = 0.9$. La línea en rojo representan el valor real mientras las líneas azules de los extremos representa el intervalo de credibilidad al 95 %.

q=0.1	Criterios	
Subescenario	RMSEE	Tiempo (seg.)
1	0.05	35.04
2	0.01	56.45
3	0.01	61.00
4	0.01	66.53

Cuadro 4.5: Evaluación del primer escenario $q = 0.1$.

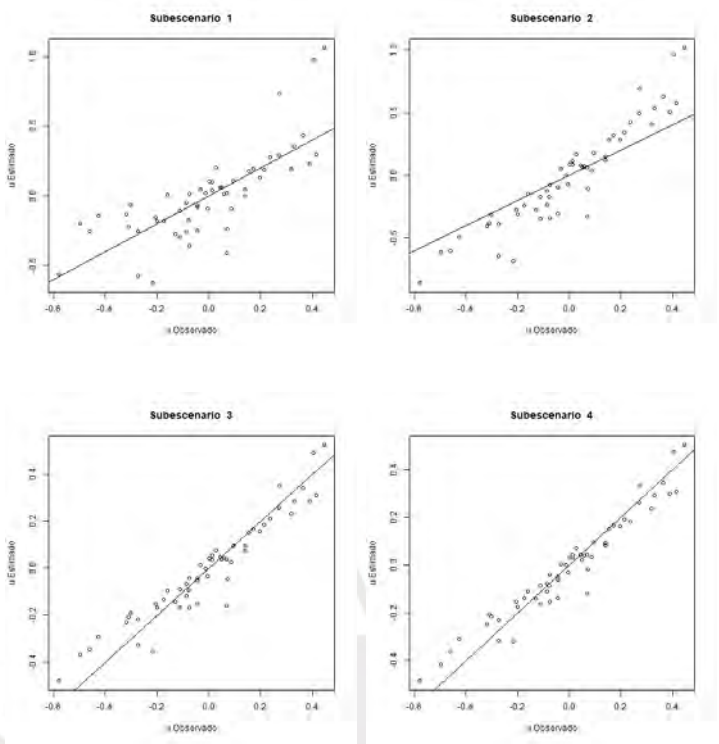


Figura 4.5: Diagramas de dispersión de los efectos espaciales reales y estimados en cada subescenario para el primer escenario $q = 0.1$.

q=0.5	Criterios	
Subescenario	RMSEE	Tiempo (seg.)
1	0.04	33.54
2	0.01	65.96
3	0.01	52.04
4	0.01	55.93

Cuadro 4.6: Evaluación del segundo escenario $q = 0.5$.

q=0.9	Criterios	
Subescenario	RMSEE	Tiempo (seg.)
1	0.04	33.32
2	0.01	52.27
3	0.01	54.87
4	0.01	55.16

Cuadro 4.7: Evaluación del tercer escenario $q = 0.9$.

Las Figuras 4.8 y 4.13 muestran diagramas de dispersión correspondientes a las estimaciones de la media a posteriori de $\boldsymbol{Y}$, así como diagramas de dispersión comparando las estimaciones y los valores simulados de $\boldsymbol{Y}$.

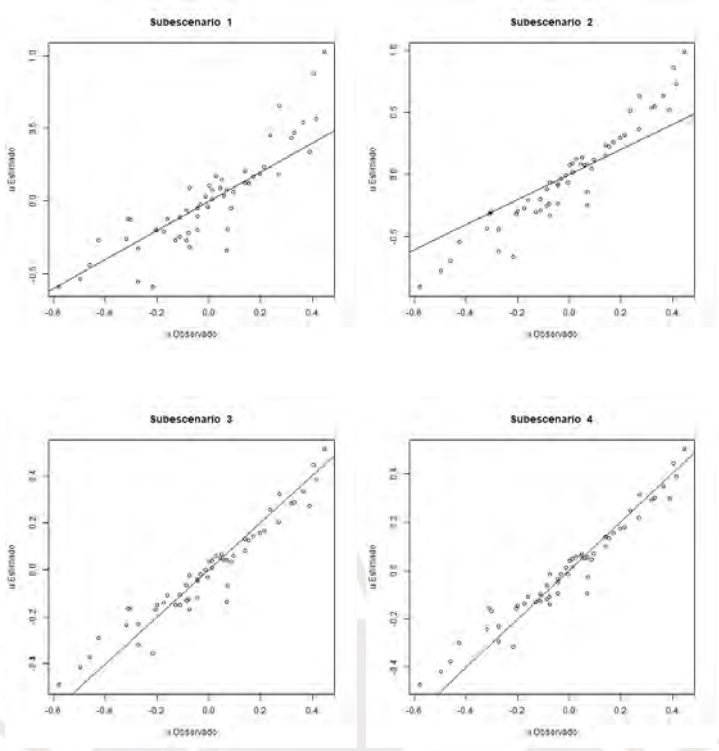


Figura 4.6: Diagramas de dispersión de los efectos espaciales reales y estimados en cada subescenario para el segundo escenario $q = 0.5$.

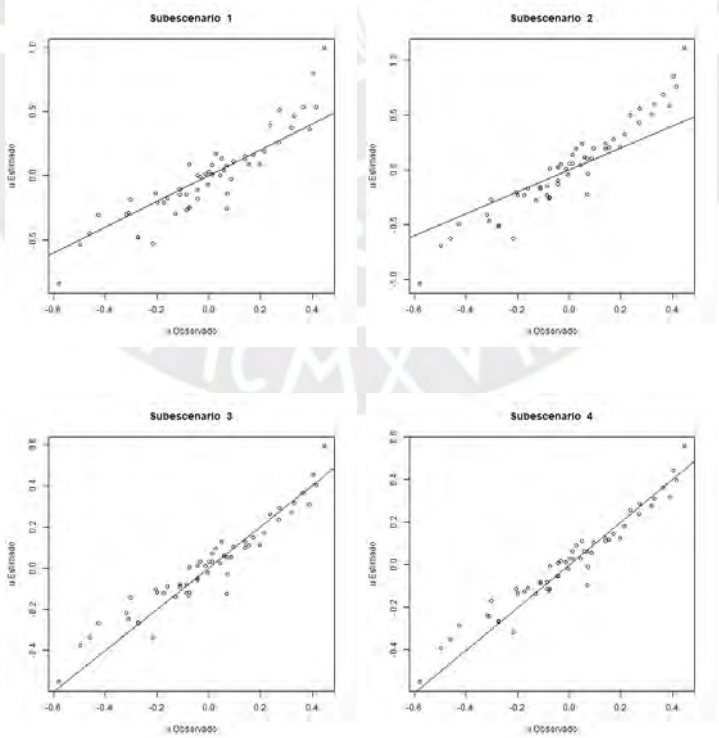


Figura 4.7: Diagramas de dispersión de los efectos espaciales reales y estimados en cada subescenario para el tercer escenario $q = 0.9$.

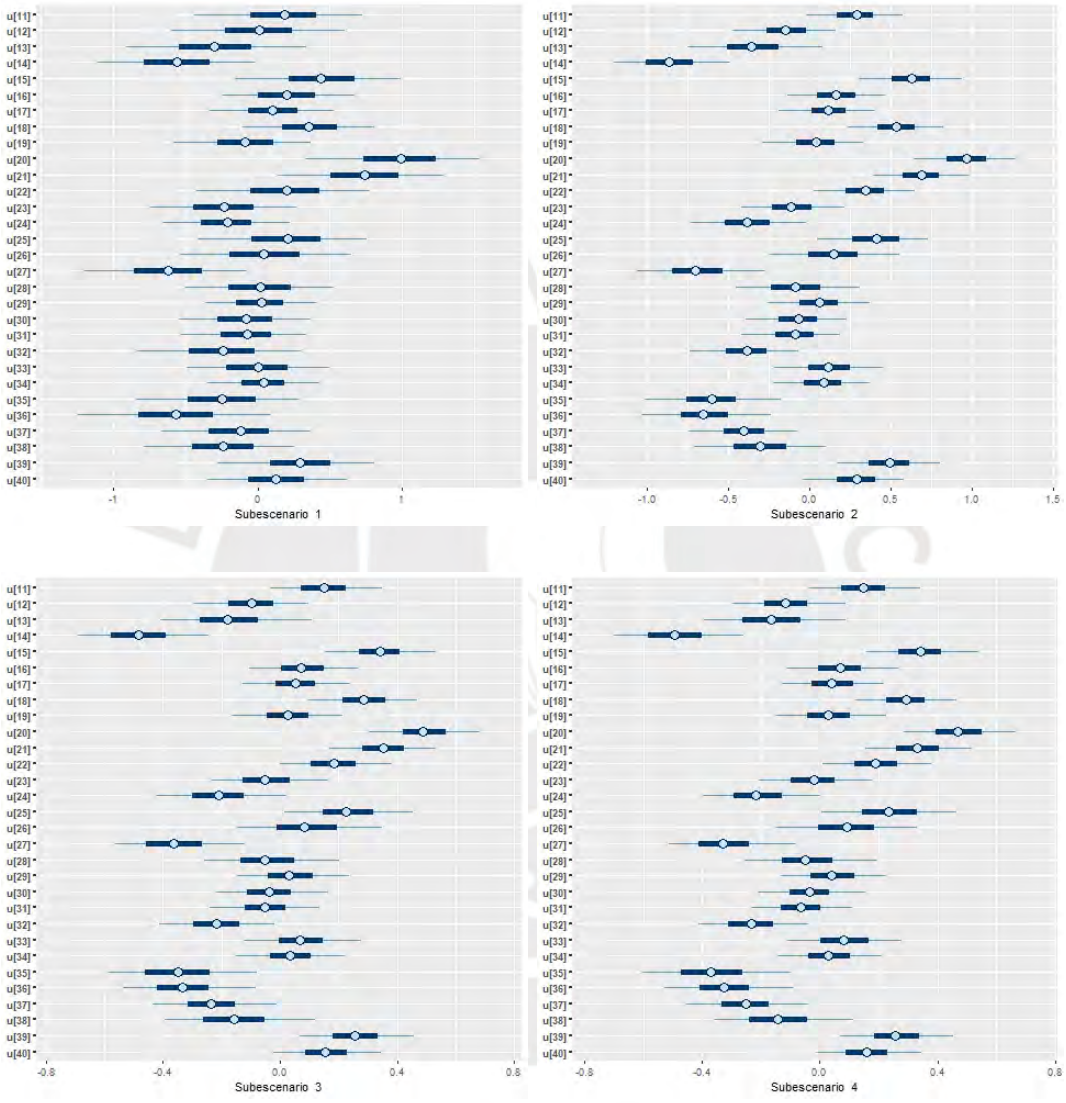


Figura 4.8: Intervalos al 50 % (barra azul gruesa) y al 90 % (línea delgada) de los efectos espaciales en cada subescenario para el primer escenario $q = 0.1$.

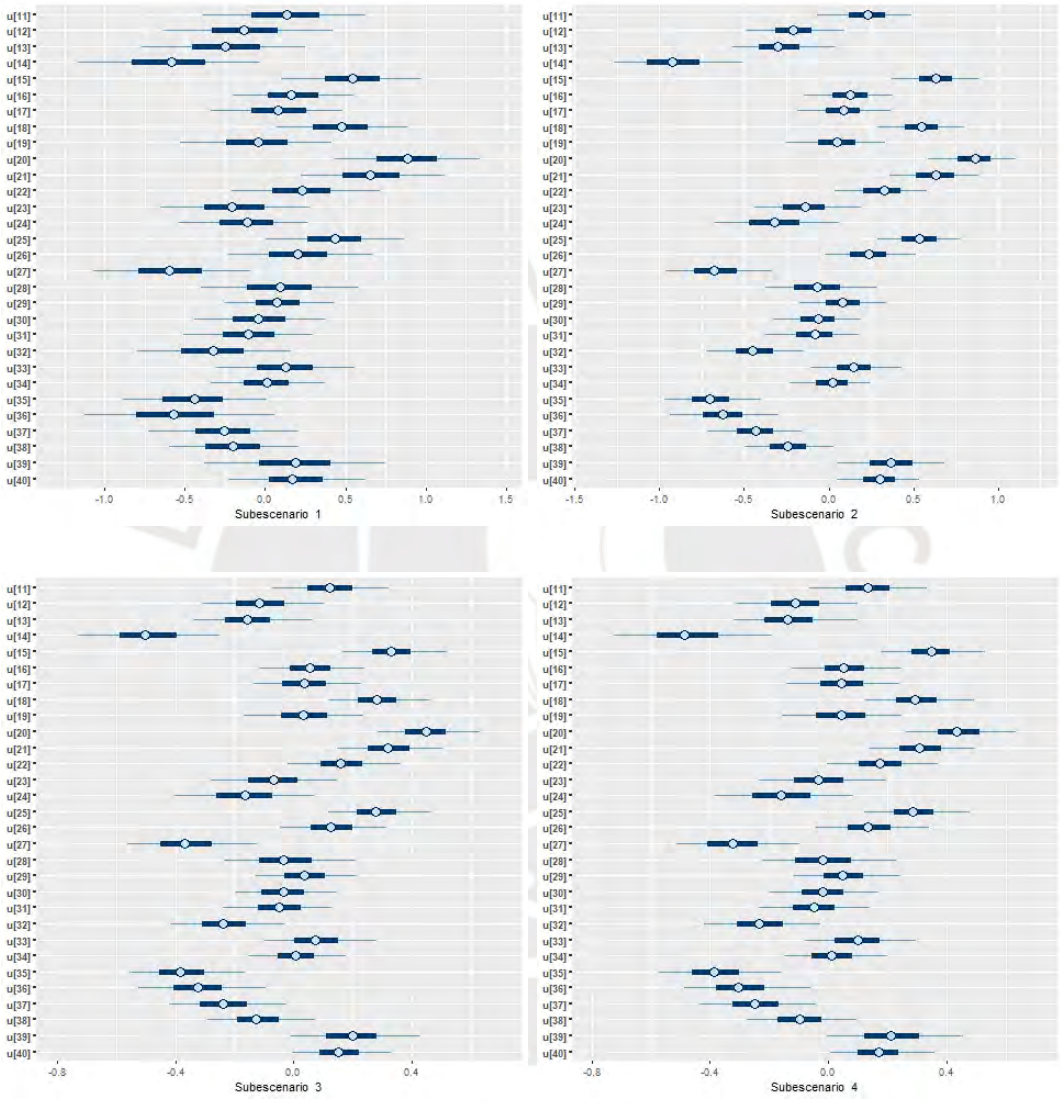


Figura 4.9: Intervalos al 50 % (barra azul gruesa) y al 90 % (línea delgada) de los efectos espaciales en cada subescenario para el segundo escenario $q = 0.5$.

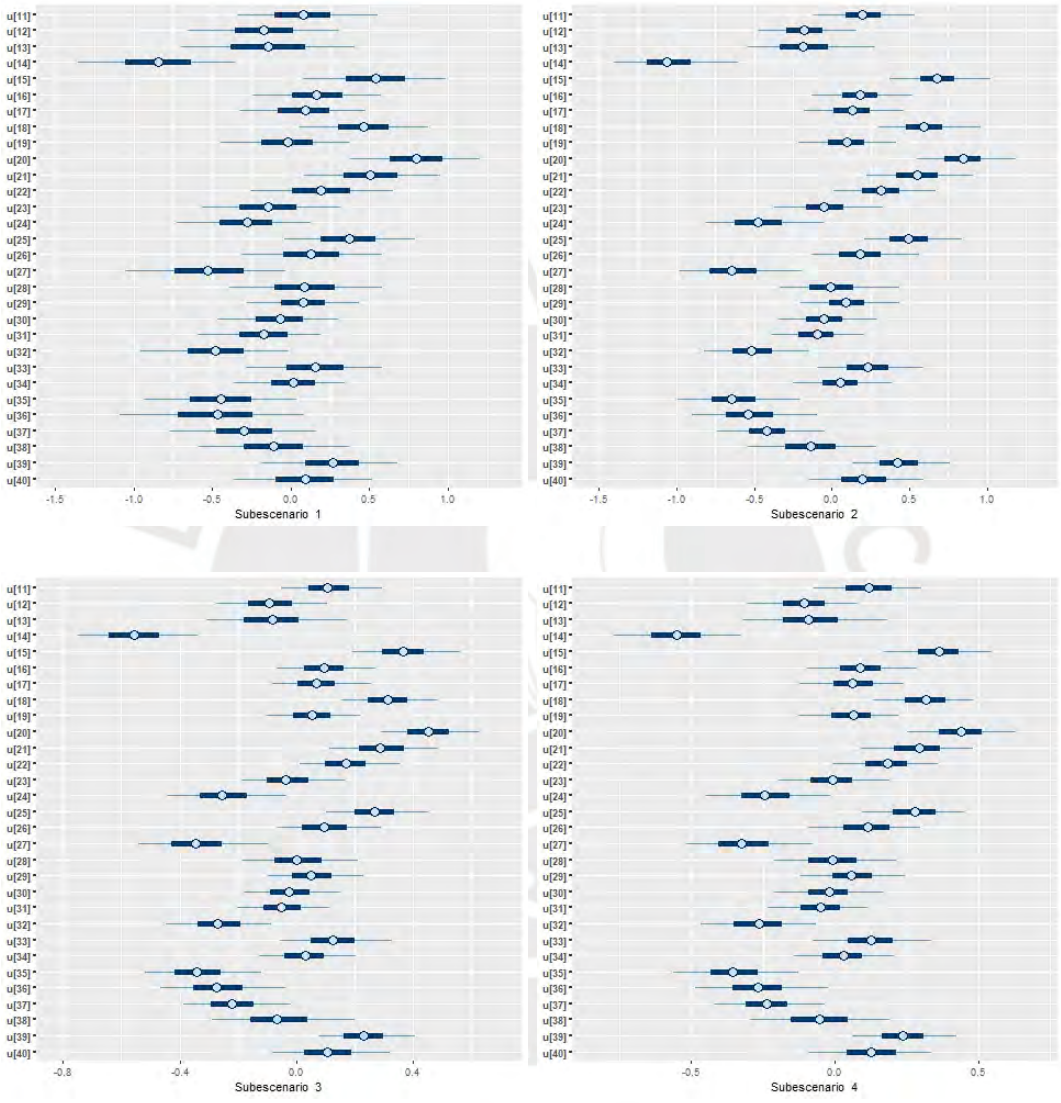


Figura 4.10: Intervalos al 50 % (barra azul gruesa) y al 90 % (línea delgada) de los efectos espaciales en cada subescenario para el tercer escenario $q = 0.9$.

