PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Título

IDENTIFICACIÓN DE CONGLOMERADOS ESPACIALES DE ACUERDO A NIVELES DE MOROSIDAD DE EMPRESAS EN EL PERÚ

TESIS PARA OPTAR POR EL GRADO ACADÉMICO DE MAGÍSTER EN ESTADÍSTICA

AUTOR

Alex Edward Tristán Gómez

ASESORA

Dra. Zaida Jesús Quiroz Cornejo

Agosto, 2021

Dedicatoria

Esta tesis esta dedicada a mis padres José y Gladys quienes con su amor, paciencia y esfuerzo me han permitido llegar a cumplir hoy un sueño más.

A mis hermanos, durante todo este proceso, por estar conmigo dándome consejos y palabras de aliento que hicieron de mi una mejor persona y de una u otra forma me acompañan en todos mis sueños y metas.

Agradecimentos

Mi profundo agradecimiento a la Pontificia Universidad Católica del Perú, a mis profesores, en especial a la Dra. Zaida Quiróz, quienes con la enseñanza de sus valiosos conocimientos hicieron que pueda crecer día a día como profesional, gracias a cada una de ustedes por su paciencia, dedicación, apoyo incondicional y amistad.

Resumen

El cumplimiento de las obligaciones financieras que tienen las empresas es respaldado por una correcta gestión de riesgo de crédito, esto evita problemas de liquidez y solvencia. Por ello es importante detectar los niveles de riesgo de morosidad en las empresas. La presente tesis tiene como objetivo identificar conglomerados de provincias del Perú, en función de la tasa de incumplimiento de pagos, conocida también como la tasa de morosidad. Para ello se propone un modelamiento en dos niveles. En el primer nivel se usan modelos aglomerativos jerárquicos para seleccionar n conglomerados candidatos a priori, donde el número final de conglomerados se escoge mediante criterios de selección de modelos. Posteriormente, en un segundo nivel, modelaremos el nivel de riesgo haciendo uso del modelo de Poisson y prioris condicionales autoregresivas en base a los conglomerados definidos en el primer nivel e incluyendo covariables. Los modelos pueden ser reescritos como modelos Gaussianos latentes, y se puede usar inferencia bayesiana para estimar sus parámetros, específicamente a través de la aproximación de Laplace anidada integrada. Finalmente, como resultado de la aproximación se obtienen conglomerados de provincias de acuerdo a sus niveles de morosidad, permitiendo clasificar las provincias en conglomerado de alto, medio y bajo nivel de riesgo de morosidad.

Palabras-clave: CAR, identificación de conglomerados, INLA, modelos gaussianos latentes, morosidad financiera.

Abstract

Compliance with the financial obligations of companies is ensured by proper credit risk management, this avoids liquidity and solvency problems. For this reason, it is important to identify the risk level of default in peruvian companies. The goal of this thesis is to identify clusters of provinces of Perú with regard to the default rate of payments, also known as probability of default. Thus it is proposed a model in two stages. In the first stage hierarchical agglomerative models select prior candidate clusters, and the final number of clusters is selected through selection criteria of models. In the second stage it is proposed the Poisson model considering autoregressive conditional prioris, the clusters defined in the first stage, and also including covariates. This model fill in the class of Gaussian latent models, therfore its paremeters were estimated using bayesian inference, specifically through integrated nested Laplace approximation. Finally, as a result, we found clusters in accordance with the default level, allowing to classify provinces into clusters of high, medium and low risk level.
Keywords: CAR, cluster identification, financial defaulting, INLA, latent Gaussian models.

Índice general

Lista de abreviaturas VIII

Índice de figuras IX

Índice de cuadros XI

1. Introducción **1**
- 1.1. Consideraciones preliminares . . . 1
- 1.2. Objetivo principal . . . 2
- 1.3. Objetivos específicos . . . 2
- 1.4. Organización del trabajo . . . 3

2. Conceptos preliminares **4**
- 2.1. Análisis de conglomerados (Clusters) . . . 4
  - 2.1.1. Métodos de análisis de conglomerados . . . 5
- 2.2. Modelo de regresión Poisson para datos de áreas . . . 7
- 2.3. Estadística espacial para datos de áreas . . . 8
  - 2.3.1. Análisis exploratorio para datos de áreas . . . 8
  - 2.3.2. Medidas de asociación espacial . . . 9
  - 2.3.3. Campos aleatorios gaussianos de Markov (GMRF) . . . 9
  - 2.3.4. Modelos autorregresivos condicionales (CAR) . . . 10
- 2.4. Inferencia bayesiana . . . 11
  - 2.4.1. Método de las cadenas de markov de Monte Carlo (MCMC) . . . 12
- 2.5. Aproximación de Laplace anidada integrada (INLA) . . . 13
  - 2.5.1. Introducción . . . 13
  - 2.5.2. La aproximación de Laplace . . . 13
  - 2.5.3. Modelos gaussianos latentes . . . 13
  - 2.5.4. La aproximación de Laplace anidada integrada (INLA) . . . 14

3. Desarrollo del modelo **16**
- 3.1. Introducción a los modelos propuestos . . . 16
- 3.2. Modelo de Poisson para datos de áreas . . . 16
  - 3.2.1. Inferencia bayesiana . . . 18
  - 3.2.2. Estimación del modelo . . . 19
- 3.3. Modelo Poisson por conglomerados para datos de áreas . . . 20
  - 3.3.1. Inferencia bayesiana . . . 24

3.3.2. Estimación del modelo . . . 25

4. Estudio de Simulación **27**
4.1. Generación de datos . . . 27

5. Aplicación **36**
5.1. Riesgo de crédito . . . 36
5.1.1. Definición e importancia . . . 36
5.1.2. Tasa de incumplimiento (PD) de empresas del Perú en el sistema financiero . . . 37
5.2. Análisis exploratorio de los datos . . . 37
5.2.1. Medidas de asociación espacial . . . 39
5.3. Resultados . . . 41
5.3.1. Modelo de Poisson para datos de provincias . . . 42
5.3.2. Modelo de Poisson por conglomerados para datos de provincias . . . 45

6. Conclusiones **56**
6.1. Conclusiones . . . 56
6.2. Sugerencias para investigaciones futuras . . . 56

A. Resultados teóricos **57**

B. Test Getis-Ord **59**

C. Código de aplicación **60**
C.0.1. Matriz adyacente para Perú . . . 60
C.0.2. Etapa 1 - Aplicando evaluación de conglomerados . . . 61
C.0.3. Etapa 2 - Seleccionando el mejor modelo usando INLA . . . 61
C.0.4. Seleccionado las variables explicativas . . . 61
C.0.5. Guardando los criterios de información . . . 62
C.0.6. Cálculo de RMSE . . . 62
C.0.7. Porcentaje de Identificación . . . 62

Bibliografía **63**

Lista de abreviaturas

fdp	Función de densidad de probabilidad .
pBF	Pseudo factor de Bayes(*Pseudo bayes factor*).
MCMC	Cadena de Markov - Monte Carlo
INLA	Aproximación de Laplace Anidada Integrada.
PD	Porcentaje de Default (impago).

Índice de figuras

1.1. Niveles de riesgo, medido por la tasa de morosidad para provincias del Perú en el año 2018, el nivel de riesgo que va de menor a mayor magnitud está representado por los colores verde y rojo , en ese orden respectivamente. . . 2

2.1. linkage simple (vecino más cercano) . . . 6
2.2. linkage centroide . . . 6
2.3. Método de Ward (método de mínima varianza) . . . 7

3.1. Organización de la propuesta conformada por 2 etapas . . . 21

4.1. Grafo de las principales regiónes de Escocia . . . 28
4.2. Matriz de vecindad de las regiónes de Escocia, considerando el vecino más próximo . . . 28
4.3. Simulación de la variable de estudio considerando los efectos espaciales en cada región de Escocia, donde se muestra niveles de riesgo relacionado a la variable de estudio. La lectura va mayor a menor riesgo (número de casos relacionados a la variable Y), es decir, está representado por matices de los colores rojo y amarillo, en ese orden respectivamente. . . . 29
4.4. Selección del número óptimo de conglomerados mediante menor (a) DIC, (b) WAIC y (c) LPML. . . . 30
4.5. Distribución marginal a posteriori del parámetro β_0 en cada uno de los escenarios. 31
4.6. Distribución marginal a posteriori del parámetro β_1 en cada uno de los escenarios. 32
4.7. Distribución marginal a posteriori del parámetro τ_ϕ en cada uno de los escenarios. 32
4.8. Valores reales del vector de parámetros α simulado versus la media a posteriori del parámetro α en cada uno de los escenarios. . . . 33
4.9. IC al 95 % de los efectos aleatorios espaciales en cada escenario . . . 33
4.10. Gráfico de dispersión de los valores simulados vs estimaciones de la variable de estudio Y en cada escenario . . . 34
4.11. Comparativo de los valores simulados vs valores obtenidos en el Escenario 1 35
4.12. Comparativo de los valores simulados vs valores obtenidos en el Escenario 2 35
4.13. Comparativo de los valores simulados vs valores obtenidos en el Escenario 3 35

5.1. Tasa de incumplimiento en el sistema financiero - Empresas del Perú (201801 - 201903) . . . 37
5.2. Tasa de incumplimiento por Departamentos - Empresas del Perú (201801 - 201903) Fuente: Elaboración propia . . . 39

5.3. Gráficos de la covariable ratio de estado del RUC no activo 41
5.4. Gráficos de la covariable antigüedad de RUC 42
5.5. Matriz de correlaciones 42
5.6. Gráfico de las distribuciones marginales a posteriori de los coeficientes β_0,β_1 y β_2 para el modelo Poisson - CAR 44
5.7. Gráfico de las distribuciones marginales a posteriori de los coeficientes β_3, β_4 e hiperparámetro de precisión τ_ϕ, para el modelo Poisson - CAR 45
5.8. Intervalos de credibilidad al 95 % de los efectos aleatorios espaciales del modelo Poisson - CAR 46
5.9. Mapa comparativo de PD reales vs PD estimadas por el modelo Poisson - CAR 46
5.10. Comparativo de los valores reales vs valores estimados con el modelo Poisson - CAR 47
5.11. Comparativo de conglomerados considerando 5 y 7 grupos respectivamente . 48
5.12. Gráfico resumen de indicadores estadísticos para la selección del número óptimo de conglomerados 49
5.13. Gráfico de las distribuciones marginales a posteriori de los coeficientes β_0,β_1 y β_2 del modelo Poisson - CAR con conglomerados 51
5.14. Gráfico de las distribuciones marginales a posteriori de los coeficientes β_3, β_4 e hiperparámetros τ_ϕ y α_j del modelo Poisson - CAR con conglomerados. . . 52
5.15. Intervalos de credibilidad al 95 % de los efectos aleatorios espaciales del modelo Poisson - CAR incluyendo conglomerados a priori 53
5.16. Mapa de las provincias del Perú - comparativo de los tasas de incumplimiento (PD) reales vs estimadas por el modelo Poisson - CAR incluyendo conglomerados a priori 53
5.17. Comparativo de los valores reales vs valores estimados con el modelo Poisson - CAR incluyendo conglomerados a priori 54
5.18. Conglomerados identificados para las provincias del Perú, formados por asociación del riesgo de morosidad, donde podemos diferenciar conglomerados de alto y bajo nivel de riesgo, representados por colores rojo y amarillo, respectivamente. 55

Índice de cuadros

4.1. Tabla resumen de resultados, DIC y tiempo de procesamiento en cada escenario 29
4.2. Tabla resumen de resultados, entre ellos el valor real, media, desviación estándar e IC al 95 % de los parámeros evaluados en los 3 escenarios. 31
4.3. Tabla resumen de resultados, porcentajes de detección de conglomerados . . . 34

5.1. Tabla resumen de departamentos del Perú con mayor PD 38
5.2. Tabla resumen de departamentos del Perú con menor PD 39
5.3. Test I de Moran 40
5.4. Test C de Geary 40
5.5. Test Getis-Ord 40
5.6. Tabla resumen de variables analizadas para el modelo propuesto 41
5.7. Tabla resumen de resultados, entre ellos los valores de la media, desviación estándar e IC al 95 % de los parámetros encontrados en la primera propuesta (modelo Poisson - CAR)) 43
5.8. Tabla resumen de resultados - RSME, DIC y WAIC del modelo Poisson - CAR 44
5.9. Tabla resumen - Comparación de coeficientes de aglomeración 47
5.10. Conglomerados bajo el criterio Ward. 48
5.11. Conglomerados bajo el criterio RCE y de Ward. 48
5.12. RSME, DIC y WAIC del modelo Poisson - CAR incluyendo conglomerados a priori 50
5.13. Tabla resumen de media, desviación estándar e IC al 95 % de los parámetros e hiperparámetros encontrados en el modelo Poisson - CAR incluyendo conglomerados. 50
5.14. Tabla resumen de resultados, porcentajes de detección de conglomerados . . . 54
5.15. Descripción de conglomerados finales 54

Capítulo 1

Introducción

1.1. Consideraciones preliminares

El sector financiero juega un papel importante en la economía peruana. Las instituciones financieras sólidas y solventes funcionan como canales transmisores de recursos financieros entre agentes económicos para de esta manera aprovechar las oportunidades de negocios y consumo. En el sistema bancario empresarial, uno de los riesgos más importantes es el que está asociado al riesgo crediticio el cual es diferenciado según la exposición financiera que tenga cada empresa y parte de este riesgo se debe a sus ciclos económicos en que la empresa pueda desarrollarse. Por ello el riesgo de crédito es el tipo de riesgo más importante al que debe hacer frente una entidad financiera, en particular el nivel de morosidad de una entidad es un indicador muy usado para medir el nivel de riesgo de crédito pues nos resume la proporción de cartera que la entidad posee en calidad de incumplimiento.