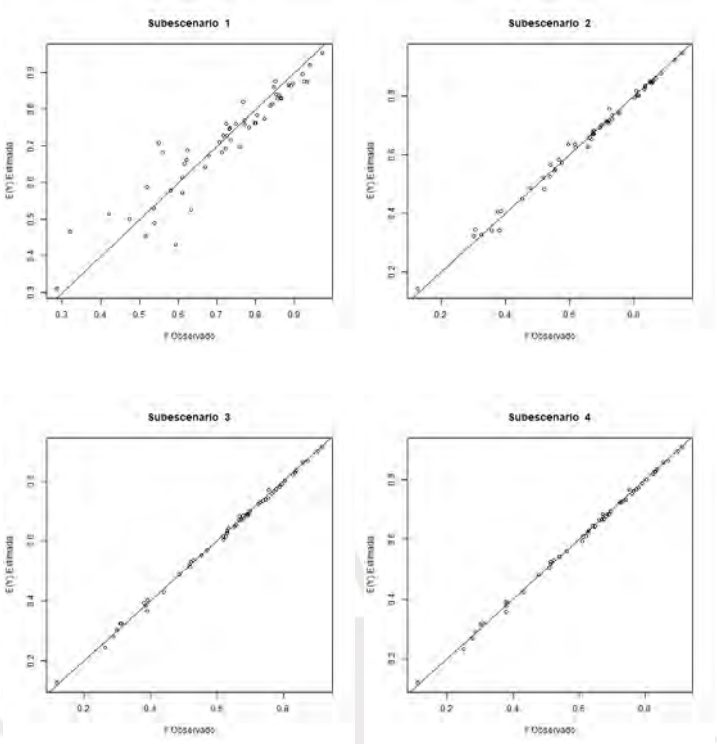


Figura 4.11: Diagramas de dispersión de las simulaciones y las estimaciones de la variable respuesta para el primer escenario $q = 0.1$.

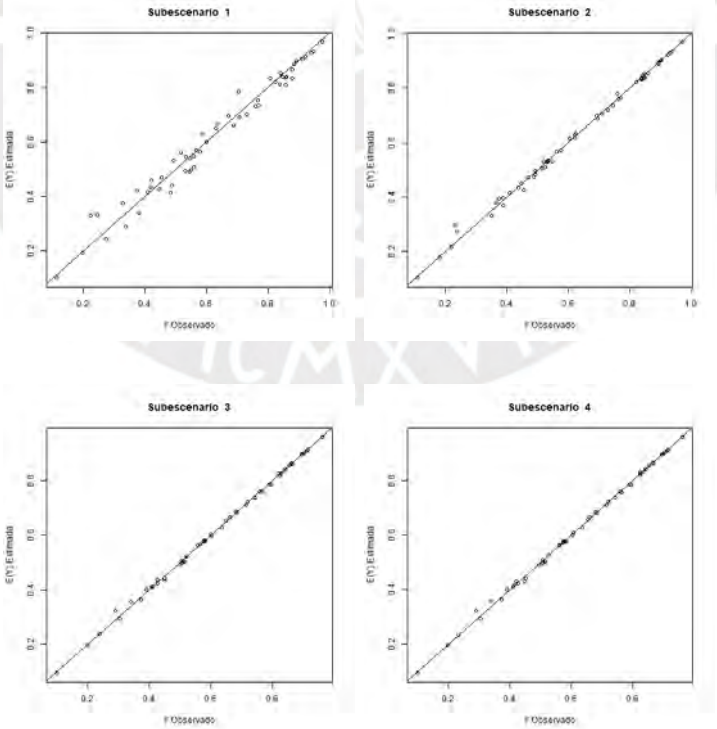


Figura 4.12: Diagramas de dispersión de las simulaciones y las estimaciones de la variable respuesta para el segundo escenario $q = 0.5$.

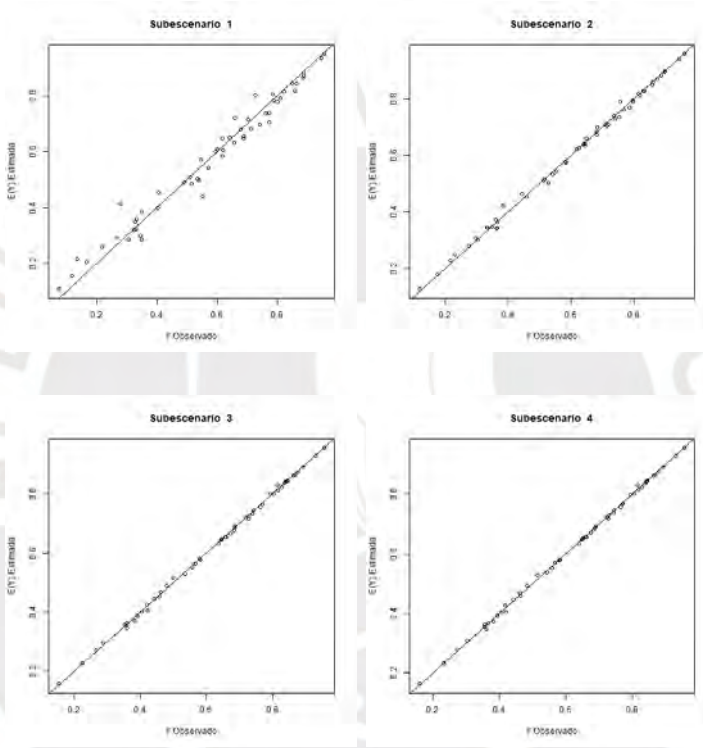


Figura 4.13: Diagramas de dispersión de las simulaciones y las estimaciones de la variable respuesta para el tercer escenario $q = 0.9$.

Capítulo 5

Aplicación a la incidencia de pobreza no extrema

En este capítulo se presentan los resultados obtenidos de la aplicación del modelo espacial cuantílico para datos de áreas entre (0,1) presentado en el Capítulo 3. Además, se realiza una comparación estadística con un modelo de regresión cuantílica que no incluye efecto espacial. El objetivo principal de la aplicación es estudiar el efecto de un grupo de covariables y de la estructura espacial en los cuantiles de la incidencia de pobreza no extrema en los distritos de la provincia de Lima. Con esta información se busca determinar cuantitativamente factores espaciales y no espaciales relacionados a la incidencia de la pobreza no extrema en los distritos de la provincia de Lima.

5.1. Incidencia de pobreza no extrema

5.1.1. Definición e importancia

La definición de pobreza puede ser realizada desde distintos enfoques y hace referencia a la situación de carencia de bienes y servicios materiales para desempeñarse en una sociedad. Principalmente, existen dos enfoques para determinar si una persona es pobre (Feres y Mancero, 2015).

El enfoque monetario determina la condición de pobreza mediante el cálculo de la línea de pobreza. Esta representa el monto suficiente para adquirir una canasta básica de alimentos y necesidades básicas no alimenticias tales como vestido, vivienda, agua, luz o educación. A aquellas personas que no pueden cubrir una canasta básica se les denomina pobres extremos, mientras que a los pobres que alcanzan la canasta básica pero no cubren sus necesidades básicas no alimenticias se les denomina pobres no extremos (https://www.mef.gob.pe/es/mapas-de-pobreza/metodos-para-medir-la-pobreza).

El enfoque no monetario para determinar la condición de pobreza toma en cuenta las condiciones de vida de una persona sin la necesidad de tomar en cuenta el monto del ingreso o gasto. Por ejemplo, a través de censos o encuestas se puede determinar si una persona tiene acceso a sus necesidades básicas.

Un enfoque que toma en cuenta tanto el monto del ingreso y el gasto así como las necesidades insatisfechas observadas puede reconocer mejor a los pobres dentro de una población. Una vez se ha definido la condición de pobreza, de acuerdo a Foster et al. (1984) se puede utilizar los siguientes indicadores de medición: La incidencia que es la proporción de la población cuyo consumo se encuentra por debajo del valor de la línea de pobreza, la brecha que mide la insuficiencia promedio del consumo de los pobres respecto de la línea de pobreza y la severidad que mide la desigualdad entre los pobres (https://www.inei.gob.pe/media/

MenuRecursivo/publicaciones_digitales/Est/Lib1370/cap03.pdf).
Sin duda, el seguimiento de la pobreza en un país permite tomar decisiones y políticas pertinentes que permitan mejorar la capacidad adquisitiva y la calidad de vida de los hogares de un país.

5.1.2. La incidencia de pobreza en el Perú

Según la evaluación del Banco Mundial y las cifras del Instituto Nacional de Estadística e Informática (INEI) la incidencia de pobreza en el Perú ha caido rápidamente entre el 2002 y 2013. Sin embargo, debido a la desaceleración económica entre el 2013 y 2019 no se ha podido reducir la pobreza en los niveles esperados, incluso, como se observa en la Figura 5.1 entre el 2016 y 2018 no se aprecia mejoría, por ello, es importante estudiar los factores que podrían influir en su comportamiento.

Figura 5.1: Evolución de la incidencia de pobreza en el Perú. Fuente: INEI

Actualmente, la elaboración de indicadores sociales referidos a pobreza monetaria están a cargo del INEI. Entre estos se incluye la incidencia de pobreza cuyo último cálculo corresponde al año 2018 (INEI, 2018a).
Los estudios sobre la incidencia de pobreza en el Perú se iniciaron a mediados de la década de 1980, cuando el Banco Central de Reserva del Perú (BCR) elaboró el primer mapa de pobreza adoptando el enfoque de Necesidades Básicas Insatisfechas (NBI)(Banco Central de Reserva, 1986). Un mapa de pobreza, por ejemplo, a nivel distrital identifica las características de pobreza por distrito de un país. Este enfoque ha sido adoptado por el INEI, sin embargo, debido a que en el Perú los censos nacionales no incluyen preguntas sobre ingresos o gastos, para la elaboración de mapas de pobreza con enfoque monetario, se utiliza la información de la encuesta nacional de hogares que se realiza periodicamente. Este metodología ha sido utilizada por el Banco Mundial en otros países como Ecuador (Hentschel y Lanjouw, 1996),

Madagascar y Mozambique (Elbers, Lanjouw, Mistiaen, Ozler y Simler, 2003). En nuestro país la primera aplicación fue realizada en 1996 (INEI, 1996) y luego en el año 2001 por el Ministerio de Economía y Finanzas (MEF, 2001). Luego, en el año 2003, el INEI publicó el mapa de pobreza 2001 (Kuiper, 2003).

En el 2007, el INEI, en colaboración con el Banco Mundial, creó un comité especializado conformado por expertos en temas de pobreza provenientes de diversos sectores en donde, por ejemplo, en el sector acádemico se incluye a la PUCP. A partir de ese año hasta la fecha, el INEI publica el informe técnico: "Evolución de la pobreza monetaria". En este informe se presentan indicadores sobre la evolución de las líneas de pobreza total y extrema e indicadores de brecha y severidad de la pobreza monetaria, así como las principales características de la población y de los hogares en pobreza (INEI, 2019).

En el 2010 se publicó un nuevo mapa de pobreza (INEI, 2010) siguiendo la metodología propuesta por los investigadores del Banco Mundial (Elbers, Lanjouw y Lanjouw, 2003). Siguiendo la misma metodología, en el 2015, se publicó un nuevo mapa debido a los importantes cambios socioeconómicos y la reducción de la pobreza que ocurrieron desde el 2009 (INEI, 2015).

5.2. Descripción de los datos

Los datos fueron obtenidos del mapa de pobreza publicado por el INEI en el año 2009 (INEI, 2010) y del mapa de necesidades básicas insatisfechas 1993, 2007 y 2017 publicado por el INEI en el año 2018 (INEI, 2018b).

Se considera como variable dependiente a la incidencia de pobreza no extrema y como covariables a la incidencia de población que se encuentran en viviendas con características físicas inadecuadas y a la incidencia de hogares en viviendas con hacinamiento.

Para considerar una vivienda inadecuada, el mapa de necesidades básicas insatisfechas (INEI, 2018b) tomó en cuenta el material de la construcción. Por ejemplo, aquellos hogares que habitan en viviendas de estera, quincha, piedra con barro, madera, cartón, lata, ladrillos y adobes superpuestos.

Para considerar una vivienda con hacinamiento se tiene a aquellas con más de 4 personas por habitación. Los problemas que conlleva una alta densidad, son la insalubridad y una alta incidencia de problemas entre sus ocupantes.

Cada observación de la incidencia de pobreza no extrema corresponde a cada uno de los 43 distritos correspondientes a la división política de la provincia de Lima en el año 2007. Las variables fueron calculadas por el INEI utilizando, como fuentes, el Censo Nacional del 2007 y la Encuesta Nacional de Hogares (ENAHO) del 2007.

En la Figura 5.2 se muestra el mapa de la incidencia de pobreza no extrema por distrito en la provincia de Lima en donde se observa una fuerte correlación espacial la cual también se validó calculando los índices de Moran y Geary los cuales son 0.519 y 0.459 respectivamente. El mapa nos permite apreciar que conforme los distritos se encuentran más alejados de la capital la incidencia de pobreza tiende a aumentar. En la Figura 5.2, al lado derecho se presenta el histograma de la incidencia de pobreza no extrema en donde se tiene una tendencia a 0.15 y la leve asimetría a la derecha indica que la mayor parte de los distritos tienen la incidencia en un nivel debajo de 0.15. En la Figura 5.3 se tiene la distribución de la incidencia

de viviendas en mal estado en donde se aprecia que la gran mayoría de distritos tienen bajos niveles de este indicador; no obstante, los distritos de Pachacamac, Pucusana, San Bartolo y Ancón presentan valores elevados de la variable; además, el histograma muestra que la distribución tiene alta asimetría positiva con una media alrededor del 8 %. En referencia a la covariable de hogares con hacinamiento, en la figura 5.4 se puede observar correlación espacial en la incidencia de distritos con hacinamiento además se tiene que los distritos presentan una marcada asimetría negativa cuyos valores en su mayoría son superiores al 10 %.

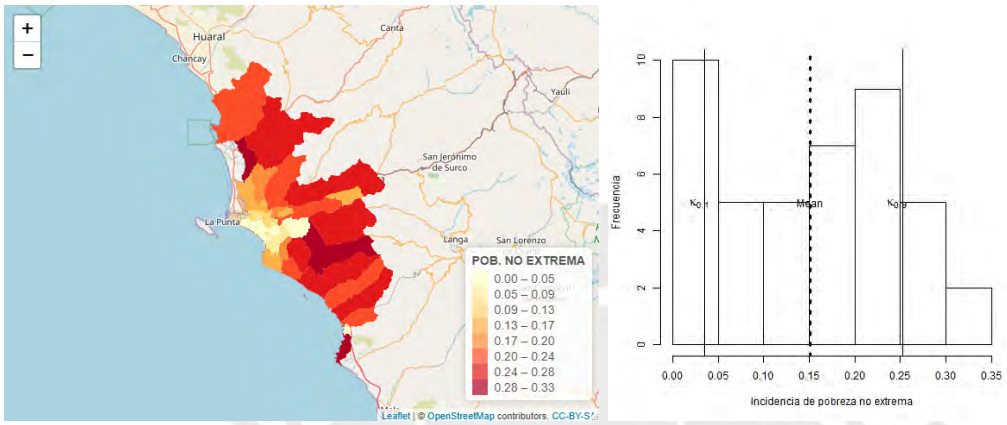


Figura 5.2: Histograma y mapa de la incidencia de la población en condición de pobreza no extrema por distrito en Lima.

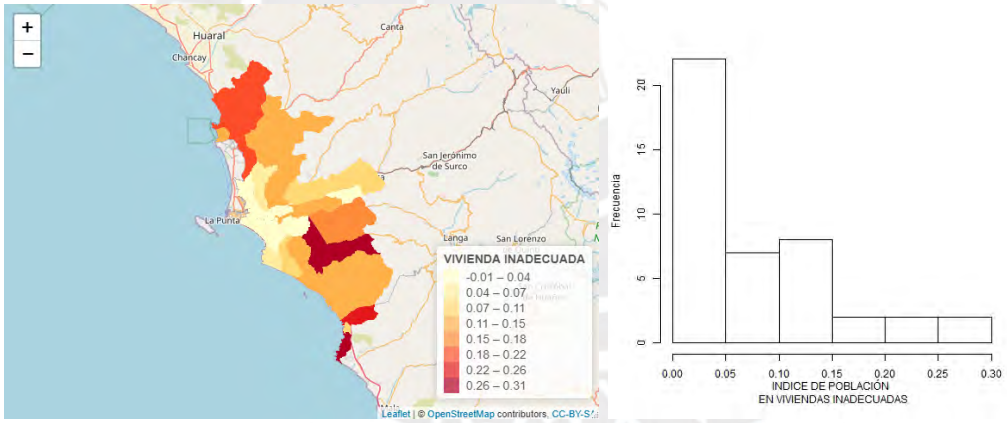


Figura 5.3: Histograma y mapa de la incidencia la población en viviendas inadecuadas por distrito en Lima.

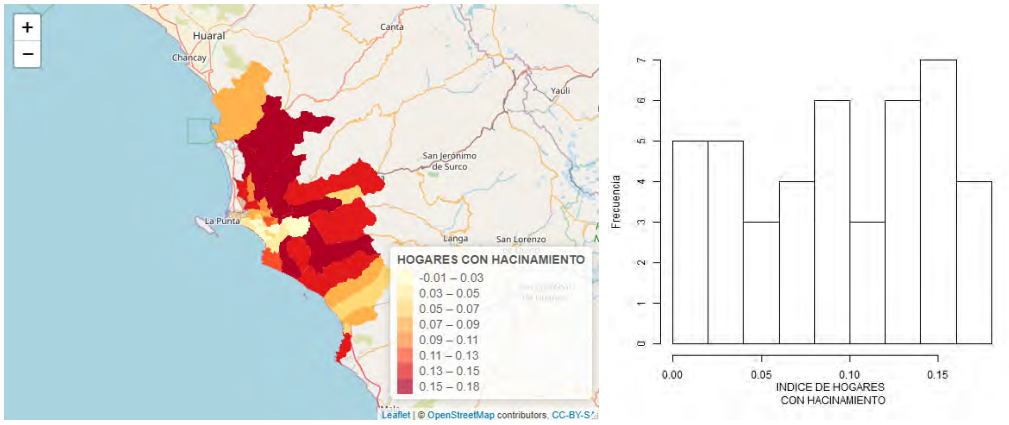


Figura 5.4: Histograma y mapa de la incidencia de hogares en vivivendas con hacinamiento por distrito en Lima.

Para la elección de las covariables a utilizar se analizó la relación de 38 posibles covariables referidas a niveles de pobreza, de acceso a servicios básicos, niveles de educación y servicios de salud en la población, todas ellas calculadas para la elaboración del Mapa de Pobreza o para el Mapa de Necesidades Básicas Insatisfechas elaborado por el INEI. En la tabla 5.1 se muestra la estructura de la base de datos que se ha utilizado y en la Figura 5.5 se presenta las correlaciones, la tendencia líneal (línea azul) y tendencia no lineal (línea roja) de parte de las variables analizadas, en donde se puede observar que las variables escogidas han presentado una correlación lineal aceptable con el índice de pobreza no extrema, 0.858 y 0.813 respectivamente.

5.3. Estructura del modelo

La estructura del modelo para la aplicación a los datos reales se define como $Y_i \sim \text{Kumar}(\kappa_i, \phi, q)$ con $0 < Y_i < 1$, $0 < \kappa_i < 1$, $\phi > 0$, $0 < q < 1$, la cual, como se ha mencionado en la subsección 2.2.1, es una reparametrización de la representación usual de la variable aleatoria Y_i con distribución Kumaraswamy en la que κ_i, es el cuantil q de Y_i, ϕ es el parámetro de precisión, la fdp queda definida por

$$f_{Y_i}(y_i|\kappa_i,\phi) = -\frac{\log(1-q)\phi}{\log(1-e^{-\phi})\log(\kappa_i)} y_i^{-\frac{\phi}{\log(\kappa_i)}-1}\{1-y_i^{-\frac{\phi}{\log(\kappa_i)}}\}^{\frac{\log(1-q)}{\log(1-e^{-\phi})}-1},$$

y

$$E(Y_i) = \frac{\log(1-q)}{\log(1-e^{-\phi})} B\left(1-\frac{\log(\kappa_i)}{\phi}, \frac{\log(1-q)}{\log(1-e^{-\phi})}\right),$$

donde $B(.,.)$ denota a la función beta, Y_i es una variable aleatoria que representa el índice de pobreza no extrema en el i-ésimo distrito de la provincia de Lima, $i = 1, 2, \ldots, n$; κ_i es el cuantil q de Y_i y ϕ es el parámetro de precisión de Y_i. El valor de n corresponde a los 43 distritos en estudio

Tomando en cuenta el predictor lineal considerado y la matriz de vecindad considerado se comparan tres modelos:

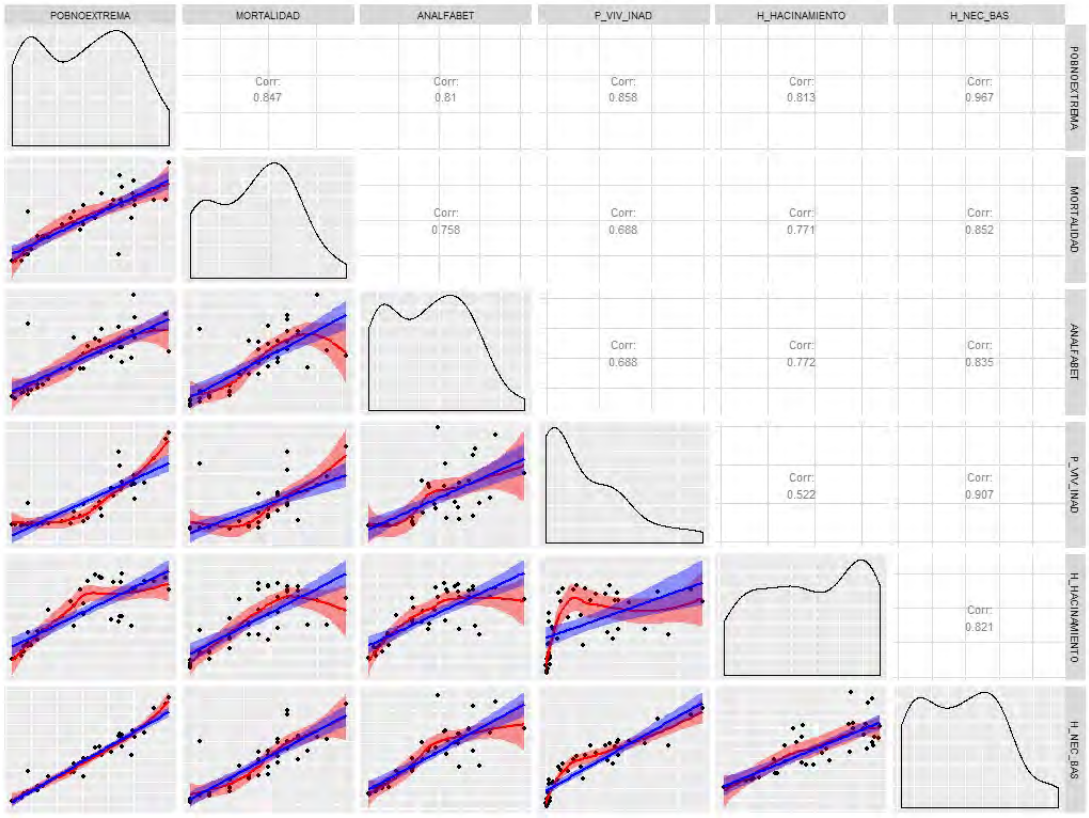


Figura 5.5: Diagramas de dispersión y correlaciones de parte de las variables analizadas para el modelo. La línea azul permite evaluar la tendencia lineal, mientras que la línea roja permite evaluar la tendencia no lineal entre variables.

5.3.1. Modelo de regresión cuantílica Kumaraswamy no espacial (KNSQ)

Se incorporan covariables al cuantil q, κ_i, pero no se le incluye el efecto espacial. Para las covariables se considera el intercepto como x_{0i}, la incidencia de población que se encuentran en viviendas con características físicas inadecuadas como x_{1i} y la incidencia de hogares en viviendas con hacinamiento como x_{2i}.

$$Y_i|\boldsymbol{\theta}, \boldsymbol{\psi} \sim \text{Kumar}(\kappa_i, \phi, q),$$

$$g_1(\kappa_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i},$$

donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\intercal$ es el vector de los coeficientes de regresión asociados a κ_i, $\boldsymbol{X_i} = (1, x_{1i}, x_{2i})^\intercal$ es el vector de las covariables anteriormente definidas; $g_1(\cdot)$ es la función de enlace que para el presente trabajo será considerada como $g_1(\cdot) = \text{logit}(\cdot)$, función de enlace logística. Definiendo $\boldsymbol{\theta} = \boldsymbol{\beta}$, $\boldsymbol{\psi} = \phi$ e $\boldsymbol{Y}^\intercal = (Y_1, ..., Y_n)$, la función de verosimilitud para el

modelo puede ser escrita como sigue

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) &= p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= f_{Y_1}(y_1|\boldsymbol{\theta_1}, \boldsymbol{\psi}) \times f_{Y_2}(y_2|\boldsymbol{\theta_2}, \boldsymbol{\psi}) \times \ldots \times f_{Y_n}(y_n|\boldsymbol{\theta_n}, \boldsymbol{\psi}) \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i|\boldsymbol{\theta_i}, \boldsymbol{\psi}) \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i|\kappa_i, \phi), \end{aligned} \tag{5.1}$$

donde $\boldsymbol{\kappa}^\intercal = (\kappa_1, \ldots, \kappa_n)$ es un vector cuyas componentes se definen como

$$\kappa_i = \frac{1}{1 + \exp(\boldsymbol{X_i}^\intercal \boldsymbol{\beta})},$$

y $f_{Y_i}(y_i|\kappa_i, \phi, q)$ es la función de densidad de probabilidad de una variable aleatoria con distribución Kumaraswamy.

5.3.2. Modelo de regresión cuantílica Kumaraswamy espacial (KSQ-CAR)

Se incorporan covariables al cuantil q, κ_i, y además se incluye el efecto espacial. Para el efecto espacial u_i se ha tomado en cuenta la dependencia espacial formada a partir de los límites por contiguidad de los distritos. El parámetro κ_i queda enlazado a las covariables y al efecto espacial mediante la siguiente función:

$$g_1(\kappa_i) = \boldsymbol{X_i}^\intercal \boldsymbol{\beta} + u_i,$$

donde $\boldsymbol{\beta}^\intercal = (\beta_0, \beta_1, \beta_2)$ es el vector de los coeficientes de regresión asociados a κ_i ; $\boldsymbol{u}^\intercal = (u_1, \ldots, u_n)$ son los efectos espaciales donde cada u_i está asociado a cada κ_i. $X_i = (1, x_{1i}, x_{2i})^\intercal$ es el vector de las covariables anteriormente definidas; $g_1(\cdot)$ es la función de enlace que para el presente trabajo será considerada como $g_1(\cdot) = \text{logit}(\cdot)$, función de enlace logística. Se considera $\boldsymbol{u}^\intercal = (u_1, \ldots, u_n)$ donde u_i es el efecto espacial del distrito i-ésimo, entonces la distribución conjunta para $\boldsymbol{u}$ tiene la distribución mencionada en (3.1) la cual es la siguiente:

$$\boldsymbol{u} \sim \text{N}_n(\boldsymbol{0}, \boldsymbol{Q_u^{-1}}),$$

donde $\boldsymbol{Q_u^{-1}} = \tau_{\boldsymbol{u}}(\boldsymbol{D} - \alpha \boldsymbol{W})$, $\boldsymbol{D}$ es la matriz diagonal (N_i), α es un parámetro que controla la dependencia espacial tal que $\alpha = 0$ implica independencia espacial mientras que $\alpha = 1$ conlleva a un modelo condicional autoregresivo intrínseco (ICAR); y $\boldsymbol{W}$ es la matriz de vecindad, cuyos componentes son definidos por

$$w_{ij} = \begin{cases} 1 & \text{, si } i \sim j \text{ } (i \text{ es vecino de } j), \\ 0 & \text{, si } i \text{ no es vecino de } j \end{cases}$$

Por lo tanto, dadas las distribuciones gaussianas independientes para $\boldsymbol{u}$ y $\boldsymbol{\beta}$, entonces $\boldsymbol{\theta}^\intercal = (\boldsymbol{u}, \boldsymbol{\beta})$ también sigue una distribución gaussiana multivariada definida por

$$\boldsymbol{\theta} \ \sim \ \mathrm{N}_k(\mathbf{0}, \boldsymbol{Q^{-1}(\psi)}), \tag{5.2}$$

donde $\boldsymbol{\psi}^\intercal = (\phi, \tau_{\boldsymbol{u}})$ es el vector de hiperparámetros y $\boldsymbol{Q}(\boldsymbol{\psi})$ es una matriz dispersa (con gran cantidad de valores 0 en sus componentes) lo que conlleva a una mayor eficiencia computacional. Definiendo $\boldsymbol{\theta}^\intercal = (\boldsymbol{u}, \boldsymbol{\beta})$, $\boldsymbol{\psi}^\intercal = (\phi, \tau_{\boldsymbol{u}}, \alpha)$ e $\boldsymbol{Y}^\intercal = (Y_1, ..., Y_n)$, la función de verosimilitud para el modelo puede ser escrita como sigue:

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) &= p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= f_{Y_1}(y_1|\boldsymbol{\theta_1}, \boldsymbol{\psi}) \times f_{Y_2}(y_2|\boldsymbol{\theta_2}, \boldsymbol{\psi}) \times \ldots \times f_{Y_n}(y_n|\boldsymbol{\theta_n}, \boldsymbol{\psi}) \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i|\boldsymbol{\theta_i}, \boldsymbol{\psi}) \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i|\kappa_i, \phi), \end{aligned} \tag{5.3}$$

donde $\boldsymbol{\kappa}^\intercal = (\kappa_1, \ldots, \kappa_n)$ es un vector cuyas componentes se definen como

$$\kappa_i = \frac{1}{1 + \exp(\boldsymbol{X_i^\intercal \beta} + u_i)},$$

y $f_{Y_i}(y_i|\kappa_i, \phi, q)$ es la función de densidad de probabilidad de una variable aleatoria con distribución Kumaraswamy.

5.3.3. Modelo de regresión cuantílica Kumaraswamy espacial usando el método SPOCK (KSQ-SPOCK)

También se evaluará el modelo propuesto en la sección anterior pero con una matriz de vecindad, denotada como $\boldsymbol{W^*}$ creada a partir del método Spock. Siguiendo el método descrito en la Sección 2.3.4, se tiene una matriz $n \times 2$ denominada $\boldsymbol{s} = [\boldsymbol{s_1}, \boldsymbol{s_2}] = [s_{i1}, s_{i2}]$, $\forall i = 1, \ldots, n$ donde s_{i1}, s_{i2} es la coordenada del centroide del distrito i-ésimo. Luego se definirá $\boldsymbol{s^*} = \boldsymbol{P^\perp s}$ donde $\boldsymbol{P^\perp} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal$ es la matriz proyectada en el espacio ortogonal $\boldsymbol{X}$. Considerando $\boldsymbol{s^*}$ se construirá la matriz $\boldsymbol{W^*}$ tomando como vecinos a los centroides mas cercanos con respecto a su distancia Euclidiana.

Para entender el razonamiento detrás de este método, se tendrá que u_i es un punto $\Psi(s_1, s_2)$ sobre una superficie suave definida para una posición arbitraria (s_1, s_2) que representa a alguno de los centroides de los distritos de la provincia de Lima y sea $\Lambda\Psi = (\gamma_1, \gamma_2)^\intercal$ la gradiente de Ψ evaluada en el punto (s_{01}, s_{02}) utilizada para realizar la siguiente expansión de Taylor

$$\begin{aligned} \Psi(s_1, s_2) &= \Psi(s_0 1, s_0 2) + (s_1 - s_{01}, s_2 - s_{02})\Lambda\Psi + R(s_1, s_2, s_{10}, s_{20}) \\ &= \gamma_0 + \gamma_1(s_1 - s_{01}) + \gamma_2(s_2 - s_{02}) + R(s_1, s_2, s_{10}, s_{20}), \end{aligned} \tag{5.4}$$

donde el resto $R(s_1, s_2, s_{10}, s_{20})$ tiene forma cuadrática dada por $\boldsymbol{hH(r)h}$, $\boldsymbol{h} = (s_1 - s_{01}, s_2 - s_{02})$, $\boldsymbol{H(r)}$ es la matriz hessiana de Ψ evaluada en el punto $\boldsymbol{r}$, que está en algún punto entre (s_1, s_2) y (s_{01}, s_{02}).
Luego evaluando 5.4 en cada uno de los centroides $\boldsymbol{s} = [\boldsymbol{s_1}, \boldsymbol{s_2}] = [s_{i1}, s_{i2}]$, $\forall i = 1, \ldots, n$, se tiene el vector

$$\begin{aligned}\boldsymbol{u} = \Psi(\boldsymbol{s}) &= (\lambda_0 - s_{01} - s_{02})\mathbf{1} + \lambda_1 \boldsymbol{s_1} + \lambda_2 \boldsymbol{s_2} + R(\boldsymbol{s_1}, \boldsymbol{s_2}, s_{01}, s_{02}) \\ &= [\mathbf{1}, \boldsymbol{s_1}, \boldsymbol{s_2}]\boldsymbol{\gamma} + \boldsymbol{R}\end{aligned}$$

donde $\boldsymbol{\gamma}$ es la gradiente de $\boldsymbol{\Psi}$ evaluada en cada uno de los puntos de referencia del mapa.
Luego el predictor lineal quedará definido por