La morosidad es el problema principal que sufren ciertas entidades de distinto tamaño (Goodhart y Schoenmaker, 1993). Niveles elevados de morosidad en un cartera se convierten en un problema relevante que llega a comprometer la viabilidad en el largo plazo de la entidad para luego afectar a su propio sistema. Como consecuencia, los altos niveles de morosidad de créditos conllevan a problemas de liquidez, los cuales si no son respaldados con planes de contingencia pasan a convertirse en problemas de solvencia, determinando probablemente la liquidación de la empresa (Freixas et al., 1994) .

En la banca empresarial, comúnmente los niveles de morosidad varían en diferentes regiones, y dentro de las regiones entre sus mismos distritos, uno de los factores que explican estos niveles de riesgo se debe al giro de negocio y formalidad del mismo. Una de las principales consecuencias es que el ciclo de vida de las empresas sea corto. El alcance y el patrón de los niveles de morosidad pueden ser ilustrados en un mapa del riesgo de morosidad para cada región del país (Figura 1.1), donde podemos observar provincias de la zona norte con niveles de morosidad superiores al resto de provincias del Perú.

Un beneficio particular de estudiar los niveles de morosidad a nivel provincial es que permite identificar conglomerados de unidades de área. Esto permitiría detectar niveles de riesgo altos y de esta forma las entidades financieras podrían gestionar estos riesgos. Permitiría, por ejemplo que las entidades financieras asesoren a las empresas de las regiones con mayores niveles de morosidad. Estas asesorías podrían coberturar planes de negocios o asesorías financieras. Por lo tanto, además del beneficio de ubicar regiones con altos niveles de morosidad, este tipo de estudios pueden ayudar también a reducir los costos de provisiones.

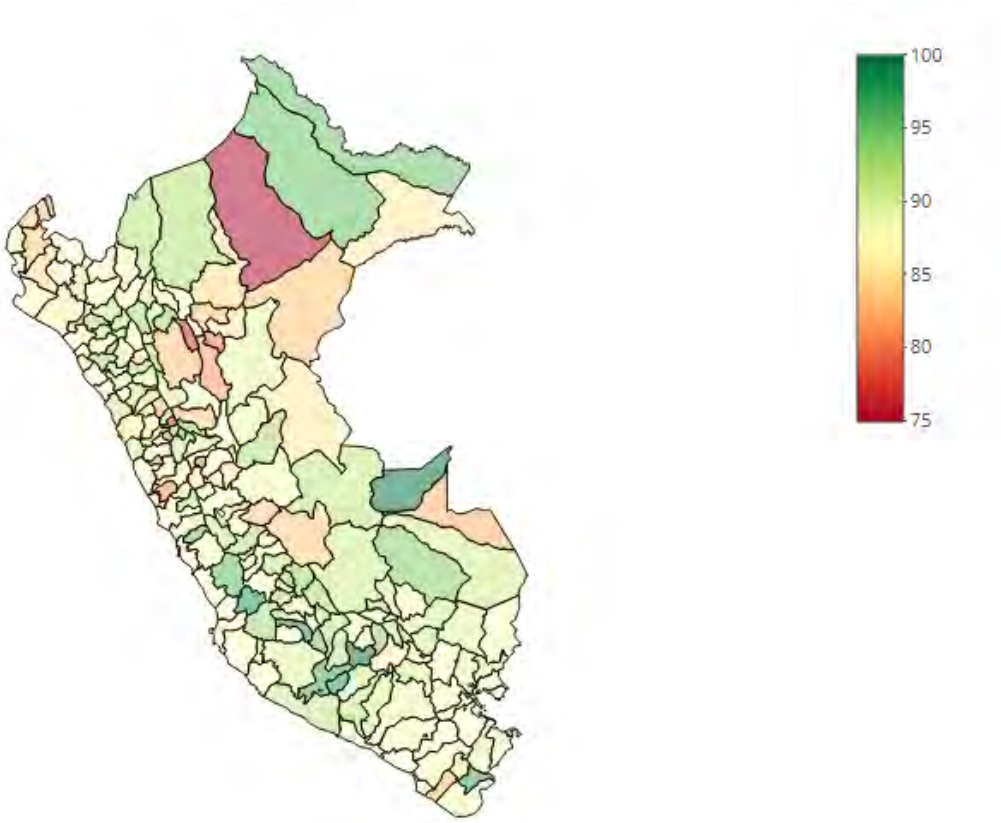


Figura 1.1: Niveles de riesgo, medido por la tasa de morosidad para provincias del Perú en el año 2018, el nivel de riesgo que va de menor a mayor magnitud está representado por los colores verde y rojo , en ese orden respectivamente.

La literatura estadística nos permite identificar patrones espaciales de riesgos, por ejemplo, a través de los modelos jerárquicos (Charras-Garrido et al., 2012), estadísticas SCAN (Kulldorff, 1997), procesos puntuales (Diggle et al., 2005), entre otros. El modelo jerárquico está basado en un modelo Poisson log-linealizado donde el nivel de morosidad es explicado por covariables y un conjunto de efectos aleatorios, los cuales toman en cuenta la autocorrelación espacial. Los más usados son los modelos condicionales autoregresivos. Estos modelos fueron propuestas por Besag et al. (1991) y desarrollados por Kulldorff (1999) y asumen que áreas geográficamente adyacentes están autocorrelacionadas espacialmente.

1.2. Objetivo principal

En esta tesis se propone aplicar una metodología estadística para estimar patrones espaciales de niveles de morosidad en el Perú y detectar aquellas regiones que pertenecen a conglomerados de alto y bajo niveles de riesgo.

1.3. Objetivos específicos

- Estudiar propiedades e implementar la estimación de conglomerados mediante modelos CAR desde la perspectiva de estadística bayesiana.

- Aplicar métodos de inferencia bayesiana considerando el algoritmo INLA.
- Realizar estudios de simulación acerca del modelo de Poisson log-linealizado considerando computación intensiva sobre diferentes escenarios.
- Aplicar el modelo a conjunto de datos reales. En particular, usar variables relevantes en los modelos propuestos que permitan evidenciar el riesgo de impago para empresas del Perú. Detectar provincias del Perú con tasas de incumplimientos (PD) diferenciadas para controlar el riesgo de crédito de las empresas. Identificar zonas y conglomerados con niveles de riesgo de impago para empresas del Perú.

1.4. Organización del trabajo

- En el Capítulo 2, presentamos el concepto de conglomerados vistos según su nivel de disimilaridad. A continuación se describen los modelos condicionales autorregresivos (CAR) como también sus aproximaciones para ajustar modelos bayesianos mediante el método aproximación de Laplace anidada integrada (INLA).
- En el Capítulo 3, se describe la estructura del modelo de conglomerados de provincias del Perú mediante la estadística espacial para datos de áreas, específicamente, se presentan dos propuestas. La primera consiste en usar un modelo Poisson loglinealizado al cual añadiremos un efecto espacial a través del modelo condicional autorregresivo (CAR) para explicar los niveles de riesgos por provincias. Por otro lado, la segunda propuesta considera un primer nivel modelos aglomerativos jerárquicos, donde se eligen n conglomerados candidatos a priori según su nivel de disimilaridad. En un segundo nivel modelaremos la variable de estudio una vez definido el conglomerado al que pertenece la provincia, es decir, calcularemos el nivel de riesgo haciendo uso del modelo de Poisson considerando prioris condicionales autoregresivas (CAR) que tomen en cuenta los conglomerados.
- En el Capítulo 4, se presentan los resultados de un estudio de simulación.
- En el Capítulo 5, se presentan los resultados de la aplicación del modelo propuesto a datos de empresas del Perú en el sistema financiero.
- Finalmente, en el Capítulo 6 discutimos algunas conclusiones obtenidas en este trabajo.

  En el anexo presentamos los programas utilizados en la aplicaciones al conjunto de datos reales.

Capítulo 2

Conceptos preliminares

En este capítulo se revisan conceptos importantes para describir el modelo aplicado a los datos de riesgo de morosidad. Previamente definimos el vector aleatorio $\mathbf{Y} = (Y_1, ..., Y_n)^t$ que representa el número de casos observados en cada región o área.

2.1. Análisis de conglomerados (Clusters)

El análisis de conglomerado, conocido como análisis de clusters, es una técnica estadística multivariada que busca agrupar elementos tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos a los que llamaremos conglomerados (Kauffman y Rouseew, 1990). La diferencia esencial con el análisis de discriminante se debe a que en este último es necesario especificar previamente los grupos por un objetivo ajeno a la medida de las variables. Mientras que el análisis de conglomerado define grupos tan distintos como sea posible en función de los propios datos. Aunque poco o nada se conoce sobre la estructura de las categorías a priori, se tiene con frecuencia algunas nociones sobre características deseables e inaceptables a la hora de establecer un determinado esquema de clasificación.

Existen dos grandes tipos de análisis de conglomerados: jerárquicos y no jerárquicos. En cuanto a los métodos no jerárquicos, también conocidos como partitivos o de optimización, se tiene como objetivo realizar una sola participación de los individuos en K grupos. Ello implica clasificar a priori los grupos que deben ser formados, siendo esta la principal diferencia respecto a los métodos jerárquicos donde la asignación de individuos a los grupos se hace mediante algún proceso que optimice el criterio de selección. Otra diferencia de éstos métodos respecto a los jerárquicos reside en que trabajan con la matriz de datos originales y no precisan su conversión en una matriz de distancias o similitudes (Massart y Kaufman, 1983).

Se conocen como métodos jerárquicos a los que tienen por objetivo agrupar un conglomerado para formar uno nuevo o separar alguno ya existente para dar origen a otros dos, de tal forma que se minimice alguna función de distancia o bien se maximice alguna medida de similitud (Macnaughton Smith et al., 1965). Los métodos jerárquicos se subdividen a su vez en aglomerativos y disociativos. Los aglomerativos comienzan con el análisis con tantos grupos como individuos hay en el estudio. A partir de ahí se van formando grupos de forma ascendente, hasta que al final del proceso todos los casos están en un mismo conglomerado. Los métodos disociativos o divisivos realizan el proceso inverso, es decir, empiezan con un conglomerado que reúne todos los individuos para después formar a partir de sucesivas

divisiones grupos cada vez más pequeños.

2.1.1. Métodos de análisis de conglomerados

Considerando el objetivo del análisis de conglomerado el siguiente paso es encontrar agrupaciones naturales del conjunto de individuos de la muestra, por lo que resulta relevante el entendimiento de agrupación natural, y por lo tanto con respecto a qué criterio se puede decir si dos grupos son más o menos similares.

Por otro lado, dentro de los métodos para medir las distancias entre conglomerados dentro de los conglomerados jerárquicos tenemos el método de linkage simple, linkage centroide y el método de Ward (Szekely y Rizzo, 2005). Se considera n áreas particionadas $A = \{A_1, ..., A_n\}$ y $(C_1, ..., C_n)$ potenciales configuraciones de conglomerados, tal que un conglomerado C_k es definido como una partición de A en k grupos espacialmente contiguos, así $C_k = \{C_k(1), ..., C_k(k)\}$, donde $C_k(j)$ representa el j-ésimo conglomerado de C_k.

Método linkage simple (vecino más cercano)

Este método busca cuáles son las áreas más próximas en cuanto a distancia o similaridad (según el caso puede ser menor distancia o mayor similaridad), posteriormente éstas áreas forman un grupo que no vuelve a separarse durante el proceso. Este proceso se hace iterativo entre todas las áreas, donde el grupo ya formado se trata como si fuera uno solo, es decir, la distancia entre el grupo formado y la siguiente área se toma como la distancia mínima de las áreas del grupo a la nueva área. Por otro lado, en términos de similitud o similaridad, la similaridad entre el grupo formado y la siguiente área se toma como la máxima similitud o similaridad, de una de las áreas del grupo y la nueva área (podemos ver un ejemplo de forma ilustrativa en la figura 2.1). En particular, en esta tesis se busca áreas próximas por similaridad. Por lo tanto, la disimilaridad se define como:

$$d_{ij} = min\left\{\|Y_f - Y_g\| : A_f \in C_k(i), A_g \in C_k(j)\right\},$$

donde $\|.\|$ denota la distancia Euclidiana, Y_f representa un valor observado de la variable respuesta que pertenece al área A_f, la cual a su vez pertenece al grupo i, de forma similar $Y_g$representa un valor observado de la variable respuesta que pertenece al área A_g, la cual a su vez pertenece al grupo j.

La Figura 2.1 muestra una representación de la disimilaridad d_{ij}, donde los puntos rojos pertenecen al grupo i y puntos verdes al grupo j. La disimilaridad en los reales se calcula como la diferencia de las variables respuesta (Y) entre un punto rojo y un punto verde, haciendo ese proceso para todos los puntos. La disimilaridad es el menor valor absoluto de esas diferencias.

Método linkage Centroide

Asume que la distancia entre dos grupos es la distancia existente entre sus centros de gravedad (centroides). El cálculo comienza encontrando el centro de gravedad de cada conglomerado, para agrupar los conglomerados cuya distancia entre centroides sea mínima. Tras la unión de dos conglomerados se hace un cálculo del nuevo centro de gravedad y se realiza

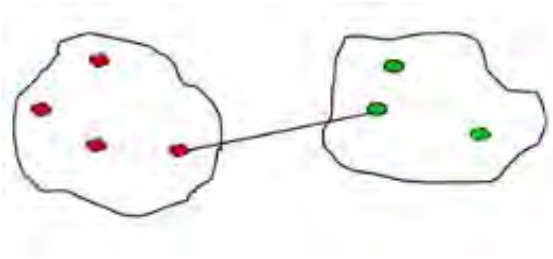


Figura 2.1: linkage simple (vecino más cercano)

el proceso de cálculo de forma similar. En este caso la disimilaridad se define como :

$$d_{ij} = \left\| \bar{C}_k(i) - \bar{C}_k(j) \right\|,$$

donde $\bar{C}_k(i) = (1/n_i) \sum_{f:A_f \in C_k(i)} Y_f$, y n_i es el número de áreas en el conglomerado $C(i)$.

El beneficio de este procedimiento es que se diluye la influencia de casos extremos, por ejemplo de forma ilustrativa en la figura 2.2 se puede interpretar que $\bar{C}$ son los puntos azules.

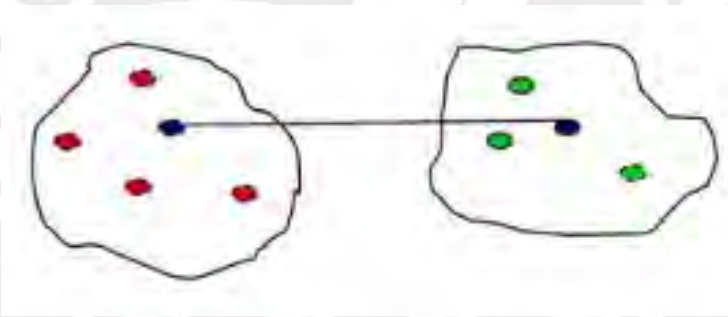


Figura 2.2: linkage centroide

Método de Ward (método de mínima varianza)

El método de Ward consiste en unir los casos anteriores, buscando minimizar la varianza dentro de cada grupo. Para ello en primera instancia se calcula la media de todas las variables en cada conglomerado, posteriormente se calcula la distancia entre cada caso y la media del conglomerado, para luego sumar las distancias entre todos los casos. Como siguiente paso se agrupan los conglomerados que generan la menor suma de las distancias dentro de cada conglomerado. En particular, en esta tesis la disimilaridad se define como:

$$d_{ij} = ESS(C_k(i,j)) - [ESS(C_k(i)) + ESS(C_k(j))],$$

donde $C_k(i,j) = C_k(i) \cup C_k(j)$ y ESS la suma de los errores al cuadrado que tiene la siguiente forma $ESS(C_k(i)) = \sum_{f:A_f \in C_k(i)} \left\| Y_f - \bar{C}_k(i) \right\|^2$, donde Y_f es la variable respuesta que representa al conjunto de individuos pertenecientes a un grupo. Cabe mencionar que el $ESS(i)$ básicamente lo que hace es calcular la diferencia al cuadrado de Y_f (valor de la

variable respuesta en el área f) y $\bar{C}_k(i)$ para el conglomerado i, el cual fue calculado en la sección anterior. $ESS(j)$ hace lo mismo para el conglomerado j, finalmente se tiene que $ESS(C_k(i,j))$ junta los conglomerados i y j, y hace el mismo cálculo de diferencias para los dos conglomerados de forma conjunta.

Este proceso tiene como beneficio crear grupos homogéneos y con tamaños similares (podemos ver un ejemplo de forma ilustrativa en la figura 2.3).

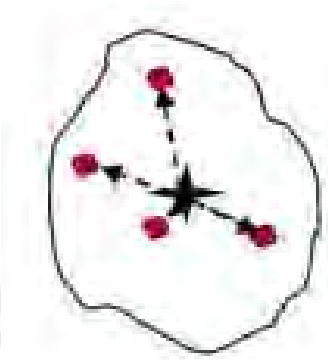


Figura 2.3: Método de Ward (método de mínima varianza)

2.2. Modelo de regresión Poisson para datos de áreas

Los modelos lineales mixtos generalizados pueden modelar variables correlacionadas, las cuales también pueden estarlo de forma espacial. En las siguientes secciones presentaremos una propuesta de modelo para el tratamiento de datos áreas.
Supongamos que tenemos n variables aleatorias $Y_1, Y_2, ..., Y_n$ que poseen distribución Poisson (λ_i), es decir:

$$Y_i \sim Poisson(\lambda_i),$$

cuya función de probabilidad se define como:

$$f_{Y_i}(y_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \quad i = 0, ..., n, \lambda_i \geq 0.$$

Al expresar dicha función de probabilidad en términos de la familia exponencial, se obtiene como resultado:

$$f_{Y_i}(y_i) = exp(y_i ln(\lambda_i) - \lambda_i - ln(y_i!)).$$

la esperanza y la varianza de Y_i se definen como:

$$E(Y_i) = \lambda_i,$$
$$Var(Y_i) = \lambda_i.$$

Los modelos lineales generalizados están conformados por tres componentes principales tal como lo indica Agresti (2015): componente sistemático, componente aleatorio y la función de

enlace $g(\cdot)$, la función de enlace asocia la media con el predictor lineal η_i, tal que:

$$g(\lambda_i) = \eta_i,$$

donde $\eta_i = X_i^T \boldsymbol{\beta}$ es el componente sistemático o predictor lineal, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ es el vector de parámetros desconocidos que serán estimados y $X_i = (x_{i1}, x_{i2}, ..., x_{ik})^T$ representan los valores de las variables explicativas observadas. Por lo general, la función de enlace usada es la función logarítmica.
Una vez definida la función de enlace logarítmica, si incrementamos en una unidad x_j, obtenemos que la media λ_i^* de la variable respuesta Y_i es:

$$\begin{aligned}
\lambda_i^* &= e^{\beta_0 + x_{i1}\beta_1 + ... + (x_{ij}+1)\beta_j + ... + x_{ip}\beta_p} \\
\lambda_i^* &= e^{\beta_j} e^{\beta_0 + x_1\beta_1 + ... + x_j\beta_j + ... + x_p\beta_p} \\
\lambda_i^* &= e^{\beta_j} \lambda_i \\
\lambda_i^*/\lambda_i &= e^{\beta_j},
\end{aligned}$$

donde e^{β_j} es conocido como el riesgo relativo. Si β_j es positivo, la media de la variable respuesta se incrementa en $100(e^{\beta_j} - 1)\,\%$, si β_j es negativo, la media se reduce en $100(1 - e^{\beta_j})\,\%$.

2.3. Estadística espacial para datos de áreas

La estadística espacial es la suma de un conjunto de metodologías apropiadas para el análisis de datos que corresponden a la medición de variables aleatorias en diversos puntos del espacio de una región.
En esta sección se describen las herramientas exploratorias y propuestas de modelos que sean aplicables para datos de unidades de áreas. Uno de los principales beneficios del uso de estadística espacial para datos de áreas es que se puede hacer uso tanto para unidades geográficas irregulares como también para grillas regulares.
Con ayuda de la estadística espacial podemos estudiar si existen patrones espaciales, además de que tan fuerte es este patrón. Intuitivamente, el concepto de patrones espaciales sugiere medidas para unidades de área donde las unidades que están cerca poseen valores similares de aquellas unidades de áreas que se encuentran alejadas de las mismas. Es importante mencionar también que algunas veces el patrón puede ser identificado visualmente, sin embargo, es importante cuantificar su intensidad.

2.3.1. Análisis exploratorio para datos de áreas

El concepto principal que debemos conocer es la matriz de proximidad W que nos da medidas de asociación entre áreas. Esta matriz está conformada por entradas w_{ij} que representa la asociación entre las unidades i y j, donde se asume como diagonal principal $w_{ii} = 0$. La matriz W queda definida de la siguiente manera:

$$W = \begin{bmatrix} w_{1,1} & \cdots & w_{1,n} \\ \vdots & \ddots & \vdots \\ w_{n,1} & \cdots & w_{n,n} \end{bmatrix},$$

dentro del abanico de posibilidades de w_{ij} se encuentran:

- Elecciones binarias, donde $w_{ij} = 1$ si las unidades de áreas i y j comparten una frontera en común, caso contrario 0.
- De forma alterna, w_{ij} podría reflejar también la distancia entre unidades de áreas la cual puede ser traducida a forma binaria.
- Otra forma es considerar $w_{ij} = 1$ para las m áreas vecinas con mayor proximidad, caso contrario 0.
- Por último, los elementos de W pueden ser vistos como pesos, donde el peso es mayor cuando i y j son áreas próximas.

2.3.2. Medidas de asociación espacial

Existen dos estadísticas que son usados para medir la fuerza de asociación espacial entre unidades de áreas, el Índice de Moran y la C de Geary (Ripley, 1981). El índice de Moran tiene la siguiente forma :

$$I = \frac{n \sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{(\sum_{i \neq j} w_{ij}) \sum_i (y_i - \bar{y})^2},$$

donde y_i representa la variable observada en la i-ésima unidad de área, e $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ $\forall i = 1, 2, ..., n$. En general el Índice de Moran I se encuentra en el intervalo $[-1, 1]$, donde $I = -1$ significa dispersión perfecta, es decir, dispersión de los valores de la variable no encontrándose patrones espaciales claros, $I = 0$ significa patrón espacial aleatorio e $I = 1$ significa autocorrelación espacial perfecta, es decir, concentración de los valores similares de una variable en unidades geográficas próximas.

Por otro lado, tenemos que el C de Geary tiene la forma:

$$C = \frac{(n-1) \sum_i \sum_j w_{ij}(y_i - y_j)^2}{2(\sum_{i \neq j} w_{ij}) \sum_i (y_i - \bar{y})^2},$$

donde C nunca es negativo, valores pequeños (entre 0 y 1) indican asociación espacial (Cliff y Ord, 1973).

2.3.3. Campos aleatorios gaussianos de Markov (GMRF)

Para construir un modelo para datos de áreas un concepto importante a tener en cuenta son los campos aleatorios Gaussianos de Markov (en adelante GMRF por sus siglas en inglés). Un vector aleatorio $\mathbf{X} = (X_1, ..., X_n)^T \in N(\mu, Q^{-1})$ es llamado GMRF con respecto a un sistema de vecindad dado si $f(x_i|x_{-i}) = f(x_i|x_{N_i})$ $\forall i$, donde $f(\cdot)$ denota función de densidad

y N_i es la vecindad de i y x_{-i} denota todos los elementos a excepción de x_i. La vecindad de i definida por la matriz de proximidad W.

Los GMRF's por construcción tienen una matriz de precisión Q (inversa de la matriz de covarianza) llena de ceros. Formalmente si $i \neq j$, tenemos que:

$$X_i \perp X_j \mid X_{-i,j} \leftrightarrow Q_{ij} = 0 \leftrightarrow j \notin N_i.$$

De la ecuación anterior se tienen las siguientes propiedades: si X_i y X_j son condicionalmente independientes, i y j no son vecinos, como consecuencia el elemento Q_{ij} de la matriz de precisión es cero. Para más detalles ver Rue y Held (2005).

2.3.4. Modelos autorregresivos condicionales (CAR)

Definido el concepto de GMRF se obtiene el lema de Brook: donde $f(x)$ representa la densidad del vector aleatorio n-dimensional $\mathbf{X}$ y $\Omega = \{\mathbf{X} \in \mathbb{R}^n : f(\mathbf{X}) > 0\}$. Sea el vector fijo $\mathbf{X}' \in \Omega$

$$\begin{aligned}\frac{f(\mathbf{x})}{f(\mathbf{x}')} &= \prod_{i=1}^{n} \frac{f(X_i \mid X_1, ..., X_{i-1}, X'_{i+1}, ..., X'_n)}{f(X'_i \mid X_1, ..., X_{i-1}, X'_{i+1}, ..., X'_n)} \\ &= \prod_{i=1}^{n} \frac{f(X_i \mid X'_1, ..., X'_{i-1}, X_{i+1}, ..., X_n)}{f(X'_i \mid X'_1, ..., X'_{i-1}, X_{i+1}, ..., X_n)}.\end{aligned}$$

De acuerdo con el lema de Brook tenemos que para un campo aleatorio de Markov gaussiano $\mathbf{X}$ y un vector fijo $\mathbf{X}'$, se logra obtener la distribución conjunta de $\mathbf{X}$ a partir de distribuciones condicionales completas, pues el lado derecho de la ecuación se puede observar que $f(\mathbf{X})$ es proporcional al producto de las condicionaes completas (Rue y Held, 2005). Tal como se cita en (Besag, 1974), a partir de las distribuciones condicionales completas se puede obtener la distribución conjunta.

Para el caso de los modelos CAR el enfoque suele ser más intuitivo, es decir, este es análogo al análisis de series de tiempo pues especifica modelos para las distribuciones de probabilidad de cada observación Y_i condicionada a los valores observados de las demás observaciones mediante $f(y_i \mid y_{-i})$ donde y_{-i} es el vector de todas observaciones menos la correspondiente a la unidad de área i, $\forall i$ y $f(\cdot)$ representa una función de densidad de una distribución normal.

En el área de las series de tiempo decimos que las variables aleatorias $Y_1, ..., Y_t$ cumplen la propiedad de Markov cuando $f(y_{t+1} \mid y_t, ..., y_1) = f(y_{t+1} \mid y_t)$, es decir, que el valor en el tiempo $t+1$ solo depende de lo ocurrido en el tiempo inmediatamente anterior t. Una consecuencia de tener variables aleatorias con la propiedad de Markov es de generar un proceso de Markov. Bajo esta premisa si lo proyectamos al campo de los datos espaciales, diremos que el valor Y_i dependerá solamente de lo que ocurra con sus vecinos. Deduciéndose que Y_i depende de Y_j únicamente si la localización j pertenece al conjunto de vecinos i (el cual llamaremos N_i). Cuando esto se cumple diremos que el proceso Y es un campo aleatorio de Markov.