$$\begin{aligned}\boldsymbol{X\beta} + \boldsymbol{u} &= \boldsymbol{X\beta} + [1, s_1, s_2]\boldsymbol{\gamma} + \boldsymbol{R}, \\ &= \boldsymbol{X\beta} + \boldsymbol{P}[\mathbf{1}, \boldsymbol{s_1}, \boldsymbol{s_2}]\boldsymbol{\gamma} + \boldsymbol{P}^{\perp}[\mathbf{1}, \boldsymbol{s_1}, \boldsymbol{s_2}]\boldsymbol{\gamma} + \boldsymbol{R},\end{aligned}$$

y el componente $\boldsymbol{P}[\mathbf{1}, \boldsymbol{s_1}, \boldsymbol{s_2}]\boldsymbol{\gamma}$ se removerá en el modelo SPOCK.
A este modelo se le denotará KSQ-SPOCK (Kumaraswamy Spatial Quantile model - SPOCK method).
Al igual que en el modelo KSQ-CAR, en el modelo KSQ-SPOCK, dadas las distribuciones gaussianas independientes para $\boldsymbol{u}$ y $\boldsymbol{\beta}$, entonces $\boldsymbol{\theta}^{\intercal} = (\boldsymbol{u}, \boldsymbol{\beta})$ también sigue una distribución gaussiana multivariada definida por

$$\boldsymbol{\theta} \ \sim \ N_k(\mathbf{0}, \boldsymbol{Q^{-1}(\psi)}), \tag{5.5}$$

donde $\boldsymbol{\psi}^{\intercal} = (\phi, \tau)$ es el vector de hiperparámetros y $\boldsymbol{Q}(\boldsymbol{\psi})$ es una matriz dispersa (con gran cantidad de valores 0 en sus componentes) generada a partir de la matriz de vecindad $\boldsymbol{W^*}$ lo que conlleva a una mayor eficiencia computacional. Definiendo $\boldsymbol{\theta}^{\intercal} = (\boldsymbol{u}, \boldsymbol{\beta})$, $\boldsymbol{\psi}^{\intercal} = (\phi, \tau_{\boldsymbol{u}}, \alpha)$ e $\boldsymbol{Y}^{\intercal} = (Y_1, ..., Y_n)$, al igual que el modelo KSQ-CAR, la función de verosimilitud puede ser escrita como sigue:

$$\begin{aligned}L(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) &= p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= f_{Y_1}(y_1|\boldsymbol{\theta_1}, \boldsymbol{\psi}) \times f_{Y_2}(y_2|\boldsymbol{\theta_2}, \boldsymbol{\psi}) \times \ldots \times f_{Y_n}(y_n|\boldsymbol{\theta_n}, \boldsymbol{\psi}) \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i|\boldsymbol{\theta_i}, \boldsymbol{\psi}) \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i|\kappa_i, \phi),\end{aligned} \tag{5.6}$$

donde $\boldsymbol{\kappa}^{\intercal} = (\kappa_1, \ldots, \kappa_n)$ es un vector cuyas componentes se definen como

$$\kappa_i = \frac{1}{1 + \exp(\boldsymbol{X_i^{\intercal}\beta} + u_i)},$$

y $f_{Y_i}(y_i|\kappa_i, \phi, q)$ es la función de densidad de probabilidad de una variable aleatoria con distribución Kumaraswamy.

5.4. Inferencia bayesiana

La fdp a posteriori para $\boldsymbol{\theta}^\intercal = (\boldsymbol{u}, \boldsymbol{\beta})$ y $\boldsymbol{\psi}^\intercal = (\phi, \tau_{\boldsymbol{u}}, \alpha)$ en los modelos KSQ-CAR y KSQ-SPOCK; o para $\boldsymbol{\theta} = \boldsymbol{\beta}$ y $\boldsymbol{\psi} = \phi$ en el modelo KNSQ, se denota como $p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y})$ y se define como sigue:

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\psi}) \times p(\boldsymbol{\theta}|\boldsymbol{\psi}) \times p(\boldsymbol{\psi})}{p(\boldsymbol{Y})},$$

donde $p(\boldsymbol{Y})$ no depende de $\boldsymbol{\theta}$ por lo tanto

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

y puede también ser expresada como

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

donde $L(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y})$ es la función de verosimilud definida en 5.1 para el modelo KNSQ, en 5.3 para el modelo KSQ-CAR o en 5.6 para el modelo KSQ-SPOCK; $p(\boldsymbol{\theta}|\boldsymbol{\psi})$ es la distribución condicional de $\boldsymbol{\theta}|\boldsymbol{\psi}$ y $p(\boldsymbol{\psi})$ es la distribución a priori de $\boldsymbol{\psi}$.
Para el presente documento se asume independencia entre $\boldsymbol{u}$, $\boldsymbol{\beta}$ y ϕ por lo que se puede tener la distribución a priori

$$p(\boldsymbol{\theta}|\boldsymbol{\psi}) \times p(\boldsymbol{\psi}) = p(\boldsymbol{\beta}) \times p(\phi),$$

para el modelo KNSQ, o ,

$$p(\boldsymbol{\theta}|\boldsymbol{\psi}) \times p(\boldsymbol{\psi}) = p(\boldsymbol{u}|\tau_{\boldsymbol{u}}, \alpha) \times p(\boldsymbol{\beta}) \times p(\phi) \times p(\tau_{\boldsymbol{u}}) \times p(\alpha),$$

para los modelos KSQ-CAR y KSQ-SPOCK donde $p(\boldsymbol{\beta}) = p(\beta_0)p(\beta_1)p(\beta_2) = \prod_{j=0}^{2} p(\beta_j)$.
Para los coeficientes $\boldsymbol{\beta}$ se asume que $\beta_j \sim N(0, 1000), j = 0, 1, 2$. Por lo tanto, $p(\beta_j) = 1/(\sqrt{2000\pi})e^{-\frac{1}{2000}\beta_j^2}$ para $j = 0, 1, 2$.
Para $\boldsymbol{u}$, como ya se ha mencionado en 2.16, se asume la distribución

$$\boldsymbol{u} \sim \mathrm{N}_n(\boldsymbol{0}, \boldsymbol{Q_u^{-1}}).$$

Tomando en cuenta que se tienen los campos aleatorios de markov gaussianos (GMRF) independientes $\boldsymbol{u}$, $\boldsymbol{\beta}$ entonces $\boldsymbol{\theta} = (\boldsymbol{u}\ ,\ \boldsymbol{\beta})$ dado el conjunto de hiperparámetros es una familia GMRF definida de la siguiente manera

$$\boldsymbol{\theta}|\boldsymbol{\psi} \sim \mathrm{N}_{n+3}(\boldsymbol{0}, \boldsymbol{Q^{-1}(\psi)}), \tag{5.7}$$

donde $\boldsymbol{Q(\psi)}$ es una matriz dispersa definida por:

$$\boldsymbol{Q(\psi)} = \begin{bmatrix} \boldsymbol{Q_\beta} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Q_u} \end{bmatrix}.$$

También $\boldsymbol{Q_u}$ como ya se ha mencionado presenta gran cantidad de ceros en sus elementos dada su definición en función de la matriz de vecindad. Para los hiperparámetros de $\boldsymbol{\psi} = (\phi$, $\tau_{\boldsymbol{u}}$ y $\alpha)$ se asumirán a priori las siguientes distribuciones

$$\begin{aligned}\alpha &\sim \text{uniforme}(0,1)\\ \tau_{\boldsymbol{u}} &\sim \text{gamma}(4, 0.5)\\ \phi &\sim \text{gamma}(4, 0.5).\end{aligned}$$

Considerando la verosimilitud definida en (5.1) y las distribuciones a priori definidas anteriormente la fdp a posteriori definida en (5.4) puede ser expresada como:

$$\begin{aligned}p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) &\propto \prod_{i=1}^{n} f_{Y_i}(y_i|\kappa_i, \phi) \times p(\boldsymbol{u}|\tau) \times p(\boldsymbol{\beta}) \times p(\phi) \times p(\tau) \times p(\alpha)\\ &\propto \prod_{i=1}^{n} f_{Y_i}(y_i|\kappa_i, \phi) \times p(\boldsymbol{\theta}|\boldsymbol{\psi}) \times p(\boldsymbol{\psi}),\end{aligned}$$

donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_i, \ldots, \theta_K)$ con $K = (43 + 3)$ indica el tamaño del campo aleatorio gaussiano $\boldsymbol{\theta}$ formado por $\boldsymbol{u}$ y $\boldsymbol{\beta}$ y cuya distribución, dados los hiperparámetros $\boldsymbol{\psi}$, se definió en (5.7). Por lo tanto

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) \propto \prod_{i=1}^{n} f_{Y_i}(y_i|\kappa_i, \phi_i) \times \frac{|\boldsymbol{Q}(\boldsymbol{\psi})|^{1/2}}{(2\pi)^{n/2}} \exp(-\frac{1}{2}\theta^t \boldsymbol{Q}(\boldsymbol{\psi})\theta) \times p(\boldsymbol{\psi}),$$

donde $|\boldsymbol{Q}(\boldsymbol{\psi})|$ es el determinante de $\boldsymbol{Q}(\boldsymbol{\psi})$. Además, para los hiperparámetros $p(\boldsymbol{\psi})$, tomando en cuenta las distribuciones a priori definidas anteriormente se tiene que

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}) \propto \prod_{i=1}^{n} f_{Y_i}(y_i|\kappa_i, \phi) \times \frac{|\boldsymbol{Q}(\boldsymbol{\psi})|^{1/2}}{(2\pi)^{n/2}} \exp(-\frac{1}{2}\theta^t \boldsymbol{Q}(\boldsymbol{\psi})\theta) \times 0.001 \exp(-0.001\phi) \times 0.0005 \exp(-0.0005\tau_u)$$

Para estimar el modelo propuesto se utilizará la libería RSTAN (https://mc-stan.org/users/interfaces/stan) que se encuentra incluida en el entorno de programación R.
La estimación es realizada mediante los algoritmos basados en MCMC y en el algoritmo HMC descrito en 2.4.4. En las Figuras B.7 y B.8, se tienen los gráficos de cadenas para cada parámetro e hiperparámetro de los modelos KSQ-CAR y KSQ-SPOCK en donde se puede apreciar una convergencia aceptable para los tres escenarios. Asímismo en las Figuras B.9 y B.10 se tienen los histogramas de los parámetros en donde se incluyen dos lineas adicionales para el intervalo de credibilidad (IC) al 95 % de las simulaciones realizadas.

5.5. Resultados

En esta sección se presentan los resultados de la estimación del modelo inicial KSQ-CAR, el modelo KSQ-SPOCK que usa la matriz de vecindad SPOCK para evitar la confusión espacial entre las variables del modelo y además el modelo similar KNSQ que no toma en cuenta los efectos espaciales.
A modo general, se ha realizado la estimación con un tiempo de ejecución razonable, con buenos resultados tanto para el modelo KSQ-CAR como para el modelo KSQ-SPOCK.
En el Cuadro 5.2 se muestra una comparación de los resultados de cada escenario tomando como criterios al WAIC, LPML, RMSE y al tiempo de procesamiento. Estos criterios han

sido definidos en la Sección 2.5.

		Modelos		
q	Criterio	KNSQ	KSQ-CAR	KSQ-SPOCK
0.1	WAIC	-156.267	**−188.109**	-181.869
	LPML	78.659	**95.681**	92.374
	RMSE	0.073	**0.046**	0.050
	Tiempo (seg.)	38.520	385.730	**325.278**
0.5	WAIC	-158.019	**−192.568**	-187.296
	LPML	79.539	**97.960**	95.090
	RMSE	0.038	**0.017**	0.018
	Tiempo (seg.)	35.083	355.616	**303.985**
0.9	WAIC	-159.369	-195.358	**−190.118**
	LPML	80.218	**99.392**	96.566
	RMSE	0.062	**0.032**	0.034
	Tiempo (seg.)	31.810	360.733	**341.748**

Cuadro 5.2: Criterios de selección y tiempos de procesamiento en segundos para cada escenario y modelo propuesto. Por cada criterio y escenario se resalta en negritas el modelo que tiene mejores resultados.

Se puede observar que el WAIC del modelo KSQ-CAR es menor en los 3 cuantiles analizados por lo que este modelo sería el más adecuado. Ello se corrobora también con los valores del LPML los cuales son más altos también para este modelo. No obstante los valores del modelo KSQ-SPOCK son muy cercanos y deberían ser considerados en caso las covariables utilizadas presenten una mayor correlación espacial. Los valores del RMSE obtenidos en cada escenario para los modelos KSQ-CAR y KSQ-SPOCK son mucho más bajos que el modelo KNSQ lo que indica un mejor ajuste para los modelos espaciales.

El tiempo de ejecución del modelo KNSQ es mucho más bajo lo que era predecible pues el número de parámetros del modelo es mucho menor. Poniendo atención en los modelos espaciales tenemos que el modelo KSQ-SPOCK tiene un tiempo de ejecución menor que el modelo KSQ-CAR en todos los escenarios.

De acuerdo a la evaluación realizada se tiene que los modelos KSQ-CAR y KSQ-SPOCK presentan resultados parecidos y mejores que el modelo KNSQ por lo que las estimaciones se presentan para ambos.

Los cuadros 5.3 y 5.4 muestra un resumen de los valores estimados de los parámetros en cada uno de los cuantiles $q = 0.1$, $q = 0.5$ y $q = 0.9$ analizados. Se muestran las estimaciones a posteriori de la media y la desviación estándar, además, se puede observar que los coeficientes de las covariables son significativos pues sus intervalos de credibilidad (IC) al 95 % no contienen al valor de 0.

Dado que se tiene valores positivos para los coeficientes de ambas covariables se infiere que el aumento en la incidencia de población que se encuentran en viviendas con características físicas inadecuadas y la incidencia de hogares en viviendas con hacinamiento ocasionan un incremento en los cuantiles $q = 0.1$, $q = 0.5$ y $q = 0.9$ de la incidencia de pobreza distrital en Lima.

Analizando el parámetro β_1 podemos notar que el valor más alto se tiene para el cuantil 0.10 lo que indica que a mayor incidencia de viviendas en condiciones inadecuadas se tiene que el cuantil 0.10 de la incidencia de pobreza no extrema se ve más afectado. De la misma manera notamos que el cuantil 0.9 de la distribución de la incidencia de pobreza no extrema se ve menos afectado ante un aumento en la incidencia de viviendas en condiciones inadecuadas.

El parámetro β_2 nos permite interpretar que la incidencia de hogares con hacinamiento tiene una influencia mayor en todos los cuantiles de la distribución de la incidencia de pobreza no extrema en comparación con la incidencia de viviendas en mal estado. Además el impacto en el aumento del porcetaje de hogares con hacinamiento es más alto en el cuantil 0.10 de la distribución en análisis.

Con respecto al parámetro de precisión ϕ se puede observar que en en los cuantiles extremos 0.1 y 0.9 la precision es más baja lo que conlleva a pensar que otros factores no explicados por los efectos espaciales podrían afectar las colas de la distribución.

Con respecto al parámetro de precisión, τ, del efecto espacial podemos notar que es más alto en el cuantil 0.90 de la distribución lo que nos permite concluir que en este punto de la distribución la dependencia espacial es menor para los cuantiles más altos ello también guarda correspondencia con el valor del parámetro α, parámetro que también permite analizar la dependencia espacial, el cual se incrementa para los cuantiles más altos.

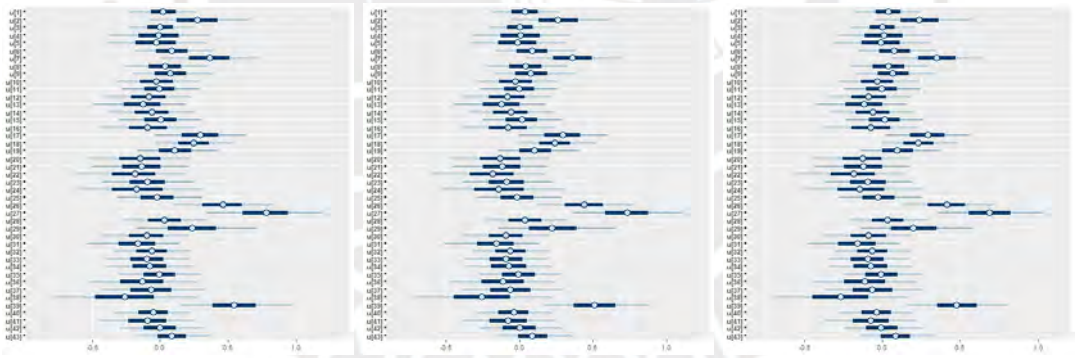
Escenario	Parámetro	Media	Desv. est.	95 %IC
$q = 0.1$	β_0	-3.85	0.23	-4.32 ; -3.45
	β_1	4.72	1.10	2.52 ; 6.88
	β_2	11.54	1.58	8.50 ;14.69
	ϕ	13.20	2.68	9.05 ;19.52
	τ	5.98	2.63	2.51 ; 12.58
	α	0.66	0.26	0.08 ; 0.99
$q = 0.5$	β_0	-3.29	0.17	-3.63 ;-2.97
	β_1	4.19	0.95	2.27 ; 6.00
	β_2	10.44	1.41	7.68 ; 13.23
	ϕ	11.86	2.74	7.65 ; 18.44
	τ	6.58	2.72	2.87 ; 13.38
	α	0.66	0.26	0.08 ; 0.99
$q = 0.9$	β_0	-2.95	0.17	-3.27 ; -2.60
	β_1	3.96	0.87	2.18 ; 5.62
	β_2	9.71	1.32	7.12 ; 12.31
	ϕ	11.01	2.73	6.73 ; 17.32
	τ	7.06	2.79	3.17 ; 13.99
	α	0.63	0.26	0.07 ; 0.98

Cuadro 5.3: Media, desviación estándar e intervalo de credibilidad (IC) al 95 % de los parámetros para el modelo KSQ-CAR.