Para un enfoque autorregresivo condicional construiremos modelos a partir de $f(y_i \mid y_j, j \in$

N_i). Suponemos distribuciones condicionales normales, con:

$$E(Y_i \mid Y_{-i}) = \sum_{j=1}^{n} b_{ij}(Y_i), \tag{2.1}$$

$$Var(Y_i \mid Y_{-i}) = \sigma_i^2, i = 1, ..., n, \tag{2.2}$$

donde $\sigma_i^2 > 0$ y $b_{ij} \geq 0$ representan constantes, además tenemos que $b_{ii} = 0$ para todo i, pues refleja una relación de la unidad de área consigo misma.
El teorema de Hammserley-Clifford (Besag, 1974) describe las condiciones para definir una distribución conjunta $f(y_1, ..., y_s)$ a partir de un conjunto de distribuciones condicionales $f(y_i \mid y_j, j \in N_i)$. Si asumimos que las distribuciones condicionales son normales con media (2.1) y varianza (2.2), nos encontramos bajo las condiciones requeridas por el teorema de Hammserley-Clifford llegandose a demostrar que dichas distribuciones condicionadas generan una distribución conjunta proporcional a una distribución normal multivariante con media cero y matriz de precisión con la siguiente forma: $Q_{CAR} = [D^{-1}(I - B)]$, donde $D^{-1} = diag\left[1/\sigma_1^2, ..., 1/\sigma_n^2\right]$. Para que la Q_{CAR} sea simétrica se impone la condición $\sigma_j^2 b_{ij} = \sigma_i^2 b_{ji}$ donde c_{ji} son elementos de la matriz C, y se asume que $c_{ij} = w_{ij}/w_{i+}$ y $\sigma_i^2 = \sigma^2/w_{i+}$, donde w_{i+} representa el número de vecinos de la i-ésima área. De esta forma la matriz de precisión se representa como:

$$Q_{CAR} = \tau(W_1 - W),$$

donde W_1 es una matriz diagonal con elementos w_{i+} y $\tau = 1/\sigma^2$. Cabe resaltar que $f(y_1, ..., y_s)$ es impropia debido a la singularidad de Q_{CAR}, pero puede ser usada como una distribución a priori para los efectos espaciales. Esta a priori es conocida como modelo CAR intrínseco. Para más detalles ver el Apéndice A.

2.4. Inferencia bayesiana

En la estadística inferencial, existen dos categorías de interpretación de las probabilidades: la inferencia bayesiana y la inferencia frecuentista, su diferencia radica en el fundamento natural de probabilidad. En inferencia bayesiana la probabilidad es la forma de mostrar una evidencia dada (Berger, 2006).

El teorema de Bayes muestra la relación entre las probabilidades condicionales. El teorema de Bayes (Bayes y Price, 1763) expresa la probabilidad condicional o probabilidad posteriori, de un evento A dado B es observado en términos de la probabilidad priori de A, probabilidad marginal de B, y la probabilidad condicional de B dado A, es decir, tenemos que :

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{2.3}$$

La base de la inferencia bayesiana es derivada del teorema de bayes (2.3), en donde procedemos a reemplazar B con observaciones $\mathbf{y}$, A con un conjunto de parámetros θ, y la función de densidad de probabilidad (en adelante fdp) a priori expresadas por p (algunas veces por π) , obteniendo el siguiente resultado:

$$p(\boldsymbol{\theta} \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}, \tag{2.4}$$

donde $p(\mathbf{y})$ es la fdp marginal de y, $p(\theta)$ es la fdp priori de los parámetros θ, $p(\mathbf{y} \mid \theta)$ es la fdp de $\mathbf{y}$ dado θ, y $p(\theta \mid \mathbf{y})$ es la fdp condicional de $(\theta \mid \mathbf{y})$, donde $(\theta \mid \mathbf{y})$ es una distribución posteriori.
El denominador viene dado por:

$$p(\mathbf{y}) = \int p(y \mid \theta)p(\theta)d\theta,$$

definido como la fdp marginal de $\mathbf{y}$. La presencia de la fdp marginal de $\mathbf{y}$ normaliza la distribución posteriori de $(\theta \mid \mathbf{y})$ asegurando que sea una distribución propia e integre 1. Reemplazando $p(\mathbf{y})$ con c, la cual representa una constante de proporcionalidad en la ecuación (2.4), pues $p(\mathbf{y})$ no depende de θ, tenemos la siguiente forma dada :

$$p(\boldsymbol{\theta} \mid y) = \frac{p(y \mid \theta)p(\theta)}{c},$$

luego,

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

A manera de resumen podemos decir que los componente de la inferencia bayesiana son:

- $p(\theta)$ representa la función de densidad de probabilidad (en adelante fdp) de las distribuciones prioris del conjunto de parámetros θ , y usa la probabilidad como la medida de la incertidumbre del conjunto de parámetros θ antes de tomar los datos en cuenta.
- $p(\mathbf{y} \mid \theta)$ representa la fdp de las variables son relacionadas con el modelo de probabilidad.
- $p(\theta \mid \mathbf{y})$ representa la fdp de la distribución a posteriori conjunta que expresa la incertidumbre acerca del conjunto de parámetros θ después de tomar en consideración la priori y los datos.

2.4.1. Método de las cadenas de markov de Monte Carlo (MCMC)

Ante la dificultad que la distribución a posteriori no presenta una forma conocida resulta complicado dar con estimaciones exactas sobre el vector de parámetros θ, como solución a estas limitaciones surge el método de simulaciones de Monte Carlo que nos permite estimar éstas variables de interés. Por otro lado, simular bajo el método de Monte Carlo suele tener sus limitantes surgiendo de esta manera un método que utilice las cadenas de Markov.

El concepto de simular cadenas de Markov implica obtener una muestra de valores dependientes solo del valor inmediato anterior cuya distribución estacionaria es la distribución a posteriori. Podemos representar la propiedad de Markov bajo la siguiente expresión:

$$p(\theta^{(t)} \mid \theta^{(1)}, \theta^{(2)}, ..., \theta^{(t-1)}) = p(\theta^{(t)} \mid \theta^{(t-1)}),$$

donde $\theta^{(1)}, ..., \theta^{(t)}$ son variables aleatorias que ocurren en un periodo $t = 1, 2, ..., n$. Es necesario que bajo el método de las cadenas de Markov se deban cumplir ciertas propiedades para lograr la estacionariedad (Robert y Casella, 2004).
Los algoritmos que se basan en MCMC más usados son el muestreador de Gibbs y el algoritmo Metropolis-Hastings; no obstante, para el presente estudio utilizaremos el método INLA dado la naturaleza del modelo en estudio.

2.5. Aproximación de Laplace anidada integrada (INLA)

2.5.1. Introducción

En el campo de la inferencia bayesiana métodos de aproximación como del cadenas de Markov de Monte Carlo han sido de gran utilidad (Gilks et al., 1996) y (Brooks et al., 2011) para simular la distribución conjunta aposteriori de los parámetros del modelo. Sin embargo, estas simulaciones son computacionalmemte muy costosas.
Rue, Martino and Chopin Rue (2009) proponen un método de estimación mediante la inferencia bayesiana que se realiza de forma rápida. Este plantea enfocarse en las distribuciones marginales de los parámetros del modelo. El método se centra en modelos que pueden ser expresados como modelos gaussianos latentes. Este enfoque produce ventajas computacionales como la reducción de tiempos para ajustar el modelo. En particular, Rue (2009) desarrollaron una nueva aproximación para las distribuciones marginales aposteriori para los parámetros del modelo basado en una aproximación de Laplace (MacKay, 2003). La versión más reciente de INLA puede ser encontrada en (Rue et al., 2017).

2.5.2. La aproximación de Laplace

Con la siguiente expresión definimos la aproximación de Laplace:

$$\int_{\alpha}^{\beta} f(x)dx \approx f(x^*)\sqrt{2\pi\sigma^{2*}}(\Phi(\beta) - \Phi(\alpha)),$$

donde x^* es la moda de $f(x)$, $\sigma^{2*} = -\frac{\delta log(f(x))}{\delta x^2}$ es evaluado en $x = x^*$ y $\Phi(\cdot)$ es la función de distribución acumulada de una variable con distribución $N(x^*, \sigma^{2*})$.

2.5.3. Modelos gaussianos latentes

En este apartado definimos el vector aleatorio $Y = (Y_1, Y_2, ..., Y_n)^t$ cuyos valores observados son de la forma $\mathbf{y} = (y_1, y_2, ..., y_n)^T$, también tenemos la distribución de Y_i caracterizada por el parámetro de posición μ_i, el cual representa la media. Considerando lo mencionado anteriormente definimos la estructura aditiva para el predictor lineal, $g(\mu_i) = \eta_i$ tal que:

$$\eta_i = \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{j=1}^{L} f_j(z_{ji}), \tag{2.5}$$

donde $\boldsymbol{\beta} = (\beta_1, ..., \beta_M)^T$ son los coeficientes de regresión de las covariables $\mathbf{x} = (1, x_2, ..., x_M)^T$ y $\mathbf{f} = (f_1(.), ..., f_L(.))^t$ es un conjunto de funciones definidas en términos de un conjunto de covariables $\mathbf{z} = (z_1, ..., z_L)^T$. Además, tenemos que los componentes de $f_j(\cdot)$ pueden asumir diferentes formas suaves, efectos temporales o espaciales, etc. En resumen, el conjunto de

parámetros $\theta = (\beta, f)$, los cuales son campos aleatorios gaussianos de Markov que dependen a su vez de un conjunto de hiperparámetros $\psi = (\psi_1, ..., \psi_k)$.
Entonces el campo gaussiano θ tendrá la siguiente distribución:

$$\theta \sim N(0, Q^{-1}(\psi)),$$

donde el componente $Q(\psi)$ tiene la característica de ser una matriz dispersa, matriz compuesta con gran cantidad de valores cero como elementos, lo que genera una mayor eficiencia computacional.

2.5.4. La aproximación de Laplace anidada integrada (INLA)

Para poder describir los modelos que puede ajustar la aproximación INLA, debemos mencionar un vector **n** observaciones $\mathbf{y} = (y_1, ..., y_n)^t$. En general, **y** es asumido condicionalmente independiente dado el campo gausiano latente θ y los hiperparámetros ψ.
Formalmente el modelo puede ser jerárquicamente escrito como:

$$\begin{aligned} y \mid \theta, \psi &\sim \prod_{i=1}^{n} \pi(y_i \mid \eta_i, \psi), \\ \theta \mid \psi &\sim N(0, Q^{-1}(\psi)), \\ \psi &\sim \pi(\psi), \end{aligned}$$

donde π es la fdp de y_i condicional al campo latente e hiperparámetros, $Q(\psi)$ es la precisión (inversa de la covarianza) matriz de campo gausiano latente y $\pi(\psi)$ representa función en términos de ψ. La dependencia de la estructura de la data es capturada principalmente por la matriz de precisión $Q(\psi)$ a través de una elección inteligente de los términos f_k reemplazados en la ecuación (2.5). Para que INLA funcione eficientemente se necesita que la matriz de precisión Q sea dispersa.
Mucho de los modelos que son comúnmente usados como priori para los términos f_k pertenecen a la clase de los llamados campos aleatorios markovianos gausianos (GMRF). Los GMRF's pueden ser usados para modelar los efectos de una covariable, efectos aleatorios, errores de medición, dependencia temporales, etc (Rue y Held, 2005). Cuando se comienza a evidenciar dependencia espacial existen modelos GMRF para datos de área como los modelos CAR o BYM (Besag et al., 1991). Los GMRF's son modelos gausianos dotados con propiedades de Markov, estos son enlazados con estructura distinta de cero de la matriz de precisión en el sentido que si dos elementos del campo son condicionalmente independientes dados los otros elementos, la entrada correspondiente a la matriz de precisión es igual a cero (Rue y Held, 2005). En la práctica escoger una a priori GMRF para f_k induce dispersión en la matriz de precisión $Q(\psi)$. Como resultado tenemos que la densidad a posteriori de **x** y ψ dado **y** es:

$$\pi(\theta, \psi \mid y) \propto exp(-\frac{1}{2}\theta^T Q(\psi)\theta + \sum_i log(\pi(y_i \mid \eta_i, \psi)) + log\pi(\psi)). \tag{2.6}$$

Como podemos ver la densidad posee una dimensión alta que dificulta la interpretación, cuando el principal interés radica en la marginal a posteriori del campo latente $\pi(\theta_i \mid \mathbf{y})$ o los

hiperparámetros $\pi(\psi_j \mid \mathbf{y})$. El método INLA posee una aproximación con densidades marginales a posteriori, las cuales pueden ser usadas para llegar a aproximaciones estadísticas de interés como la media, varianza o cuantiles posterioris.

A manera de resumen podemos decir que el método INLA puede ser aplicado para modelos gausianos latentes que cumplen los siguientes supuestos:

- Cada dato dependa solo de uno de los elementos de los campos gausianos latentes ψ, el predictor lineal, de modo que la fdp puede ser escrita como la productoria ($\prod(.)$), tal como se muestra a continuación:

$$y \mid \theta, \psi \sim \prod \pi(y_i \mid \eta_i, \psi). \tag{2.7}$$

- El tamaño del vector de los hiperparámetros ψ es pequeño (menos a 15).
- El campo latente θ, puede ser grande pero dotado con algunas propiedades de dependencias condicionales (Markov) para que la matriz de precisión $Q(\psi)$ sea dispersa.
- El predictor lineal depende linealmente de una función de covariables desconocida.
- El interés inferencial radica en las marginales posterioris univariantes $\pi(\theta_i \mid \theta)$ y $\pi(\psi_j \mid \mathbf{y})$ más que en la conjunta a posteriori $\pi(\theta_i, \psi_j \mid \mathbf{y})$.

Capítulo 3

Desarrollo del modelo

En este capítulo hablaremos del modelo propuesto para identificar los conglomerados de provincias de acuerdo al riesgo de morosidad.

3.1. Introducción a los modelos propuestos

El objetivo de esta tesis determinar patrones espaciales de riesgos e identificar conglomerados por niveles de riesgo, para ello se propone dos modelos organizados de la siguiente forma :

- La primera propuesta consiste en considerar un modelo de Poisson log-linealizado, donde el parámetro nivel de morosidad es representado por covariables y/o un conjunto de efectos aleatorios, los cuales toman en cuenta la autocorrelación espacial y son modelados generalmente por una priori condicional autoregresiva (CAR). Estas prioris propuestas por Besag et al. (1991) y desarrollada por Kulldorff (1999) trabajan con campos aleatorios Gaussianos de Markov. Estas prioris asumen que las áreas geográficamente adyacentes están autocorrelacionadas espacialmente. En resumen, usaremos un modelo Poisson considerando prioris condicionales autoregresivas (CAR).