Escenario	Parámetro	Media	Desv. est.	95 %IC
$q = 0.1$	β_0	-3.89	0.22	-4.35 ; -3.50
	β_1	4.91	1.04	2.88 ; 6.98
	β_2	11.60	1.49	8.67 ; 14.52
	ϕ	12.45	2.63	8.58 ; 18.91
	τ	5.46	2.90	1.93 ; 13.00
	α	0.64	0.28	0.05 ; 0.99
$q = 0.5$	β_0	-3.30	0.19	-3.62 ; -2.98
	β_1	4.34	0.88	2.61 ; 6.06
	β_2	10.40	1.33	7.76 ; 12.97
	ϕ	11.20	2.62	7.27 ; 17.51
	τ	5.74	2.76	2.23 ; 12.85
	α	0.62	0.29	0.04 ; 0.99
$q = 0.9$	β_0	-2.94	0.18	-3.27 ; -2.59
	β_1	4.05	0.82	2.44 ; 5.65
	β_2	9.70	1.26	7.17 ; 12.15
	ϕ	10.34	2.67	6.36 ; 16.75
	τ	6.10	2.87	2.36 ; 13.16
	α	0.60	0.28	0.04 ; 0.99

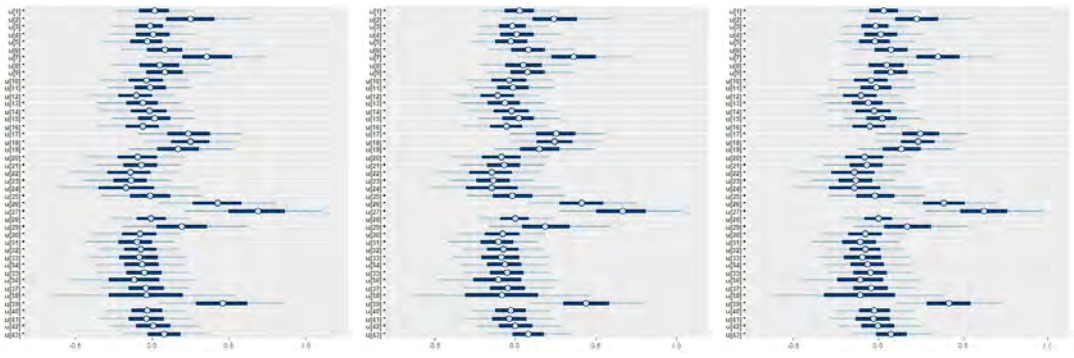
Cuadro 5.4: Media, desviación estándar e intervalo de credibilidad (IC) al 95 % de los parámetros para el modelo KSQ-SPOCK.

Con respecto a los efectos espaciales, en las Figuras 5.6 y 5.7, se puede observar los intervalos de credibilidad de los efectos espaciales de los distritos de la provincia de Lima. La barra azul gruesa muestra el IC al 50 % de los efectos espaciales mientras las líneas delgadas corresponden al IC al 90 %.



(a) q = 0.10 (b) q = 0.50 (c) q = 0.90

Figura 5.6: Intervalos de credibilidad (IC) al 95 % de los efectos espaciales para cada cuantil del modelo KSQ-CAR.



(a) q = 0.10 (b) q = 0.50 (c) q = 0.90

Figura 5.7: Intervalos de credibilidad (IC) al 95 % de los efectos espaciales para cada cuantil del modelo KSQ-SPOCK.

Las Figuras 5.8 y 5.9 muestran mapas correspondientes a las estimaciones de los efectos espaciales de la incidencia de pobreza no extrema. Podemos notar que los efectos espaciales son más altos en los distritos más alejados del centro de la ciudad, en particular se observa que el efecto espacial afecta sobretodo al cuantil 0.10 de la distribución. Los distritos que se ven más afectos son los de Santa María del Mar, Punta Negra, Santa Rosa y Ancón.

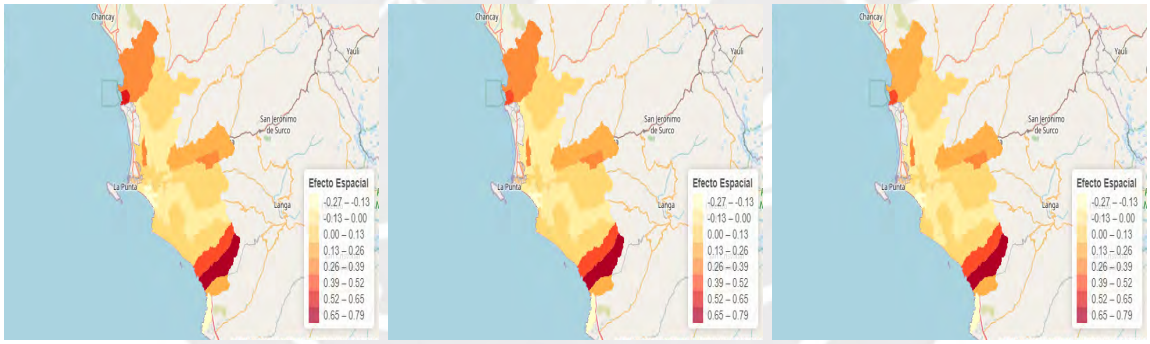


Figura 5.8: Mapa de las estimaciones para los efectos espaciales u_i para $q = 0.1$ (izquierda), $q = 0.5$ (centro) y $q = 0.9$ (derecha) del modelo KSQ-CAR.

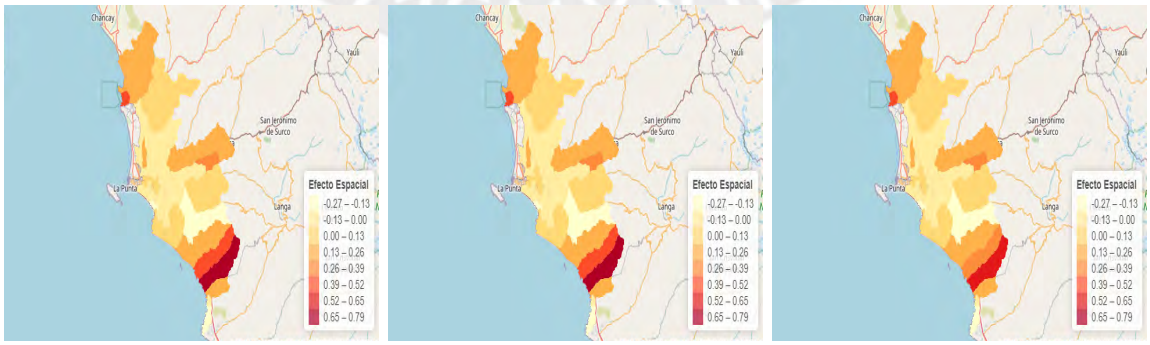


Figura 5.9: Mapa de las estimaciones para los efectos espaciales u_i para $q = 0.1$ (izquierda), $q = 0.5$ (centro) y $q = 0.9$ (derecha) del modelo KSQ-SPOCK.

En las Figuras 5.10 y 5.11 se presentan las estimaciones de los cuantiles para los modelos KSQ-CAR y KSQ-SPOCK. Se observa que el distrito de Pachacamac requiere especial

atención pues el cuantil 0.10 de la incidencia de pobreza no extrema en este distrito es muy elevada. De la misma manera se puede notar que el cuantil 0.90 de la incidencia de pobreza es similar en la mayoría de distritos excepto por aquellos que se encuentran en el centro de la ciudad.

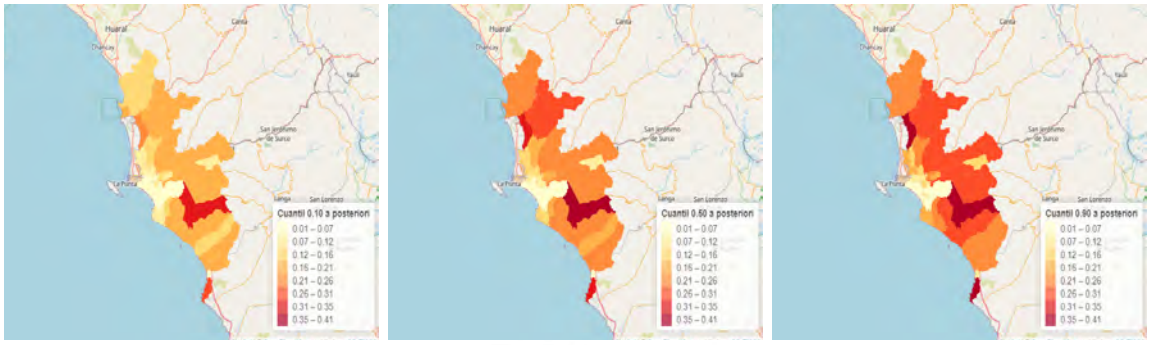


Figura 5.10: Mapa de las estimaciones de los cuantiles $q = 0.1$ (izquierda), $q = 0.5$ (centro) $q = 0.9$ (derecha) para el modelo KSQ-CAR.

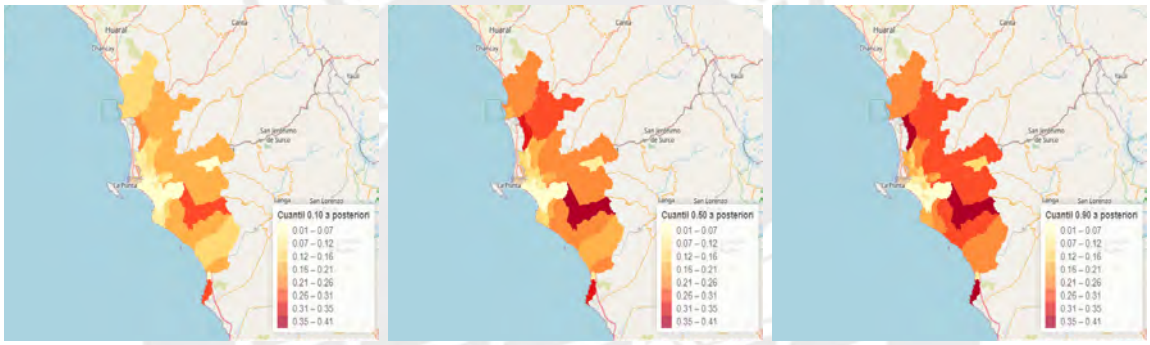
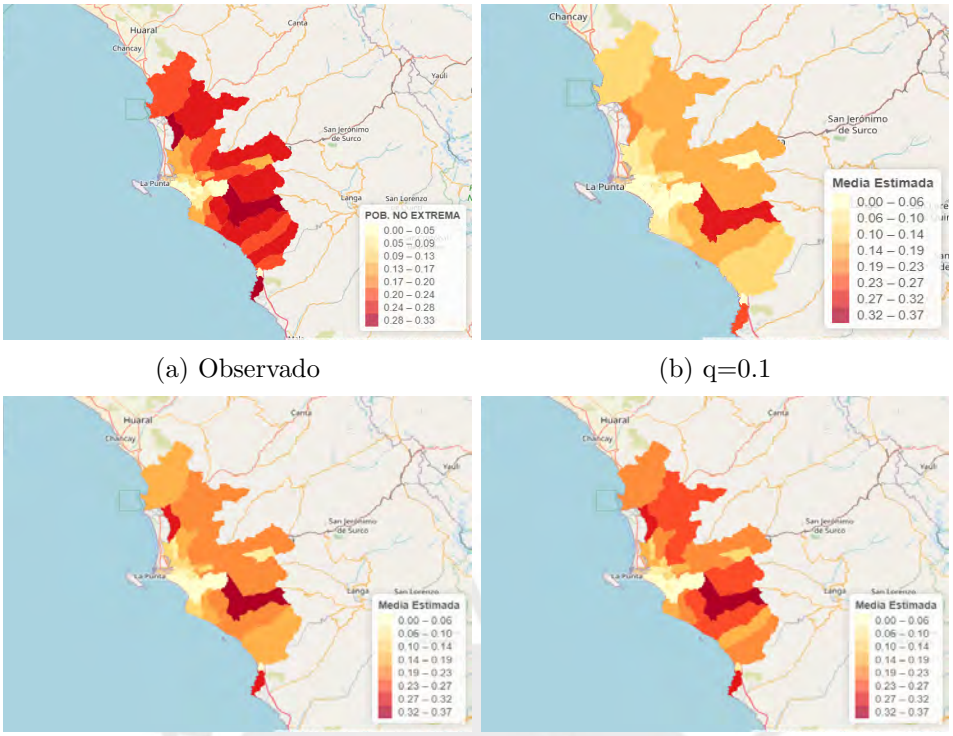


Figura 5.11: Mapa de las estimaciones de los cuantiles $q = 0.1$ (izquierda), $q = 0.5$ (centro) $q = 0.9$ (derecha) para el modelo KSQ-SPOCK.

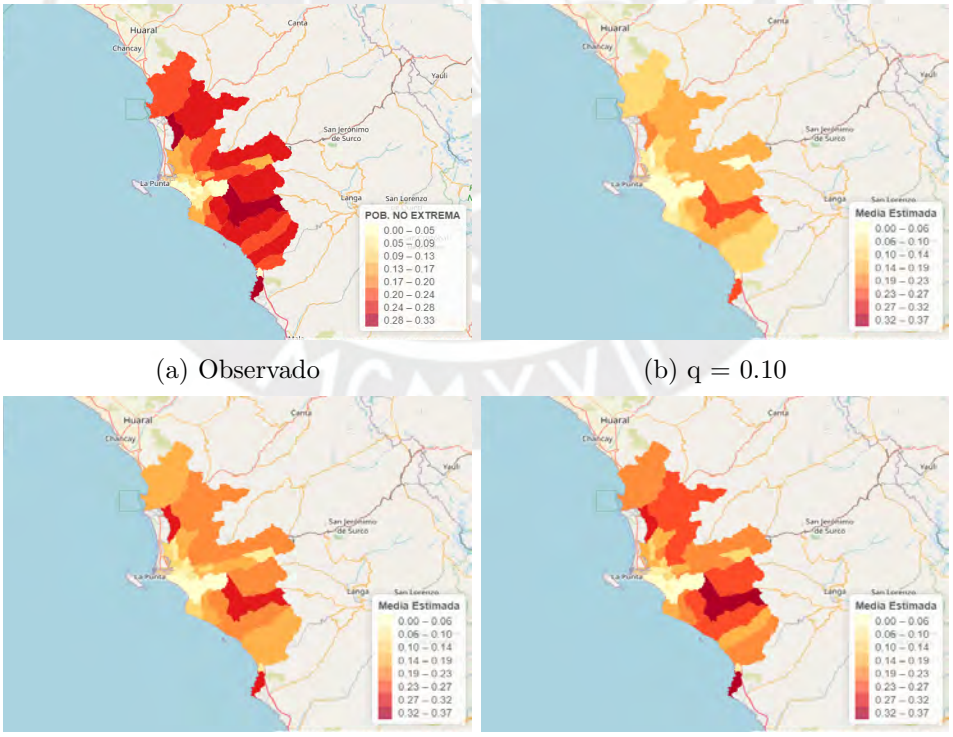
En las Figuras 5.12 y 5.13 se presentan las estimaciones de la incidencia de la pobreza no extrema para los modelos KSQ-CAR y KSQ-SPOCK. Se puede observar que las estimaciones tienen valores muy parecidos a los observados por lo que ambos modelos ajustan bien a los datos.



(a) Observado (b) q=0.1

(c) q=0.5 (d) q=0.9

Figura 5.12: Mapa de las estimación de la incidencia de pobreza no extrema para los tres escenarios del modelo KSQ-CAR.

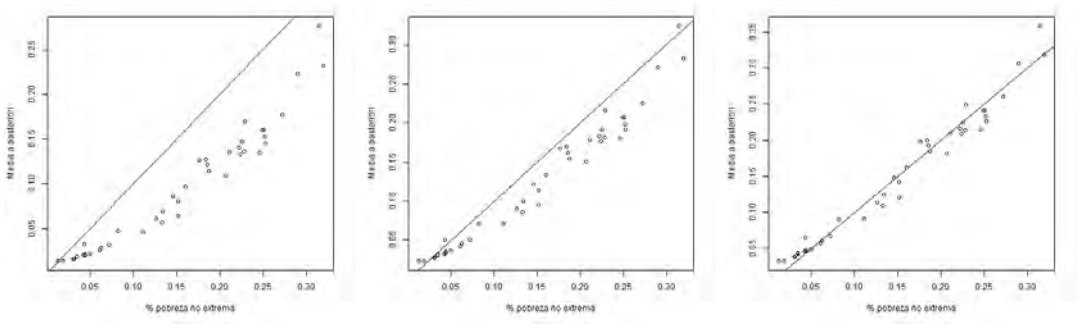


(a) Observado (b) q = 0.10

(c) q = 0.50 (d) q = 0.90

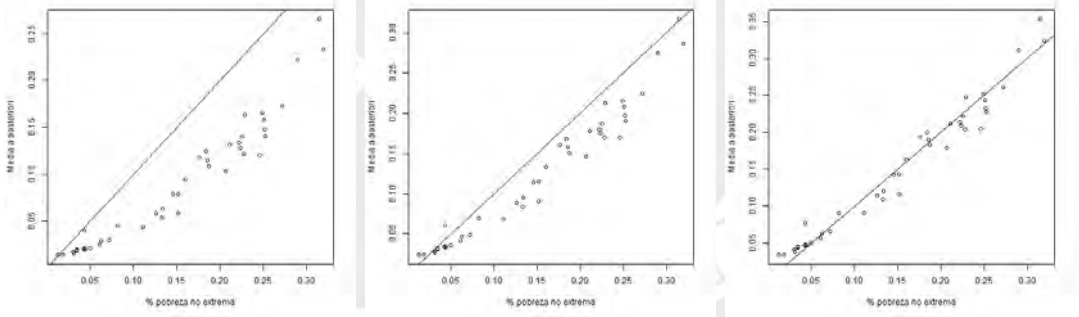
Figura 5.13: Mapa de las estimación de la incidencia de pobreza no extrema para los tres escenarios del modelo KSQ-SPOCK.