- La segunda propuesta consta de dos pasos. En la primera etapa se realizará la generación de conglomerados usando criterios de agrupación aglomerativa jerárquica, que consiste en juntar aquellas áreas de unidad que son similares mientras separamos aquellas que son diferentes de acuerdo a la variable de estudio. Tomando en cuenta los estudios de Anderson et al. (2014) este proceso es beneficioso considerarlo porque deseamos identificar grupos de áreas o regiones con similar riesgo de morosidad. La segunda etapa consistirá en modelar la variable de estudio una vez definido el conglomerado al que pertenece la unidad de área, es decir, calcularemos el nivel de riesgo haciendo uso del modelo Poisson considerando prioris condicionales autoregresivas (CAR) tomando en cuanta los conglomerados identificados a priori.

3.2. Modelo de Poisson para datos de áreas

En esta primera propuesta de tesis se postula estimar el riesgo de morosidad teniendo en cuentra los patrones espaciales asociados a cada unidad de área, lo cual es posible mediante el uso de prioris condicionales autoregresivas (CAR). La metodología consiste en ajustar un modelo de Poisson bayesiano log-linealizado acompañado de efectos del modelos CAR, es decir, consideraremos la estructura de contiguidad espacial de la región de estudio.

El diseño del modelo consiste en el estudio de una región A la cual es particionada en n áreas disjuntas $A_1, ..., A_n$. También definimos el vector aleatorio $Y = (Y_1, ..., Y_n)^t$ y el vector de parámetros $E = (E_1, ..., E_n)^t$ que representan el número de casos observados y esperados en cada unidad durante el periodo de estudio, respectivamente. Estos últimos son construidos por estandarización externa basada en variables que permitan describir la características de cada unidad de área. El modelo Poisson log-linealizado es comunmente usado para estimar el riesgo asociado. En esta tesis se asume que Y_i es una v.a. que representa el número de empresas morosas en la región i-ésima A_i tal que

$$\begin{aligned} Y_i \mid E_i, \lambda_i &\sim Poisson(E_i\lambda_i), \quad i = 1, ..., n, \\ log(\lambda_i) &= x_i^T\beta + \phi_i, \end{aligned}$$

donde λ_i representa el riesgo en cada área A_i, denominada como subregión i, E_i es el número esperado de las empresas en la región A_i que dentro del modelo vendría a ser el offset, y $x_i^T = (1, x_{i1}, ..., x_{ip})$ es un vector de covariables, con vector de coeficientes de regresión $\beta = (\beta_0, ..., \beta_p)^t$, y ϕ_i como un efecto espacial. Los efectos espaciales $\phi = (\phi_i, ..., \phi_n)^t$ explican la autocorrelación espacial. Estos son modelados a través de una a apriori CAR, específicada como un conjunto de n v.a.s con fdp $f(\phi_i \mid \phi_{-i})$, donde $\phi_{-i} = (\phi_i, ..., \phi_{i-1}, \phi_{i+1}, ..., \phi_n)$. La CAR a priori más simple es el modelo intrínseco propuesto por Besag et al. (1991), y está dado por

$$\phi_i \mid \phi_{-i} \sim N\left(\frac{\sum_{j=1}^n w_{ij}\phi_j}{\sum_{j=1}^n w_{ij}}, \frac{1}{\tau_\phi(\sum_{j=1}^n w_{ij})}\right), \quad i = 1, ..., n, \tag{3.1}$$

donde τ_ϕ es un parámetro de precisión condicional. La media condicional de ϕ_i es la media de los efectos aleatorios en áreas vecinas, mientras la varianza condicional es inversamente proporcional al número de unidades vecinales. Luego la distribución conjunta de los efectos espaciales es proporcional a una distribución gaussiana multivariada con media cero y una matriz de precisión impropia, la cual esta dada por $Q_\phi = \tau_\phi(W1 - W)$, donde $W1$ es una matriz diagonal que contiene el número de vecinos por cada unidad de área y W es una matriz de vecindad cuyos componentes son definidos por

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

Definiendo $\boldsymbol{\theta} = (\phi, \boldsymbol{\beta}^T)^T$, ψ=(τ_ϕ) e $\mathbf{Y} = (Y_1, ..., Y_n)^t$, tenemos que la función de verosimilitud para el modelo Poisson para datos de área puede ser escrita de la siguiente forma:

$$\begin{aligned} L(\theta, \psi \mid \mathbf{y}) &= p(\mathbf{y} \mid \theta, \psi) \\ &= f_{Y_1}(y_1 \mid \theta, \psi) \times f_{Y_2}(y_2 \mid \theta, \psi) \times ... \times f_{Y_n}(y_n \mid \theta, \psi) \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i \mid \theta, \psi) \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i \mid \lambda_i), \end{aligned} \tag{3.2}$$

donde tenemos que $f_{Y_i}(y_i \mid \lambda_i)$ es la fdp de una variable aleatoria con distribución Poisson, con media λ_i dada por

$$\lambda_i = e^{X_i^T \beta + \phi_i},$$

3.2.1. Inferencia bayesiana

La distribución de probabilidad a posteriori para $\boldsymbol{\theta} = (\phi, \beta)^T$ y $\psi = (\tau_\phi)$ se define como $p(\theta, \psi \mid \mathbf{Y})$, bajo la siguiente forma que se muestra a continuación:

$$p(\theta, \psi \mid \mathbf{Y}) = \frac{p(\mathbf{y} \mid \theta, \psi) \times p(\theta \mid \psi) \times p(\psi)}{p(\mathbf{y})}.$$

Dado que $p(\mathbf{y})$ no depende de θ ni ψ, por lo tanto tenemos

$$p(\theta, \psi \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta, \psi) \times p(\theta \mid \psi) \times p(\psi).$$

La ecuación anterior puede ser expresada de la siguiente forma

$$p(\theta, \psi \mid \mathbf{y}) \propto L(\theta, \psi \mid \mathbf{y}) \times p(\theta \mid \psi) \times p(\psi), \tag{3.3}$$

donde $L(\theta, \psi \mid \mathbf{y})$ es la función de verosimilitud, $p(\theta \mid \psi)$ representa la fdp de la distribución condicional de $\theta \mid \psi$ y $p(\psi)$ representa la fdp de la distribución a priori de ψ. Para el presente documento se asume independencia entre ϕ, β, τ_ϕ por lo que se puede tener la distribución a priori

$$p(\theta \mid \psi) \times p(\psi) = p(\phi) \times p(\beta) \times p(\tau_\phi),$$

donde $p(\beta) = p(\beta_1)p(\beta_2)...p(\beta_p) = \prod_{j=1}^{p} p(\beta_j)$. Se asume que los coeficientes β poseen una distribución normal, es decir, $\beta_j \sim N(0, 1000)$ donde $j = 1, ..., p$ y p hace referencia al número de covariables. Por otro lado, tenemos que para el efecto espacial ϕ asumimos una distribución a priori CAR, la cual tiene la forma de una normal multivariada con media cero y matriz de precisión impropia Q_ϕ. Es importante mencionar que la matriz de precisión $Q_\phi = \tau_\phi(W_1 - W)$, está compuesta por W_1 la matriz diagonal cuyos elementos corresponden al número de vecinos de cada distrito, y W es una matriz de vecindades.
Considerando el campos aleatorio gaussiano de Markov $\theta = (\phi, \beta)$ cuya dimensión es $(n+p)$, depende del conjunto de hiperparámetros ψ, entonces tenemos que la familia de campos aleatorios Markovianos Guassianos (GMRF) queda de la siguiente forma

$$\theta \mid \psi \sim N(0, Q_{\psi}^{-1}),$$

donde la matriz de precisión $Q(\psi)$ queda definida bajo la siguiente forma :

$$Q_{\psi} = \begin{bmatrix} Q_{\phi} & 0 \\ 0 & Q_{\beta} \end{bmatrix},$$

y la matriz de precisión de β tiene la siguiente forma:

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

donde $\tau_{\beta_j} = \frac{1}{100}$ es la precisión de cada coeficiente de regresión j-ésimo. Por otro lado, tenemos que Q_{ϕ} es una matriz dispersa. En el caso de los hiperparámetros de $\psi = (\tau_{\phi})$ se asumirá a priori la siguiente distribución:

$$\tau_{\phi} \sim \text{gamma}(0.5, 0.5). \tag{3.4}$$

Considerando la función de verosimilitud definida en (3.2) y las distribuciones aprioris mencionadas anteriormente, la función de distribución de probabilidad a posteriori definida en (3.3) puede ser expresada de la siguiente forma:

$$\begin{aligned} p(\theta, \psi \mid \mathbf{y}) &\propto \prod_{i=1}^{n} f_{Y_i}(y_i \mid \lambda_i) \times p(\phi \mid \tau) \times p(\beta) \times p(\tau) \\ &\propto \prod_{i=1}^{n} f_{Y_i}(y_i \mid \lambda_i) \times p(\theta \mid \psi) \times p(\psi), \end{aligned}$$

De esta forma obtenemos la siguiente expresión:

$$p(\theta, \psi \mid \mathbf{y}) \propto \prod_{i=1}^{n} f_{Y_i}(y_i \mid \lambda_i) \times \frac{|Q(\psi)|^{1/2}}{2\pi^{n/2}} e^{-\frac{1}{2}\theta^t Q(\psi)\theta} \times p(\psi),$$

donde definimos $|Q(\psi)|$ como el determinante de $Q(\psi)$.

3.2.2. Estimación del modelo

La metodología INLA trabaja con distribuciones marginales de los parámetros, a diferencia de estimaciones como algoritmos ya conocidos como el de MCMC que hacen uso de las distribuciones condicionales completas de los parámetros. El campo aleatorio Markoviano Gaussiano está definido por :

$$\begin{cases} \beta_i & i=1,\ldots,p \\ \phi_i & i=1,\ldots,n, \end{cases}$$

donde p hace referencia al número de coeficientes de regresión y n es el número de unidades de área. La metodología INLA, propone calcular las distribuciones marginales de los parámetros

$$p(\theta_i \mid \psi, \mathbf{y}) = \begin{cases} p(\beta_i \mid \psi, \mathbf{y}) & i=1,\ldots,p \\ p(\phi_i \mid \psi, \mathbf{y}) & i=1,\ldots,n, \end{cases}$$

donde $p(\beta_i \mid \psi, \mathbf{y})$ es la fdp de la distribución de probabilidad condicional completa de β_i y $p(\phi_i \mid \psi, \mathbf{y})$ es la fdp de la distribución de probabilidad condicional completa de ϕ_i. Por otro lado, para el cálculo de $p(\psi \mid \mathbf{y}) = p(\tau \mid \mathbf{y})$ usa la aproximación de Laplace

$$p(\psi \mid \mathbf{y}) \approx \frac{p(\mathbf{y},\psi|\theta)p(\theta|\psi)p(\psi)}{p_G(\theta|\psi,\mathbf{y})}\Big|_{\theta=\theta^*(\psi)} = \tilde{p}(\psi \mid \mathbf{y}),$$

donde cabe mencionar que p_G hace referencia a la aproximación Gaussiana, además $\theta^*(\psi)$ representa la moda de la condicional completa de θ, también tenemos que $p(\mathbf{y} \mid \psi, \theta)$ es la verosimilitud definida en (3.2), y $p(\theta \mid \psi)$ hace referencia a la distribución conjunta de los campos aleatorias markoviano gasussiano (GMRF) $\theta \sim N(0, Q_\psi^{-1})$, donde como mencionamos en pasos previos Q_ψ es la matriz de precisión formada por Q_β y Q_ϕ.
A partir de $\tilde{p}(\psi \mid \mathbf{y})$ y $\tilde{p}(\theta_i \mid \psi, \mathbf{y})$, en las distribuciones marginales a posteriori de los parámetros, son aproximados por:

$$\tilde{p}(\theta_i \mid \mathbf{Y}) = \int \tilde{p}(\theta_i \mid \psi, \mathbf{Y})\tilde{p}(\psi, \mathbf{Y})d\psi,$$

Luego se realizan cálculos mediante métodos numéricos para obtener las distribuciones marginales de los parámetros. Dada esas condiciones se hará uso de las sumas finitas ponderadas

$$\tilde{p}(\theta_i \mid \mathbf{Y}) \approx \sum_j \tilde{p}(\theta_i \mid \psi^{(j)}, \mathbf{Y})\tilde{p}(\psi^{(j)} \mid \mathbf{Y})\Delta_i,$$

donde tenemos que para ciertos puntos $\psi^{(j)}$ son correspondidos por sus ponderaciones Δ_i.
Finalmente, para el cálculo de las distribuciones marginales de los hiperparámetros tenemos la siguiente ecuación que viene dada de la siguiente forma

$$\tilde{p}(\psi_i \mid \mathbf{Y}) = \int \tilde{p}(\psi \mid \mathbf{Y})d\psi_{-i}.$$

Un detalle importante a considerar es que cuando el número de parámetros es grande el método INLA permite estimar el modelo con mayor eficiencia computacional.

3.3. Modelo Poisson por conglomerados para datos de áreas

En esta tesis tenemos como segunda propuesta aplicar y extender el modelo Anderson et al. (2014) para estimar patrones espaciales de niveles de morosidad y detectar el alcance de conglomerados de alto y bajo niveles de riesgo. La metodología consiste en la fusión de técnicas de agrupamiento aglomerativo jerárquico con modelos CAR, considerando un proceso de dos etapas. La primera etapa consiste en un algoritmo de conglomerado aglomerativo jerárquico que se extiende respetando la estructura de contiguidad espacial de la región de estudio. La

segunda etapa es ajustar un modelo Poisson Bayesiano log-linealizado para cada estructura de conglomerado candidato. La estructura del conglomerado final será seleccionada a través de un criterio de comparación de modelos, es decir, criterio de información de la desviación (DIC), información de Watanabe (WAIC), entre otros criterios de selección de modelos. A manera de resumen en forma gráfica tenemos lo siguiente:

Figura 3.1: Organización de la propuesta conformada por 2 etapas

A continuación detallaremos cada una de las etapas:

Etapa 1 : Generación de conglomerado usando agrupación aglomerativa jerárquica

El proceso de formar conglomerados consiste en juntar aquellas provincias que son similares mientras separamos aquellas que son muy diferentes considerando la magnitud de la variable de estudio. Este proceso es beneficioso porque deseamos identificar grupos de provincias con similar riesgo de morosidad. Se usan los datos de estudio para identificar un conjunto de m potenciales conglomerados. Para ello primero se asumen inicialmente n conglomerados $(C_1, ..., C_n)$. Dada la naturaleza de los datos, para el conglomerado $C_k = \{C_k(1), ..., C_k(k)\}$ se tienen n áreas particionadas $A = \{A_1, ..., A_n\}$ dentro de k grupos espacialmente contiguos, tal que y $C_k(j)$ es el j ésimo conglomerado de C_k.
Es relevante considerar que la data agrupada fue utilizada usando el ratio de incidencia log-estandarizado (SIR), $\Psi_j = ln(Y^{(j)}/E^{(j)})$, donde $E^{(j)}$ es el número total de empresas morosas en la provincia j. Para el proceso de formar conglomerados a los datos de estudio usaremos un algoritmo de agrupación aglomerativa jerárquica modificada (Hastie et al., 2001), el cual inicialmente considera cada punto como si fuera su propio conglomerado, y luego junta los conglomerados con disimilaridad mínima de cada etapa para formar un conglomerado más grande. Este proceso es repetido hasta que quede un conglomerado que contenga todos los puntos. Entonces con k conglomerados de disimilaridad (d_{ij}) entre conglomerados i $(C_k(i))$

y j ($C_k(j)$) puede ser medido por un número de métricas que consideraremos en esta tesis. Entre las métricas que usaremos tenemos:

- La primera métrica consiste en el enlace único que mide la disimilaridad como $d_{ij} = min\{\|\Psi_f - \Psi_g\| : A_f \in C_k(i), A_g \in C_k(j)\}$, donde $\|.\|$ representa la distancia Euclidiana entre dos observaciones $\Psi_{(.)}$ en las provincias (f y g), y $C_k(i)$ es un conglomerado de orden k al que pertenece cada provincia.

- La segunda métrica se refiere al enlace centroide que mide la disimilaridad como $d_{ij} = \left\|\bar{C}_k(f) - \bar{C}_k(g)\right\|$, donde $\bar{C}_k(g)$ representa la media respecto a $\Psi_{(.)}$ de acuerdo a los conglomerados en la provincia g, es decir, esta media está en función de todas las n provincias que pueden componer el conglomerado.

- El enlace Ward mide la disimilaridad como el incremento en la suma de los errores al cuadrado (ESS) cuando se junta dos conglomerados pequeños en un conglomerado más grande, es decir, $d_{ij} = ESS(C_k(i,j)) - [ESS(C_k(i)) + ESS(C_k(j))]$ donde donde representa la unión de conglomerados definidos en un paso anterior como conglomerados independientes y $ESS(C_k(i))$representa la suma de los errores al cuadrado de una área de estudio (Ψ_f) con respecto a la media del conglomerado al área de estudio ($\bar{C}_k(i)$).

En este último paso el par (i,j) queda representados por la menor disimilaridad, es decir, se identifica los conglomerados que tienen disimilaridad mínima según el método de enlace que haya sido seleccionado previamente. Por otro lado, cabe mencionar que ante la casuística de que la disimilaridad sea igual, el par (i,j) se selecciona de forma aleatoria uno de estos conglomerados.

Etapa 2 : Estimación del conglomerado usando técnicas de comparación de modelos

La data de estudio viene denotada por las variables (Y,E) (número de empresas observadas y esperadas respectivamente que tienen la siguiente forma $Y = (Y_1,...,Y_n)^t$ y $E = (E_1,...,E_n)^t$) y el mejor conglomerado de un conjunto de m candidatos $C_1,...,C_m$ para los datos de estudios es elegido en la primera etapa, el cual es estimado usando comparación de modelos. Específicamente, el modelo bayesiano poisson loglinealizado que se describe a continuación, el cual se ajusta a los datos basados en cada conglomerado C_k, y el conglomerado para los datos se estima eligiendo el modelo que minimiza algún criterio de selección de modelos como el DIC (criterio de información de la devianza), WAIC (índice de Watanabe - Akaike), entre otros.

Definimos λ_i que representa el riesgo en cada área A_i, denominada como subregión i, que se encuentra en un conglomerado dado C_k (donde k hace referencia al número de conglomerados), la propuesta del modelo está dada por:

$$
\begin{aligned}
Y_i \mid E_i, \lambda_i &\sim Poisson(E_i\lambda_i), \quad i = 1, ..., n, \\
ln(\lambda_i) &= X_i^T\beta + \phi_i + \sum_{j=1}^{k} I[A_i \in C_k(j)]\alpha_j, \\
\beta_j &\sim N(0, 1000), \;\; j = 1, ..., p, \\
\alpha_j &\sim N(0, 10), \;\; j = 1, ..., k, \\
\phi_i \mid \phi_{-i} &\sim N\left(\frac{\sum_{j=1}^{n} w_{ij}\phi_j}{\sum_{j=1}^{n} w_{ij}}, \frac{1}{\tau_\phi(\sum_{j=1}^{n} w_{ij})}\right), \\
\tau &\sim \text{gamma}(0.5, 0.5),
\end{aligned}
\tag{3.5}
$$

donde $x_i = (1, x_{i1}, ..., x_{ip})^T$ es un vector de covariables, con vector de coeficientes de regresión $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^T$, y ϕ_i es un efecto espacial. Los efectos espaciales $\boldsymbol{\phi} = (\phi_i, ..., \phi_n)^T$ explican la autocorrelación espacial, además $I[.]$ denota una función en el cual $I[A_i \in C_k(j)]$ es igual 1 si el área A_i se encuentra en el conglomerado j, y cero en otros casos. Por lo tanto, el modelo de conglomerado constante es una covariable categórica con k niveles, donde cada conglomerado representa un nivel diferente. Notamos que cuando el área A_i es un conglomerado, este modelo esencialmente incluye una variable para esa área. Consideramos modelar los parámetros del conglomerado $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)^t$ como aleatorios en vez de que sean efectos fijos.
Una vez definidas las distribuciones condicionales en la ecuación (3.5) se asume para ϕ una a priori CAR, la cual tiene la forma de una normal multivariada con media cero y matriz de precisión Q_ϕ. Es importante mencionar que la matriz de precisión $Q_\phi = \tau_\phi(W_1 - W)$, está compuesta por W_1 la matriz diagonal cuyos elementos corresponden al número de vecinos de cada distrito, y W es una matriz de vecindades cuyos componentes son definidos por

$$
w_{ij} = \begin{cases} 1, & si\ i\ es\ vecino\ de\ j \\ 0, & otros\ casos. \end{cases}
$$

Definiendo $\boldsymbol{\theta} = (\phi, \beta^T, \alpha)^T$, ψ=(τ_ϕ) e $\mathbf{y} = (y_1, ..., y_n)^t$, tenemos que la función de verosimilitud para el modelo Poisson por conglomerados para datos de áreas puede ser escrita de la siguiente forma:

$$
\begin{aligned}
L(\theta, \psi \mid \mathbf{y}) &= p(\mathbf{y} \mid \theta, \psi) \\
&= \prod_{i=1}^{n} f_{Y_i}(y_i \mid \theta, \psi)
\end{aligned}
\tag{3.6}
$$

donde tenemos que $f_{Y_i}(y_i|\theta, \psi)$ es la fdp de una variable aleatoria con distribución de Poisson con medias $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)^T$ definidas por

$$
\lambda_i = e^{X_i^T\beta + \phi_i + \sum_{j=1}^{k} I[A_i \in C_k(j)]\alpha_j}.
$$

3.3.1. Inferencia bayesiana

La distribución de probabilidad a posteriori para $\boldsymbol{\theta} = (\phi, \beta^T, \alpha)^T$ y $\psi = (\tau_\phi)$ se define como $p(\theta, \psi \mid \mathbf{Y})$, bajo la siguiente forma que se muestra a continuación:

$$p(\theta, \psi \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \theta, \psi) \times p(\theta \mid \psi) \times p(\psi)}{p(\mathbf{y})}.$$

Dado que $p(\mathbf{Y})$ no depende de θ y ψ por lo tanto tenemos

$$p(\theta, \psi \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta, \psi) \times p(\theta \mid \psi) \times p(\psi).$$

La ecuación anterior puede ser expresada de la siguiente forma

$$p(\theta, \psi \mid \mathbf{y}) \propto L(\theta, \psi \mid \mathbf{y}) \times p(\theta \mid \psi) \times p(\psi), \tag{3.7}$$

donde $L(\theta, \psi \mid \mathbf{y})$ es la función de verosimilitud, $p(\theta \mid \psi)$ representa la fdp de la distribución condicional de $(\theta \mid \psi)$ y $p(\psi)$ representa la fdp de la distribución a priori de ψ. Para el presente documento se asume independencia entre ϕ,β,α y τ_ϕ por lo que se puede tener la distribución a priori

$$p(\theta \mid \psi) \times p(\psi) = p(\phi) \times p(\beta) \times p(\alpha) \times p(\tau_\phi), \tag{3.8}$$

donde $p(\beta) = p(\beta_1)p(\beta_2)...p(\beta_p) = \prod_{j=1}^{p} p(\beta_j)$. Se asume que los coeficientes β poseen una distribución normal, es decir, $\beta_j \sim N(0, 1000)$ donde $j = 1, ..., p$ y p hace referencia al número de covariables; para el vector de coeficientes α se asume que posee una distribución normal, es decir, $\alpha_j \sim N(0, 10)$ donde $j = 1, ..., k$ y k hace referencia al número de conglomerados que se decidió utilizar.
Por otro lado, tenemos que para el efecto espacial ϕ asumimos una distribución a priori CAR, la cual tiene la forma de una normal multivariada con media cero y matriz de precisión Q_ϕ. Considerando el campo aleatorio gaussiano de Markov (GMRF) $\theta = (\phi, \alpha, \beta)$, cuya dimensión es de $(n+k+p)$, depende del conjunto de hiperparámetros ψ, la familia de campos aleatorios gaussianos de Markov (GMRF) queda de la siguiente forma

$$\theta \mid \psi \sim N(0, Q_\psi^{-1}),$$

donde la matriz de precisión $Q(\psi)$ queda definida bajo la siguiente forma :

$$Q_\psi = \begin{bmatrix} Q_\phi & 0 & 0 \\ 0 & Q_\alpha & 0 \\ 0 & 0 & Q_\beta \end{bmatrix},$$

donde las matrices de precisiones de β y α (Q_β y Q_α respectivamente) de la matriz mencionada anteriormente tienen la siguiente forma:

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

$$Q_\alpha = \begin{bmatrix} \tau_{\alpha_1} & 0 & \cdots & 0 & 0 \\ 0 & \tau_{\alpha_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & \tau_{\alpha_{k-1}} & 0 \\ 0 & 0 & \cdots & 0 & \tau_{\alpha_k} \end{bmatrix},$$

siendo $\tau_\beta = \frac{1}{100}$ y $\tau_\alpha = \frac{1}{100}$ la precisión de cada coeficiente de regresión y efecto de los conglomerados respectivamente.

Por otro lado, tenemos que Q_ϕ como mencionamos anteriormente es una matriz dispersa. En el caso de los hiperparámetros de $\psi = (\tau_\phi)$ se asumirá a priori la siguiente distribución:

$$\tau_\phi \sim \text{gamma}(0.5, 0.5).$$

Considerando la función de verosimilitud definida (3.2) y las distribuciones aprioris mencionadas anteriormente, la función de distribución de probabilidad a posteriori definida en (3.7) puede ser expresada de la siguiente forma:

$$p(\theta, \psi \mid \mathbf{Y}) \propto \prod_{i=1}^{n} f_{Y_i}(y_i \mid \lambda_i) \times p(\phi \mid \tau) \times p(\beta) \times p(\alpha) \times p(\tau_\phi)$$

De esta forma obtenemos la siguiente expresión:

$$p(\theta, \psi \mid \mathbf{Y}) \propto \prod_{i=1}^{n} f_{Y_i}(y_i \mid \lambda_i) \times \frac{|Q(\psi)|^{1/2}}{2\pi^{n/2}} e^{-\frac{1}{2}\theta^t Q(\psi)\theta} \times p(\psi),$$

donde definimos $|Q(\psi)|$ como el determinante de $Q(\psi)$.

3.3.2. Estimación del modelo

El campo aleatorio Markoviano Gaussiano es definido por:

$$\begin{cases} \beta_i, & i=1,\ldots,p \\ \alpha_j, & j=1,\ldots,k \\ \phi_i, & i=1,\ldots,n, \end{cases}$$

donde p hace referencia al número de coeficientes de regresión del parámetro λ_i y n es el número de unidades de área.