Al igual que los mapas de las estimaciones para la incidencia de pobreza no extrema, los diagramas de dispersión entre los datos observados y estimados ratifican el buen ajuste del modelo propuesto.



(a) q = 0.10 (b) q = 0.50 (c) q = 0.90

Figura 5.14: Diagramas de dispersión de las observaciones y estimaciones de la incidencia de pobreza no extrema para el modelo KSQ-CAR.



(a) q = 0.10 (b) q = 0.50 (c) q = 0.90

Figura 5.15: Diagramas de dispersión de las observaciones y estimaciones de la incidencia de pobreza no extrema para el modelo KSQ-SPOCK.

De acuerdo a los resultados obtenidos el modelo que presenta mejores estimaciones, tomando como criterios de evaluación al WAIC, LPML y RMSE, es el KSQ-CAR para el tercer escenario (q=0.9) lo que indica que ambas covariables (la incidencia de población en viviendas inadecuadas y la incidencia de hogares en viviendas con hacinamiento) y efecto espacial tienen mayor impacto en el cuantil 0.9 de la distribución de la incidencia de pobreza no extrema a nivel distrital. También cabe resaltar que los modlelos KSQ-CAR para el segundo escenario (q=0.5) y el modelo KSQ-SPOCK para el tercer escenario (q=0.9) presentan muy buenos resultados por lo que también se deben considerar los resultados de las estimacioens obtenidas en ellos. Cabe recalcar también que el modelo KSQ-SPOCK en el tercer escenario presenta un tiempo de ejecución menor que el modelo KSQ-CAR para el mismo escenario por lo que en caso de ser determinante la eficiencia de la estimación se escogería el modelo KSQ-SPOCK.

Sector	Variable	Descripción	Tipo
Demográfico	UBIGEO	Ubigeo del distrito	Categórica
	DISTRITO	Distrito observado	Categórica
	PROVINCIA	Provincia al que pertenece el distrito	Categórica
	DEPARTAMENTO	Departamento al que pertenece el distrito	Categórica
	POBLACION	Población en el distrito	Continua
Económico	POBNOEXTREMA	% de población en condición de pobreza no extrema	Continua
	POBREZA	% de población en condición de pobreza	Continua
	POBEXTREMA	% de población en condición de pobreza extrema	Continua
	GINI	Coeficiente de Gini del gasto	Continua
	P_DEP_ECONOM	% de población con alta dependencia económica	Continua
	H_DEP_ECONOM	% de hogares con alta dependencia económica	Continua
	RANKPOBREZA	Posición del distrito referida a su nivel de pobreza	Continua
Servicios	H_AGUA	% de hogares sin acceso a red pública de agua	Continua
	H_DESAGUE	% de hogares sin acceso a red pública de alcantarillado	Continua
	H_ALUMB	% de hogares sin alumbrado por red pública	Continua
	P_AGUA	% de población sin acceso a red publica de agua	Continua
	P_DESAGUE	% de poblacion sin acceso a red pública de alcantarillado	Continua
	P_VIV_INAD	% de poblacion en viviendas inadecuadas	Continua
	P_HACINAMIENTO	% de poblacion en viviendas con hacinamiento	Continua
	P_SSHH	% de poblacion en viviendas sin servicios higiénicos	Continua
	P_NEC_BAS	% de pob. con al menos una necesidad básica insatisfecha	Continua
	P_1_NEC_BAS	% de población con 1 necesidad básica insatisfecha	Continua
	P_2_NEC_BAS	% de población con 2 necesidades básicas insatisfechas	Continua
	P_3_NEC_BAS	% de población con 3 necesidades básicas insatisfechas	Continua
	P_4_NEC_BAS	% de población con 4 necesidades básicas insatisfechas	Continua
	P_5_NEC_BAS	% de población con 5 necesidades básicas insatisfechas	Continua
	H_VIV_INAD	% de hogares en viviendas inadecuadas	Continua
	H_HACINAMIENTO	% de hogares en viviendas con hacinamiento	Continua
	H_SSHH	% de hogares en viviendas sin servicios higiénicos	Continua
	H_NEC_BAS	% de hogares con al menos una nec. básica insatisfecha	Continua
	H_1_NEC_BAS	% de hogares con 1 necesidad básica insatisfecha	Continua
	H_2_NEC_BAS	% de hogares con 2 necesidades básicas insatisfechas	Continua
	H_3_NEC_BAS	% de hogares con 3 necesidades básicas insatisfechas	Continua
	H_4_NEC_BAS	% de hogares con 4 necesidades básicas insatisfechas	Continua
	H_5_NEC_BAS	% de hogares con 5 necesidades básicas insatisfechas	Continua
	P_ALUMB	% de población sin alumbrado por red pública	Continua
Salud	MORTALIDAD	Tasa de mortalidad infantil	Continua
	FECUNDIDAD	Tasa global de fecundidad	Continua
	DESNUTRICION	Desnutrición crónica en menores de 5 años	Continua
Educacion	P_SIN_ESCUELA	% de poblacion en hogares con niños sin escuela	Continua
	H_SIN_ESCUELA	% de hogares con niños sin la escuela	Continua
	ANALFABET	Tasa de analfabetismo	Continua
	ANALFABET_H	Tasa de analfabetismo en hombres	Continua
	ANALFABET_M	Tasa de analfabetismo en mujeres	Continua

Cuadro 5.1: Variables analizadas para el modelo propuesto

Capítulo 6

Conclusiones

6.1. Conclusiones

En este trabajo se presentó un modelo de regresión cuantílica para variables acotadas entre (0,1) en datos de áreas. La ventaja principal es que permite modelar efectos fijos y efectos espaciales en los cuantiles de una variable respuesta la cual tiene la particularidad de representar una característica de un área y cuyo valor está acotado entre (0,1). Debido a que el modelo es muy complejo pues se presenta un efecto espacial por cada área observada la inferencia se realizó desde la perspectiva bayesiana y se ha optado por usar el método MCMC mediante el paquete RSTAN. Para representar a la variable respuesta se utilizó la distribución Kumaraswamy sobre la que se ha profundizado en sus propiedades y en la demostración de la reparametrización que permite expresarla en función de los cuantiles.

Basados en el estudio de simulación realizado sobre diferentes escenarios de los parámetros del modelo, se concluye que el método de inferencia utilizado permite obtener un muy buen ajuste y en tiempos razonables a pesar de la complejidad del modelo. Se ha podido confirmar que ante distintos valores para la precisión y los cuantiles del modelo se obtienen buenos resultados.

El modelo propuesto ha sido aplicado a datos reales utilizando una base de datos construida a partir de la publicación del mapa de pobreza en el año 2009 del INEI. Se ha relacionado el porcentaje de pobreza no extrema con el porcentaje de viviendas en mal estado y el porcentaje de hogares con hacinamiento en los distritos de la provincia de Lima. El modelo permitió identificar la fuerte relación que existe entre estas tres variables e incluyendo el efecto espacial se puede obtener mucho mejores estimaciones para el modelo.

6.2. Sugerencias para investigaciones futuras

- Se podrían considerar otras distribuciones además de la Kumaraswamy de manera que para ciertos datos acotados entre (0,1) se pueden obtener modelos más robustos.
- Se podría considerar modelos inflacionados en 0 o en 1.
- Se podría considerar también la influencia temporal para el modelo además de la espacial.

Apéndice A

Resultados teóricos

A.1. Demostración de la reparametrización de la distribución Kumaraswamy

Sea la variable aleatoria y con distribución Kumaraswamy de parámetros α y β, su función de densidad de probabilidad es dada por

$$f_Y(y \mid \alpha, \beta) = \alpha\beta(y)^{\alpha-1}(1-(y)^{\alpha})^{\beta-1}, 0 \leq y \leq 1 \quad , \alpha, \beta > 0. \tag{A.1}$$

Su función de distribución acumulada es dada por

$$F(y) = 1 - (1-(y)^{\alpha})^{\beta}.$$

Y la función del cuantil q es dada por

$$\kappa(q) = F^{-1}(q) = \{1-(1-q)^{1/\beta}\}^{1/\alpha}, \tag{A.2}$$

donde q es un valor prefijado.
De (A.2) se puede despejar los parámetros α y β:

$$\alpha = \frac{\log(1-(1-q)^{1/\beta})}{\log(\kappa)} \tag{A.3}$$

$$\beta = \frac{\log(1-q)}{\log(1-\kappa^{\alpha})}, \tag{A.4}$$

por lo tanto

$$\frac{1}{\alpha} = \frac{\log(\kappa)}{\log(1-(1-q)^{1/\beta})} \tag{A.5}$$

$$\frac{1}{\beta} = \frac{\log(1-\kappa^{\alpha})}{\log(1-q)} \tag{A.6}$$

De acuerdo a las demostraciones realizadas en Mitnik y Baek (2011), se tiene que (A.5) y (A.6) son medidas de dispersión de la distribución Kumaraswamy. Dado que en (A.5) y (A.6) se mide la dispersión de la distribución, y quitando las partes constantes para un q fijo, se

tendrían expresiones para el segundo parámetro ϕ de la función de distribución alternativa. Para el presente estudio, tal como también se ha realizado en Bayes et al. (2017), se utilizará como medida de precisión la expresión obtenida de (A.5).
Por lo tanto se tiene la siguiente expresión para ϕ:

$$\phi = -\log(1-(1-q)^{1/B}). \tag{A.7}$$

Usando (A.7) se puede despejar β en función de ϕ

$$\beta = \frac{\log(1-q)}{\log(1-e^{-\phi})}. \tag{A.8}$$

Además de (A.7) y (A.3) ya se tiene la siguiente expresión para α en función de ϕ y κ

$$\alpha = -\frac{\phi}{\log(\kappa)}. \tag{A.9}$$

Se reemplaza α y β en la función de distribución de Kumaraswamy usual (A.1) usando las igualdades de (A.8) y (A.9) obteniendo la reparametrización mencionada en 2.2.1

$$f_Y(y|\kappa,\phi) = -\frac{\log(1-q)\phi}{\log(1-e^{-\phi})\log(\kappa)} y^{-\frac{\phi}{\log(\kappa)}-1}\{1-y^{-\frac{\phi}{\log(\kappa)}}\}^{\frac{\log(1-q)}{\log(1-e^{-\phi})}-1}.$$

Apéndice B

Figuras

B.1. Estudio de simulación: gráficos de cadenas

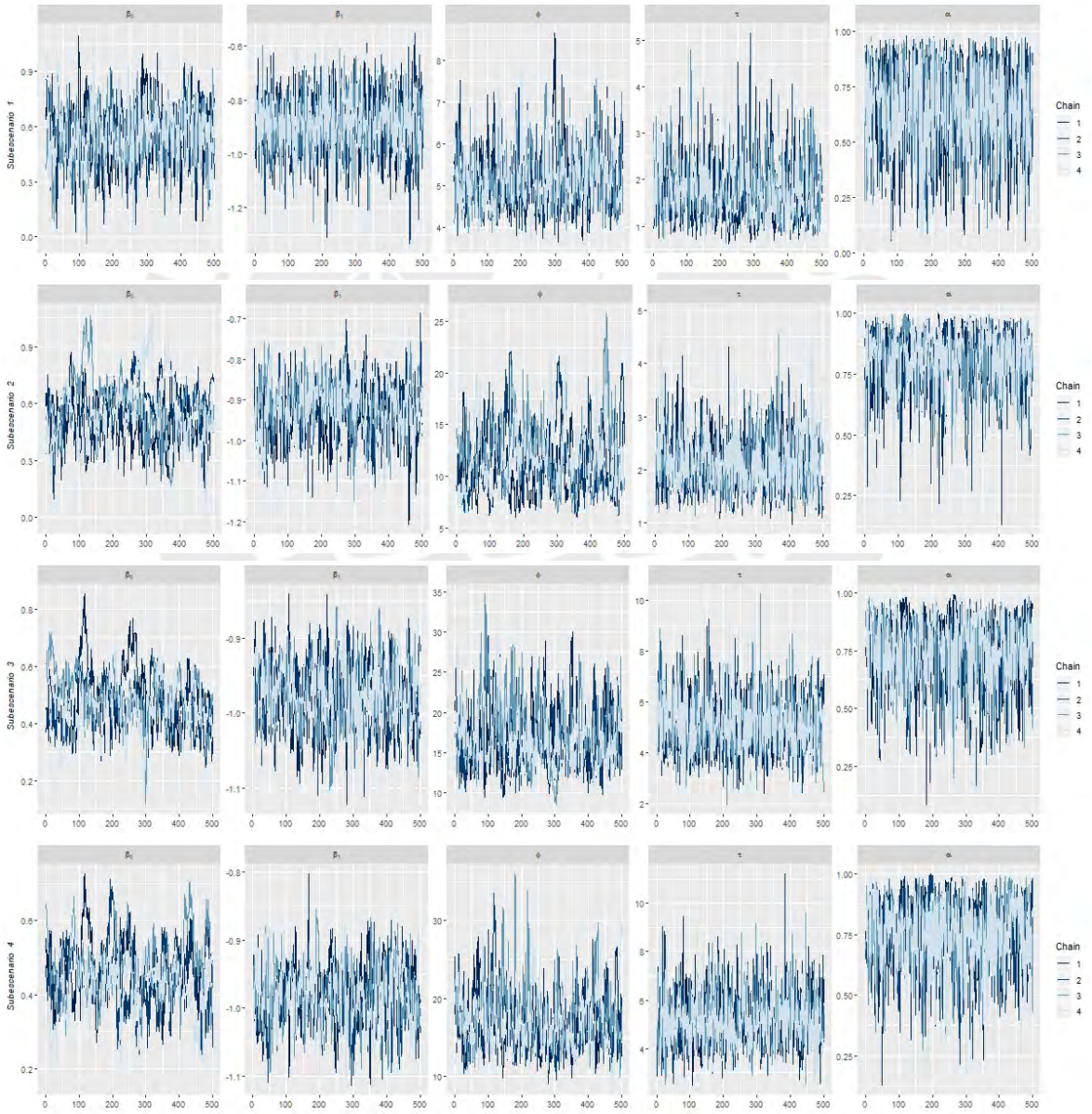


Figura B.1: Gráfico de cadenas de las simulaciones a posteriori de los parámetros para el primer escenario $q = 0.1$.

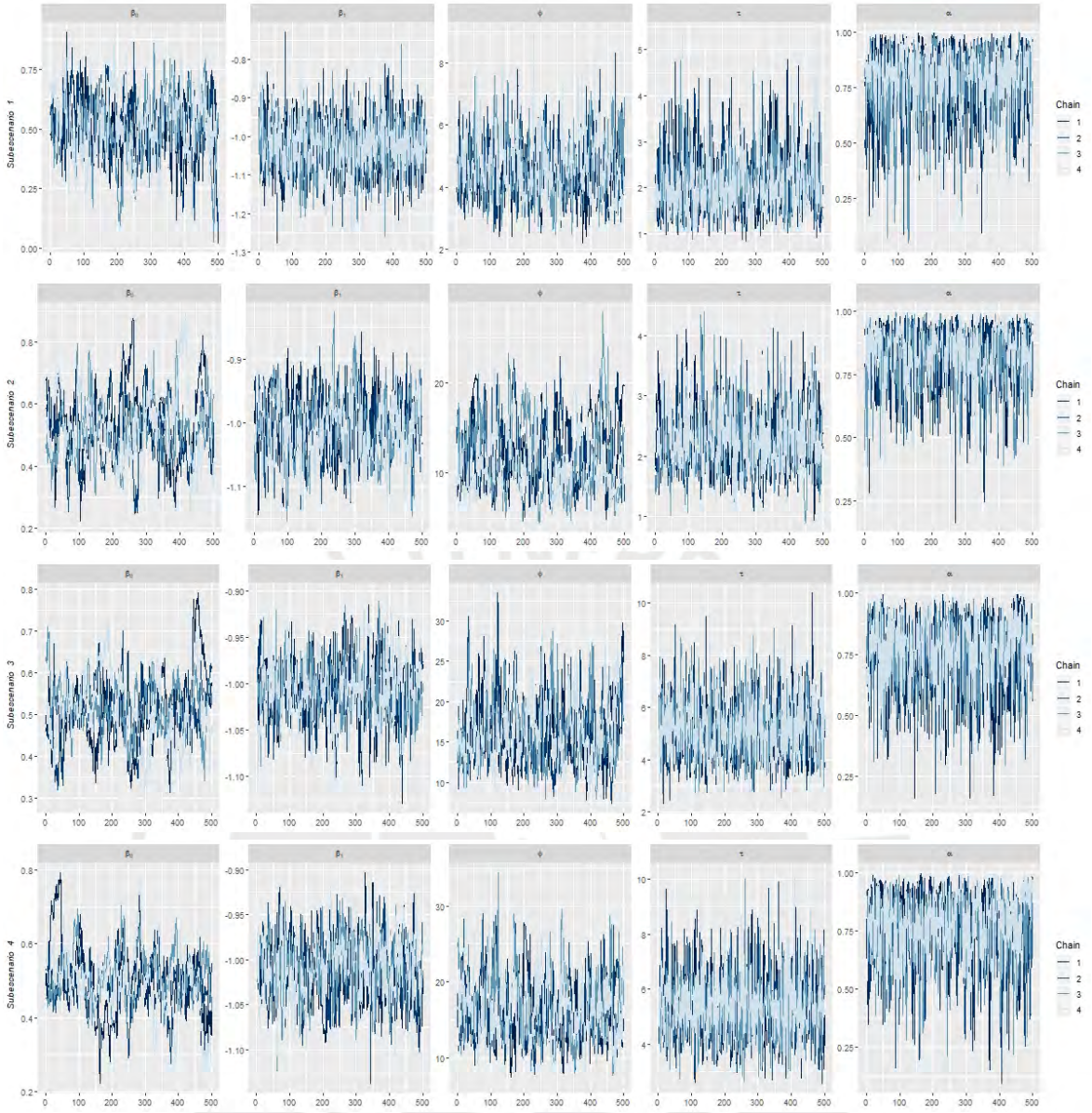


Figura B.2: Gráfico de cadenas de las simulaciones a posteriori de los parámetros para el segundo escenario $q = 0.5$.