Las distribuciones marginales de los parámetros son:

$$p(\theta_i \mid \psi, \mathbf{Y}) = \begin{cases} p(\beta_i \mid \psi, \mathbf{Y}), & i=1,...,p \\ p(\alpha_i \mid \psi, \mathbf{Y}), & i=1,...,k \\ p(\phi_i \mid \psi, \mathbf{Y}), & i=1,...,n, \end{cases}$$

donde $p(\beta_i \mid \psi, \mathbf{Y})$ es la función de distribución de probabilidad condicional completa de β_i, $p(\alpha_i \mid \psi, \mathbf{Y})$ es la función de distribución de probabilidad condicional completa de α_i y $p(\phi_i \mid \psi, \mathbf{Y})$ es la función de distribución de probabilidad condicional completa de ϕ_i.
Por otro lado, para el cálculo de $p(\psi_i \mid \mathbf{Y}) = p(\tau \mid \mathbf{Y})$ se usa la aproximación de Laplace

$$p(\psi \mid \mathbf{Y}) \approx \left. \frac{p(\mathbf{Y}|\theta)p(\theta|\psi)p(\psi)}{\tilde{p}_G(\theta|\psi,\mathbf{Y})} \right|_{\theta = \theta^*(\psi)} = \tilde{p}(\psi, \mathbf{Y}),$$

cabe mencionar que p_G hace referencia a la aproximación Gaussiana, además $\theta^*(\psi)$ representa la moda de la condicional completa de θ, también tenemos que $p(\mathbf{Y} \mid \theta)$ es la verosimilitud definida 3.2, y $p(\theta \mid \psi)$ hace referencia a la distribución conjunta de los campos aleatorias markoviano gasussiano (GMRF) $\theta \sim N(0, Q_\psi^{-1})$, donde como mencionamos en pasos previos Q_ψ es la matriz de precisión formada por Q_β,Q_α y Q_ϕ.

A partir de $\tilde{p}(\psi \mid \mathbf{Y})$ y $\tilde{p}(\theta_i \mid \psi, \mathbf{Y})$, se usa en la ecuación encontrada para obtener las distibuciones marginales a posteriori de los parámetros, obteniendo :

$$\tilde{p}(\theta_i \mid \mathbf{Y}) = \int \tilde{p}(\theta_i \mid \psi, \mathbf{Y}) \tilde{p}(\psi, \mathbf{Y}) d\psi.$$

Luego se realizan cálculos mediante métodos numéricos para calcular las distribuciones marginales de los parámetros. Dadas esas condiciones se hará uso de las sumas finitas ponderadas

$$\tilde{p}(\theta_i \mid \mathbf{Y}) \approx \sum_j \tilde{p}(\theta_i \mid \psi^{(j)}, \mathbf{Y}) \tilde{p}(\psi^{(j)} \mid \mathbf{Y}) \Delta_i,$$

donde tenemos que para ciertos puntos $\psi^{(j)}$ son correspondidos por sus ponderaciones Δ_i.
Finalmente, para el cálculo de las distribuciones marginales de los hiperparámetros tenemos la siguiente ecuación que viene dada de la siguiente forma

$$\tilde{p}(\psi_i \mid \mathbf{Y}) = \int \tilde{p}(\psi \mid \mathbf{Y}) d\psi_{-i}.$$

Un detalle importante a considerar es que si el número de parámetros es grande al hacer uso del método INLA permite estimar el modelo con mayor eficiencia computacional.

Capítulo 4

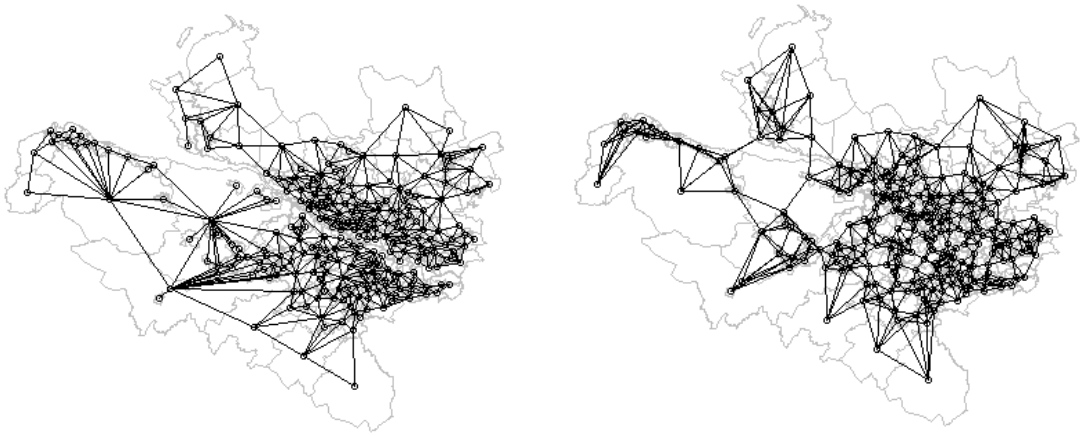
Estudio de Simulación

En este capítulo se generarán datos aleatorios que sigan cierta distribución fijando los parámetros que se vayan a utilizar. Posteriormente se seleccionará el número óptimo de conglomerados utilizando criterios de informaciíon, partiendo de la elección del número óptimo de conglomerados se evalúan los resultados de los parámetros simulados bajo tres escenarios. Como último paso, se mostrará los valores estimados comparados a los datos simulados.

4.1. Generación de datos

Para la simulación de los datos, se consideran $n = 271$ regiones que hacen referencia a una región de Escocia. Se asume que Y_i representa la variable aleatoria de interés en la i-ésima región de estudio. Para la simulación de las covariables, tenemos que la matriz de covariables viene dada por $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$ donde $x_1 = (1, ..., 1)^t$ y la variable x_2 es simulada de una variable aleatoria con distribución normal $N(0, 1)$. El grafo de las 271 regiones (ver figura 4.1(a)) representa las regiones vecinas, las cuales comparten aristas y vértices, donde los vértices representan cada región y las aristas unen a las regiónes vecinas. Cabe mencionar que la matriz de vecindad puede variar de acuerdo al criterio del número de regiones vecinas. Por ejemplo la figura 4.1(b) muestra el grafo que representa diferentes vecindades.
Dado el grafo, a partir de este podemos construir la matriz de vecindad **W** de dimensión igual a 271 (número que representa el total de regiones). En la Figura 4.2 podemos observar la matriz **W**, donde los cuadrados blancos representan que las regiones i y j son vecinas, además de contar con la diagonal con cuadrados blancos. En este estudio el efecto espacial aleatorio por región se define como $\boldsymbol{\phi} = (\phi_1, ..., \phi_{271})$ donde ϕ_i es el efecto espacial de la región i-ésima. De esta forma se asume para ϕ una priori CAR, la cual tiene la forma de una normal multivariada con media cero y matriz de precisión Q_ϕ, donde tenemos que $Q_\phi^{-1} = \tau_\phi(W_1 - W)$, está compuesta por W_1 matriz diagonal. Además para el estudio de simulación se asigna al parámetro τ_ϕ un valor igual a 0.01.

En la siguiente gráfica (ver figura 4.3) se puede observar la simulación de los efectos aleatorios espaciales para cada región de estudio (regiones de Escocia).
En esta parte estableceremos parámetros que se fijaron previamente (se relizaron 3 escenarios) donde se colocaron valores para los coeficientes en cada uno de los escenarios $\beta = (\beta_0, \beta_1)^t$=$(0{,}1, 0{,}08)^t$ para el escenario 1, $\beta = (\beta_0, \beta_1)^t$=$(-0{,}15, 0{,}4)^t$ para el escenario 2 y $\beta = (\beta_0, \beta_1)^t$=$(-0{,}09, 0{,}2)^t$ para el escenario 3. En referencia a los conglomerados se simula el parámetro $\alpha \sim N(0, 1)$ para cada escenario.



(a) Considerando el vecino más próximo (b) Considerando los 5 vecinos más cercanos

Figura 4.1: Grafo de las principales regiónes de Escocia

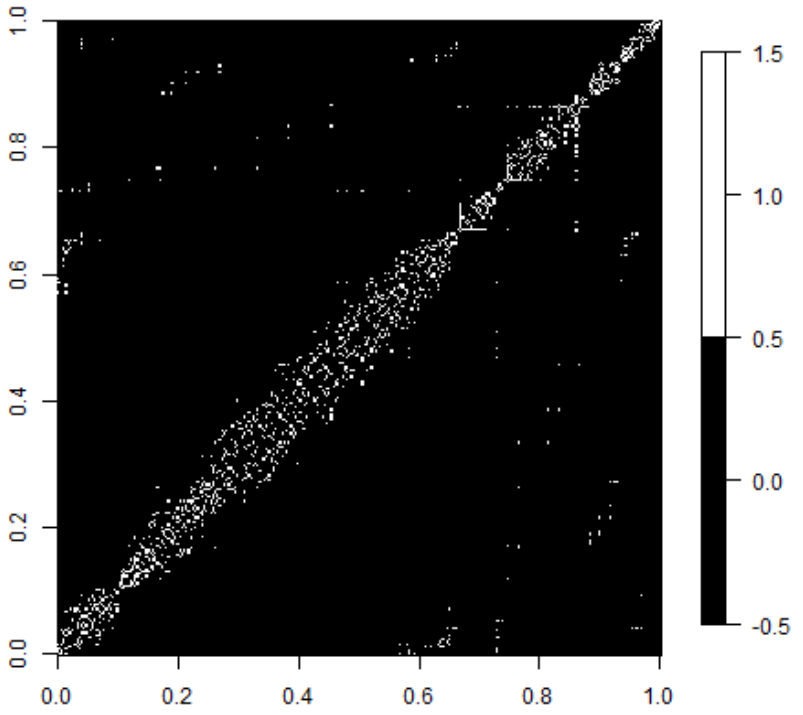


Figura 4.2: Matriz de vecindad de las regiónes de Escocia, considerando el vecino más próximo

Es importante mencionar que se ajustó el modelo por conglomerados para datos de áreas. El algoritmo definió el mejor modelo a partir de $m = 100$ posibles conglomerados candidatos mediante criterios de selección, específicamente el Criterio de Información de la Devianza (DIC), Criterio de Información de Watanabe Akaike (WAIC) y logaritmo de verosimilitud

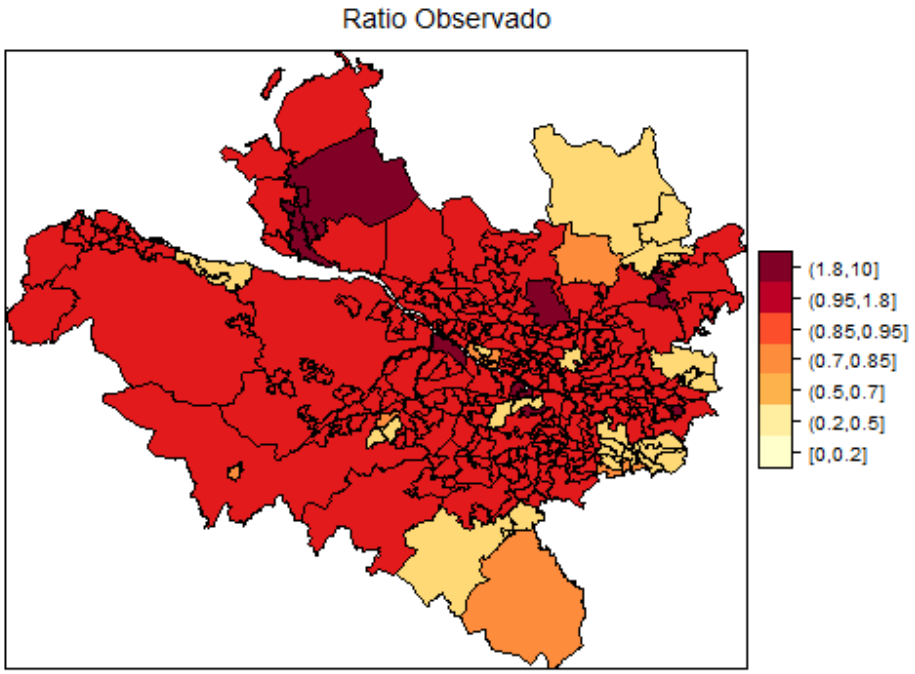


Figura 4.3: Simulación de la variable de estudio considerando los efectos espaciales en cada región de Escocia, donde se muestra niveles de riesgo relacionado a la variable de estudio. La lectura va mayor a menor riesgo (número de casos relacionados a la variable Y), es decir, está representado por matices de los colores rojo y amarillo, en ese orden respectivamente.

pseudo-marginal (LPML), (ver figura 4.4). A partir de estos resultados el número óptimo se logra con $k = 31$ conglomerados. Encontrado el número óptimo de conglomerados, en el siguiente cuadro (4.1) de comparación evaluamos en los criterio mencionados para cada escenario luego de haber escogido previamente el número de conglomerados óptimo. Adicional a ello mostraremos también el tiempo de proceso para el algoritmo.

	Tabla de resultados		
	DIC	WAIC	Tiempo (min.)
Escenario 1	2061.196	2034.232	25.32
Escenario 2	2025.188	2042.661	20.22
Escenario 3	2077.938	2071.687	23.97

Cuadro 4.1: Tabla resumen de resultados, DIC y tiempo de procesamiento en cada escenario

De los resultados mostrados, podemos evidenciar que los tiempos de ejecución en cada escenario se consideran bajos, comprobando de esta forma la eficiencia del algoritmo INLA para modelos gausianos latentes con varios parámetros a estimar.