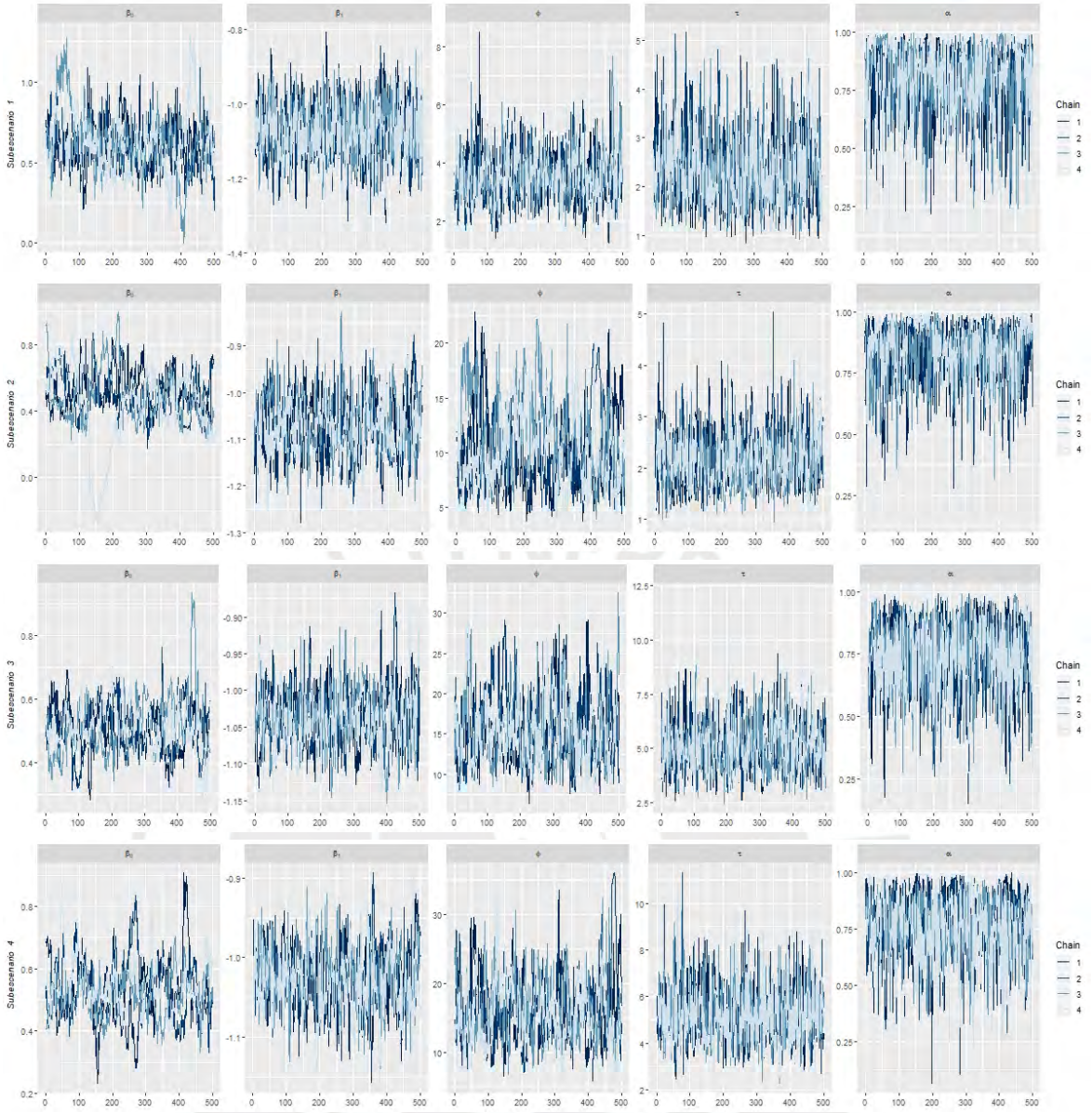


Figura B.3: Gráfico de cadenas de las simulaciones a posteriori de los parámetros para el tercer escenario $q = 0.9$.

B.2. Estudio de simulación: histogramas

En las Figuras B.4 hasta B.6, se puede observar los histogramas de las simulaciones a posteriori de cada parámetro e hiperparámetro. Las líneas en azul corresponden a los intervalos de credibilidad al 95 % mientras que la línea roja indica el valor real.

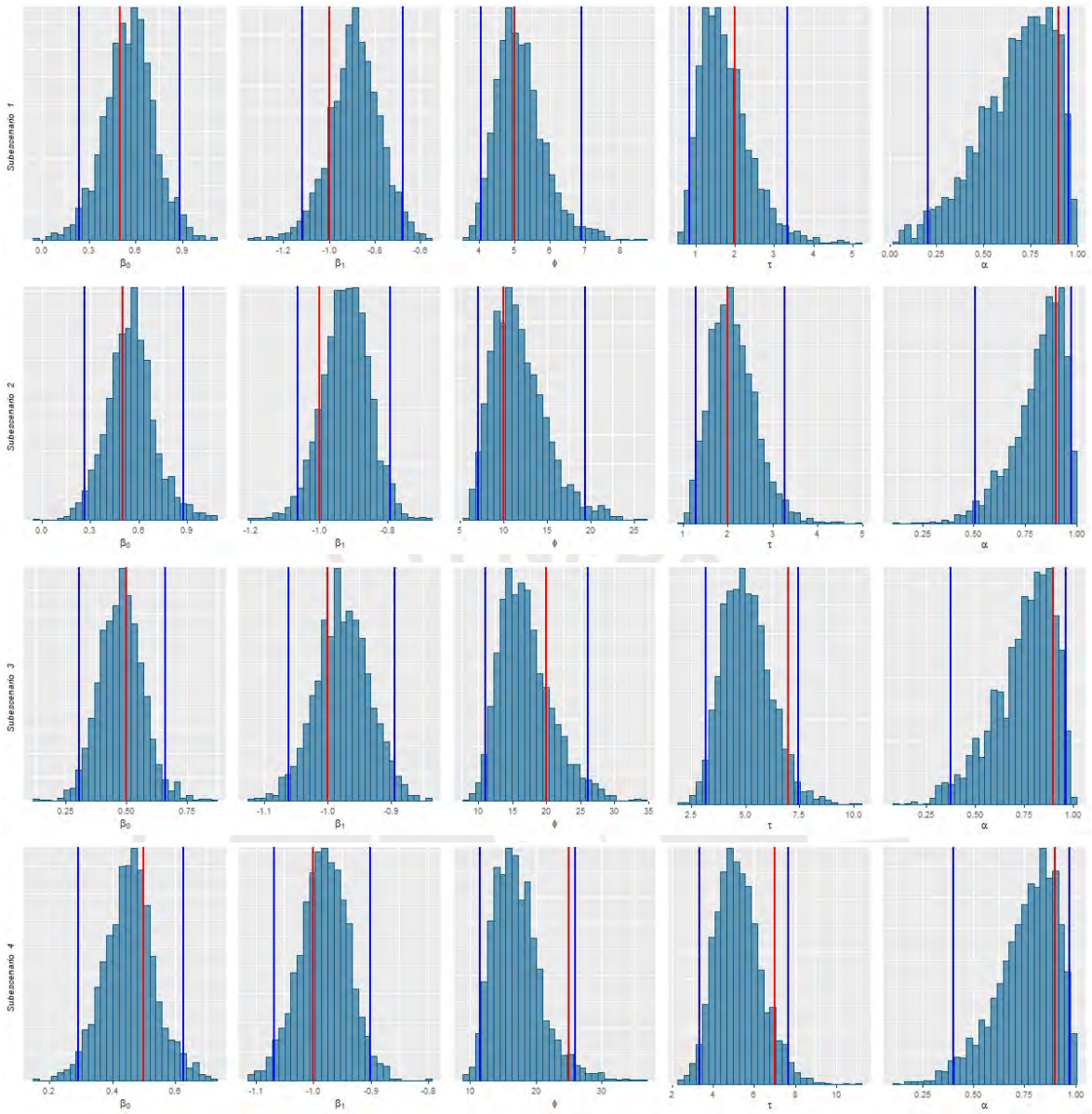


Figura B.4: Histograma de las simulaciones a posteriori de los parámetros para el primer escenario $q = 0.1$. La línea en rojo representa el valor real mientras las lineas azules de los extremos representa el intervalo de credibilidad al 95 %.

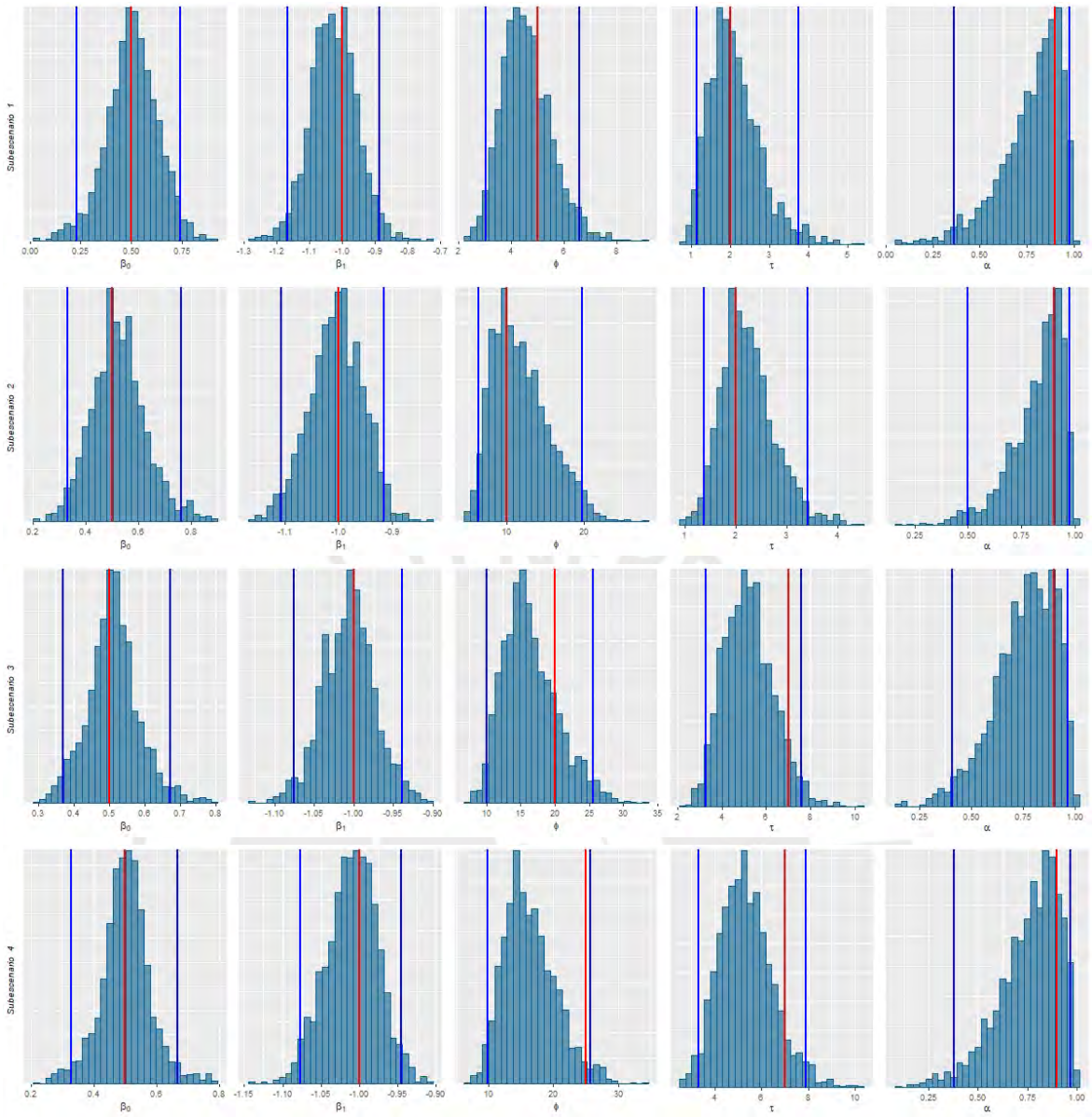


Figura B.5: Histograma de las simulaciones a posteriori de los parámetros para el segundo escenario $q = 0.5$. La línea en rojo representa el valor real mientras las lineas azules de los extremos representa el intervalo de credibilidad al 95 %.

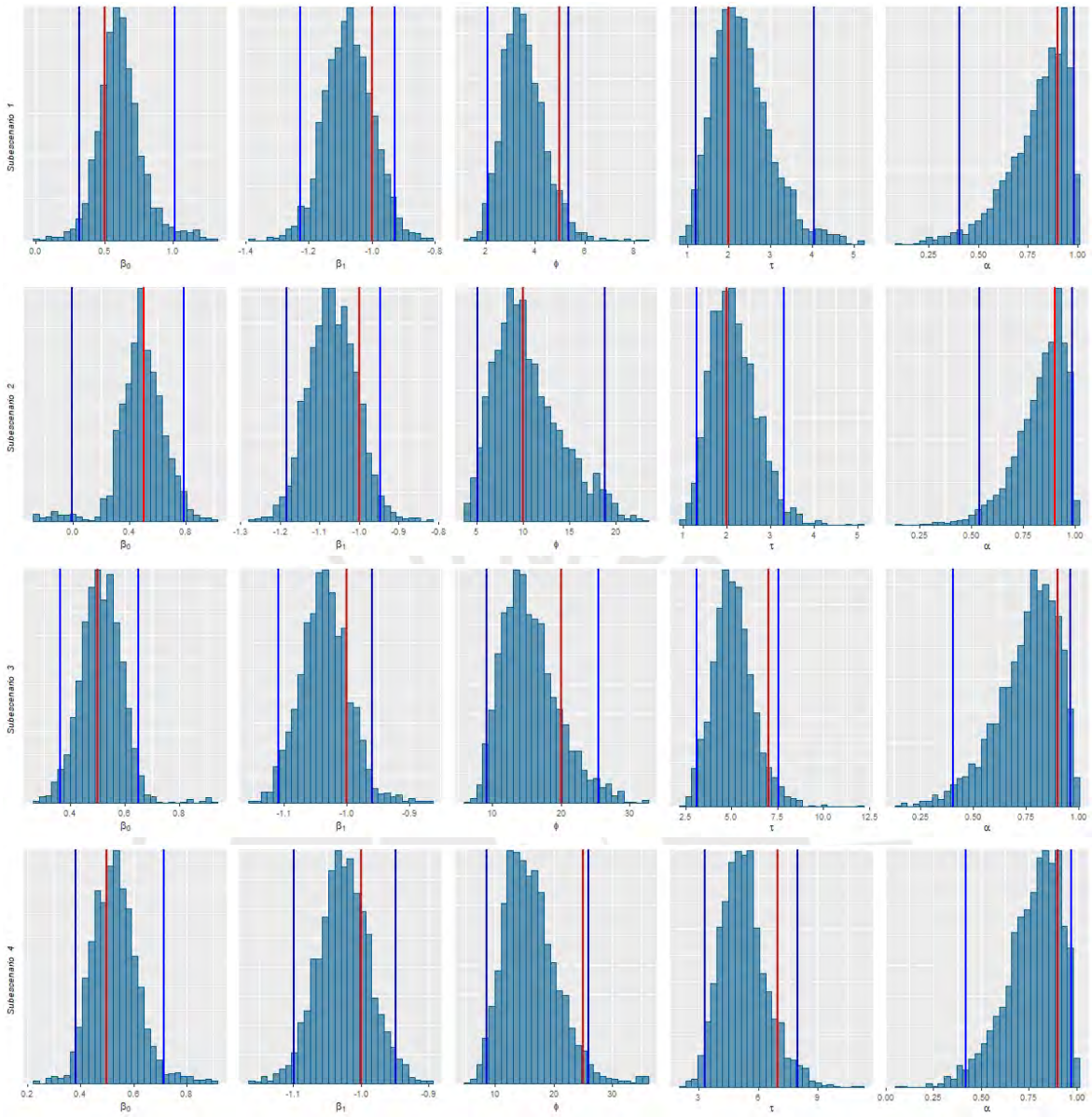


Figura B.6: Histograma de las simulaciones a posteriori de los parámetros para el tercer escenario $q = 0.9$. La línea en rojo representa el valor real mientras las lineas azules de los extremos representa el intervalo de credibilidad al 95 %.

B.3. Aplicación a la incidencia de pobreza no extrema

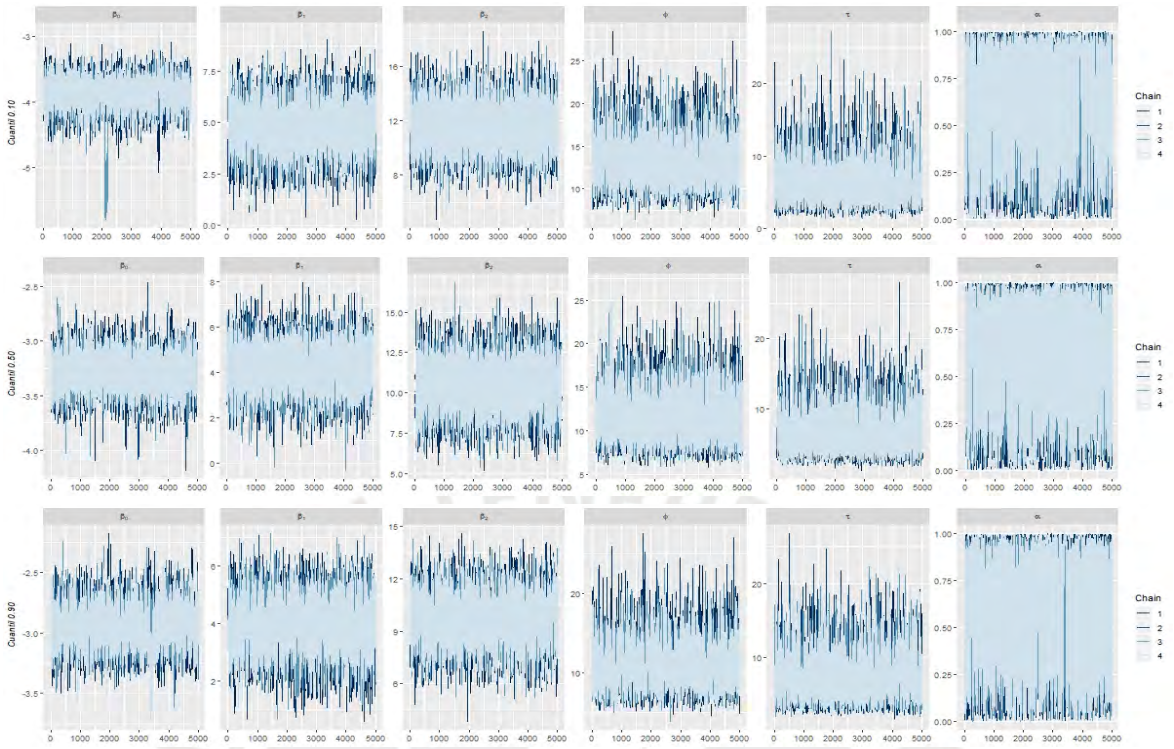


Figura B.7: Gráfico de cadenas de las simulaciones a posteriori de los parámetros para el modelo KSQ-CAR.

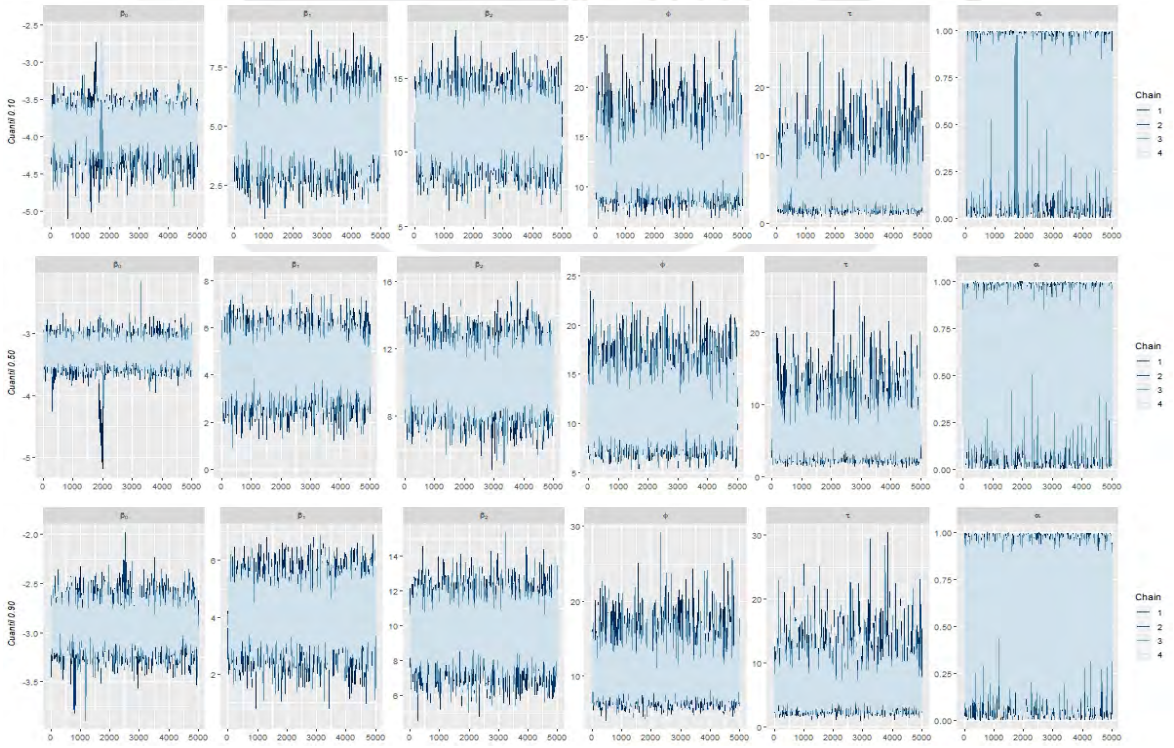


Figura B.8: Gráfico de cadenas de las simulaciones a posteriori de los parámetros para el modelo KSQ-SPOCK.

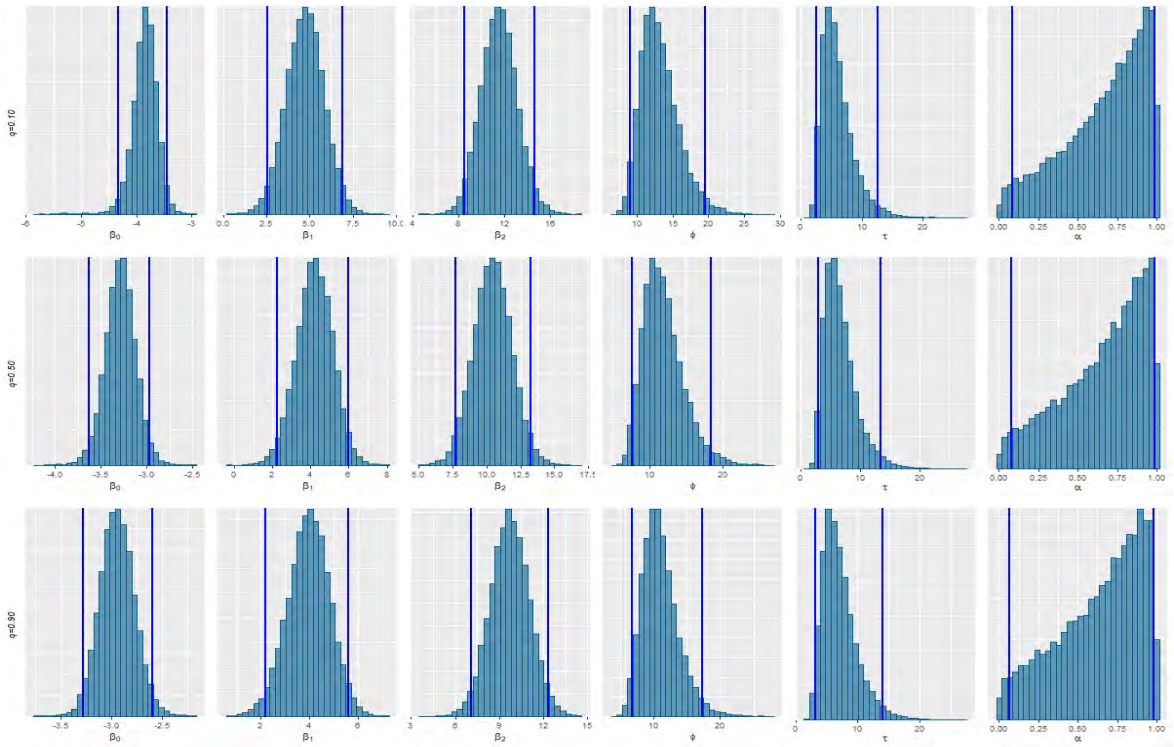


Figura B.9: Histograma de las simulaciones a posteriori de los parámetros para el modelo KSQ-CAR.

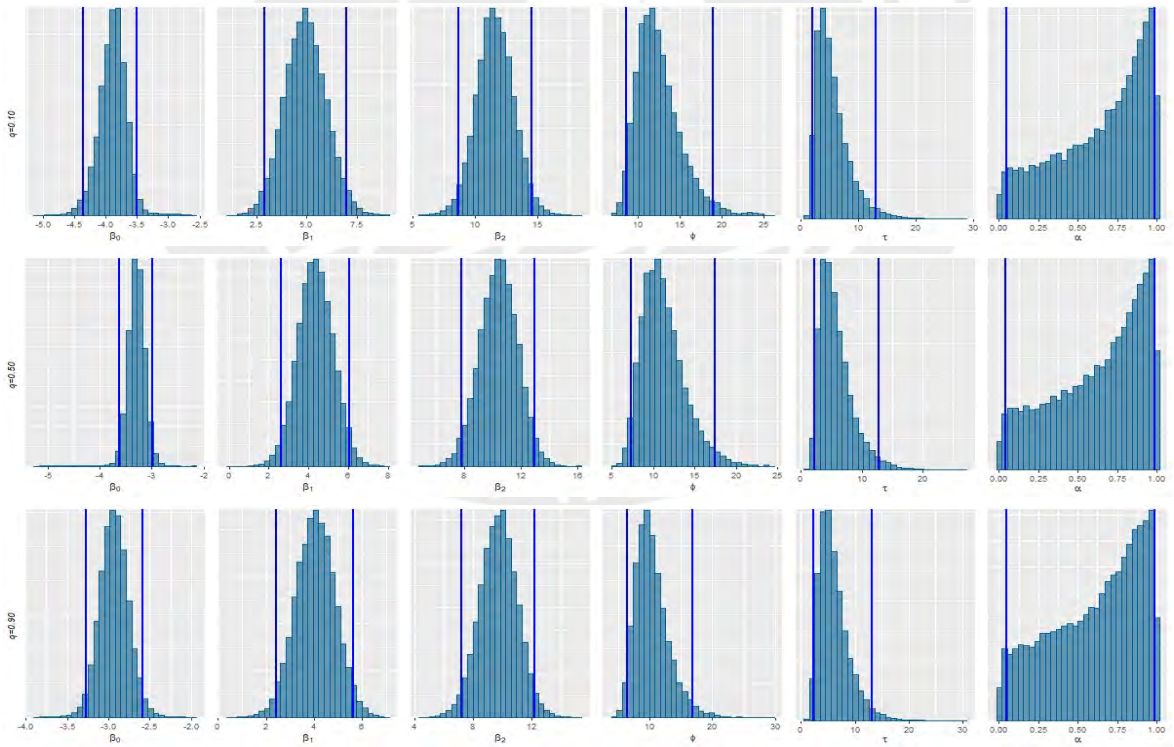


Figura B.10: Histograma de las simulaciones a posteriori de los parámetros para el modelo KSQ-SPOCK.

Bibliografía

Ali, M., Mahmoud, M. R. y ElSherbin, E. A. (2015). The new kumaraswamy family of generalized distributions with application, *Pakistan Journal of Statistics and Operation Research* **11**(2): 159–180.

Banco Central de Reserva (1986). Mapa de pobreza del perú 1981, *Technical report*, Banco Central de Reserva. Subgerencia de ingreso y producto.

Bayes, C., Bazan, J. y de Castro, M. (2017). A quantile parametric mixed regression model for bounded response variables, *Statistics and its interface* **10**: 483–493.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society. Series B* **36**(2): 192–236.

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

Carrasco, J. M., Ferrari, S. L. y Cordeiro, G. M. (2010). A new generalized kumaraswamy distribution, *Technical report*, Departamento de Estatistica, Universidade de São Paulo.

Dey, D. K., Chen, M.-H. y Chang, H. (1997). Bayesian approach for nonlinear random effects models, *Biometrics* **53**(4): 1239–1252.
URL: *http://www.jstor.org/stable/2533493*

Elbers, C., Lanjouw, J. O. y Lanjouw, P. (2003). Micro–Level Estimation of Poverty and Inequality, *Econometrica* **71**(1): 355–364.
URL: *https://ideas.repec.org/a/ecm/emetrp/v71y2003i1p355-364.html*

Elbers, C., Lanjouw, P., Mistiaen, J., Ozler, B. y Simler, K. (2003). Are Neighbours Equal? Estimating Local Inequality in Three Developing Countries, *WIDER Working Paper Series 052*, World Institute for Development Economic Research (UNU-WIDER).
URL: *https://ideas.repec.org/p/unu/wpaper/dp2003-52.html*

Fahrmeir, L., Kneib, T., Lang, S. y Marx, B. (2013). *Regression. Models, Methods and Applications*, 1 edn, Springer-Verlag Berlin Heidelberg.

Feres, J. y Mancero, X. (2015). Enfoques para la medición de la pobreza: Breve revisión de la literatura, *Technical report*, OECD Publishing, Paris.

Foster, J., Greer, J. y Thorbecke, E. (1984). A class of decomposable poverty measures, *Econometrica* **52**(3): 761–766.
URL: *http://www.jstor.org/stable/1913475*

Geisser, S. y Eddy, W. F. (1979). A predictive approach to model selection, *Journal of the American Statistical Association* **74**(365): 153–160.
URL: *http://www.jstor.org/stable/2286745*

Gelfand, A. E., Diggle, P., Guttorp, P. y Fuentes, M. (2010). *Handbook of Spatial Statistics*, CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S. y Rubin, D. B. (2004a). *Bayesian data analysis*, Texts in statistical science, Chapman & Hall.

Gelman, A., Carlin, J. B., Stern, H. S. y Rubin, D. B. (2004b). *Bayesian Data Analysis*, 2nd ed. edn, Chapman and Hall/CRC.

Hentschel, J. y Lanjouw, P. (1996). Constructing an indicator of consumption for the analysis of poverty. principles and illustrations with reference to ecuador, *Working papers*, World Bank - Living Standards Measurement.
URL: *https://EconPapers.repec.org/RePEc:fth:wobali:127*

Hoff, P. D. (2010). *A First Course in Bayesian Statistical Methods*, Springer.

INEI (1996). Metodología para determinar el ingreso y la proporción de hogares pobres, *Technical report*, Instituto Nacional de Estadística e Informática.

INEI (2010). Mapa de pobreza provincial y distrital 2009, *Technical report*, Instituto Nacional de Estadística e Informática.

INEI (2015). Mapa de pobreza provincial y distrital 2013, *Technical report*, Instituto Nacional de Estadística e Informática.

INEI (2018a). Evolucion de la pobreza monetaria 2008-2017, *Technical report*, Instituto Nacional de Estadística e Informática.

INEI (2018b). Mapa de necesidades básicas insatisfechas 1993,2007 y 2017, *Technical report*, Instituto Nacional de Estadística e Informática.

INEI (2019). Evolución de la pobreza monetaria 2007-2018, *Technical report*, Instituto Nacional de Estadística e Informática.

Koenker, R. y Bassett, G. (1978). Regression quantiles, *Econometrica* **46**(1): 33–50.

Kuiper, J. (2003). Modelización del mapa de pobreza para el año 2001, *Technical report*, Instituto Nacional de Estadística e Informática.

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes, *Journal of Hydrology* **46**(1,2): 79–88.

Kunsch, H. R. (1987). Intrinsic autoregressions and related models on the two-dimensional lattice, *Biometrika* **74**(3): 517–524.

MEF (2001). Hacia la búsqueda de un nuevo instrumento de focalización de recursos destinados a la inversión social adicional en el marco de la lucha contra la pobreza, *Technical report*, Ministerio de Economía y Finanzas.

Mitnik, P. y Baek, S. (2011). The kumaraswamy distribution: Median-dispersion reparameterizations for regression modeling and simulation-based estimation, *Statistical Papers* **54**.

Mkhandi, S., Kachroo, R. y Guo, S. G. (1996). Uncertainty analysis of flood quantile estimates with reference to tanzania, *Journal of Hydrology* **185**(1-4): 317–333.

Neal, R. (2012). Mcmc using hamiltonian dynamics, *Handbook of Markov Chain Monte Carlo* .

Prates, M. O., Assunção, R. M. y Rodrigues, E. C. (2019). Alleviating spatial confounding for areal data problems by displacing the geographical centroids, *Bayesian Analysis* **14**(2): 623–647.
URL: *https://doi.org/10.1214/18-BA1123*

Robert, C. y Casella, G. (2004). *Monte Carlo Statistical Methods*, Springer.

Rue, H. y Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall Monographs on Statistics and Applied Probability, Chapman & HallCRC.

Shoukri, M. M., Mian, I. U. M. y Tracy, D. S. (1988). Sampling properties of estimators of the log-logistic distribution with application to canadian precipitation data, *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **16**(3): 223–236.
URL: *http://www.jstor.org/stable/3314729*

Taillardat, M. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics, *The American Meteorological Society* . Journals online.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research* **11**.

Yu, K. y Moyeed, R. A. (2001). Bayesian quantile regression, *Statistics and Probability Letters* **54**(4): 437–447.