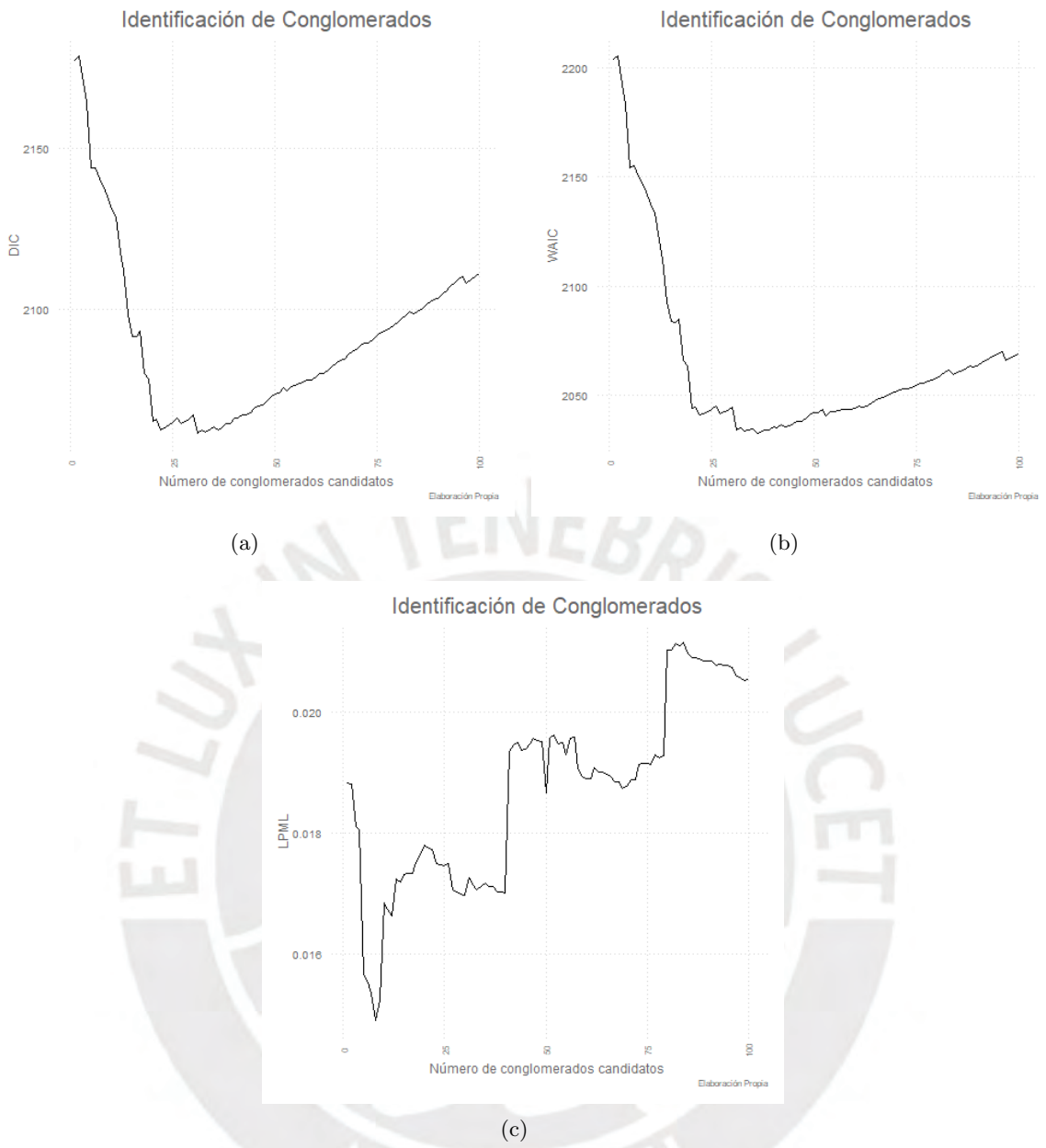


Figura 4.4: Selección del número óptimo de conglomerados mediante menor (a) DIC, (b) WAIC y (c) LPML.

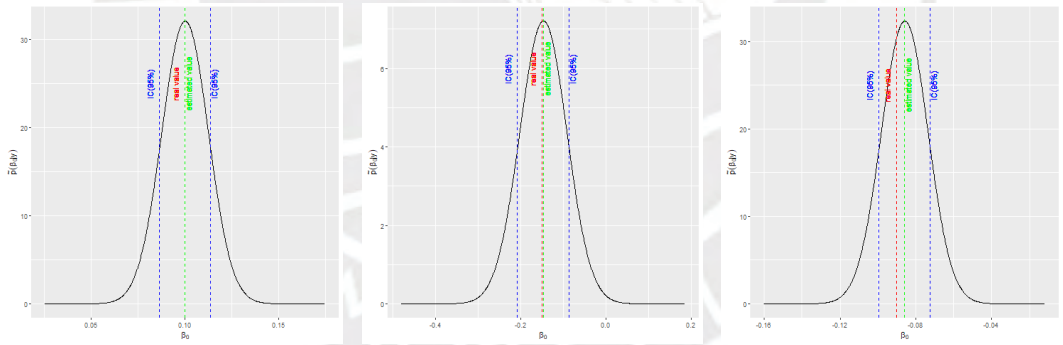
En el cuadro 4.2 se muestra las estimaciones a posteriori de los parámetros, se encontraron resultados deseados pues se logró recuperar los parámetros e hiperparámetros predeterminados, es decir, se obtuvieron estimaciones aproximadas a los valores reales. Además se observa que los tiempos de ejecuciones son reducidos.

De los resultados obtenidos observamos que los intervalos de credibilidad (en adelante IC) contienen al valor real del parámetro. A continuación mostramos en las gráficas 4.5, 4.6, 4.7 y 4.8 las distribuciones marginales a posteriori de los parámetros e hiperparámetros, donde observaremos la media a posteriori (línea verde), el valor real del parámetro (línea roja) y

	Tabla de resultados				
	Parámetro	Real	Media	Desv. est.	95 % IC
Escenario 1	β_0	0.100	0.100	0.012	(0.076 , 0.124)
	β_1	0.080	0.079	0.010	(0.059 , 0.099)
	τ_ϕ	0.010	0.011	0.004	(0.005 , 0.015)
	α_1	0.500	0.567	0.093	(0.384 , 0.750)
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	($\vdots$, $\vdots$)
	α_{31}	0.500	0.409	0.125	(0.164 , 0.655)
Escenario 2	β_0	-1.500	-0.146	0.055	(-0.255 , -0.038)
	β_1	0.400	0.398	0.019	(0.362 , 0.435)
	τ_ϕ	0.010	0.006	0.003	(0.004 , 0.012)
	α_1	-0.500	-0.736	0.358	(-1.487 , -0.081)
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	($\vdots$, $\vdots$)
	α_{31}	0.500	0.446	0.171	(0.111 , 0.783)
Escenario 3	β_0	-0.09	-0.086	0.012	(-0.110 , -0.062)
	β_1	0.200	0.195	0.015	(0.165 , 0.225)
	τ_ϕ	0.010	0.011	0.005	(0.009 , 0.023)
	α_1	-0.500	-0.455	0.226	(-0.914 , -0.027)
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	($\vdots$, $\vdots$)
	α_{31}	0.500	0.552	0.142	(0.274 , 0.831)

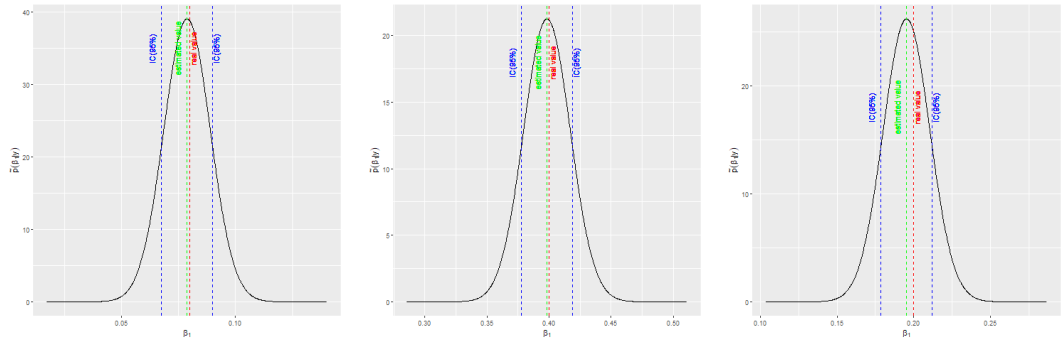
Cuadro 4.2: Tabla resumen de resultados, entre ellos el valor real, media, desviación estándar e IC al 95 % de los parámeros evaluados en los 3 escenarios.

finalmente los intervalos de credibilidad (líneas azules).



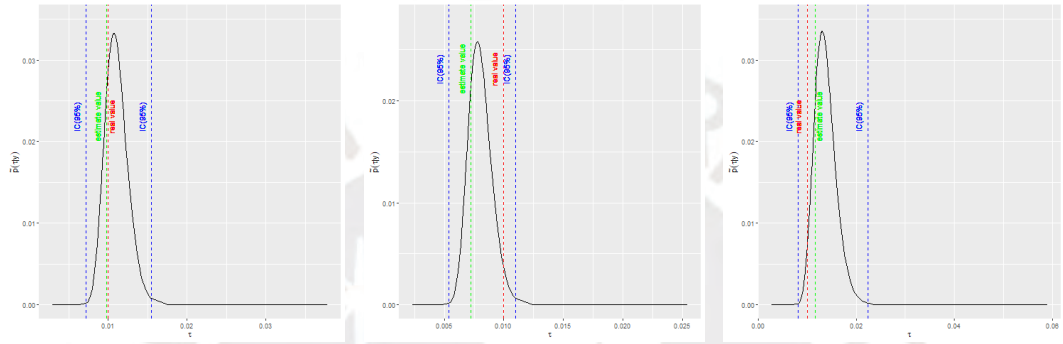
(a) Escenario 1 (b) Escenario 2 (c) Escenario 3

Figura 4.5: Distribución marginal a posteriori del parámetro β_0 en cada uno de los escenarios.



(a) Escenario 1 (b) Escenario 2 (c) Escenario 3

Figura 4.6: Distribución marginal a posteriori del parámetro β_1 en cada uno de los escenarios.

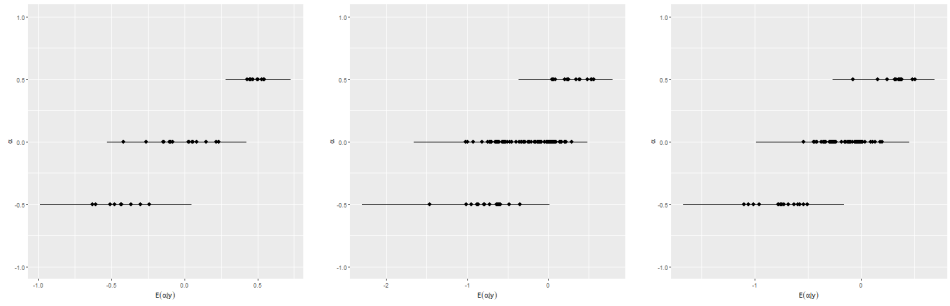


(a) Escenario 1 (b) Escenario 2 (c) Escenario 3

Figura 4.7: Distribución marginal a posteriori del parámetro τ_ϕ en cada uno de los escenarios.

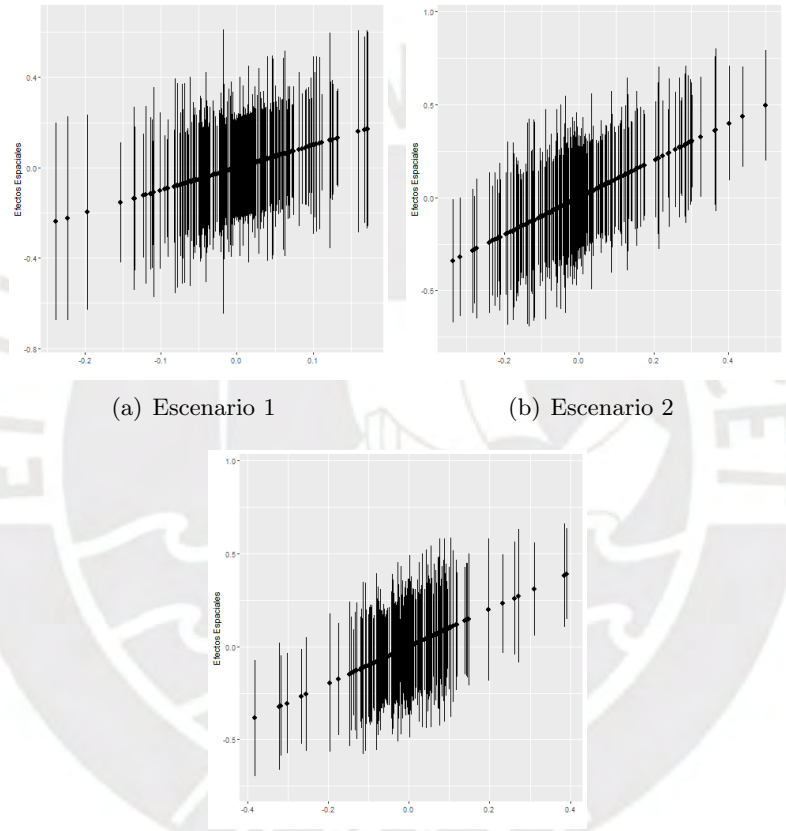
Por otra lado, también se realizaron cálculos con respecto a los efectos espaciales, evidenciamos lo mencionado en la Figura 4.9 tenemos los IC de los efectos espaciales de las 271 regiones de Escocia para la simulación. Dla figura podemos en el eje de las abscisas los efectos espaciales reales de cada región, mientras que en el eje de las ordenadas representa los límites de intervalo de credibilidad al 95 %, además notar que los puntos negritos hacen referencia a la media posteriori del efecto espacial en cada región de Escocia. En conclusión podemos observar que se han podido recuperar los efectos espaciales simulados de forma exitosa.
Posteriormente, se procede a mostrar las estimaciones de la media a posteriori de Y versus los valores simulados. En todos los casos las estimaciones fueron satisfactorias (ver figura 4.10).

A continuación, evidenciamos mediante las Figuras 4.11, 4.12 y 4.13 la presencia de autocorrelación espacial entre las regiones de Escocia. La estimación se ajusta a los valores simulados en cada uno de los escenarios evaluados, cabe indicar que los niveles de colores están relacionadas de forma directa a la variable Y, es decir, a mayor valor que encontremos en las regiónes mayor es el riesgo.
Estos niveles de riesgo logran ser reconocido por los colores, es decir, altos niveles de riesgo están relacionados a los matices del color rojo, mientras que niveles menores de riesgo están relacionados a los matices color amarillo.



(a) Escenario 1 (b) Escenario 2 (c) Escenario 3

Figura 4.8: Valores reales del vector de parámetros α simulado versus la media a posteriori del parámetro α en cada uno de los escenarios.

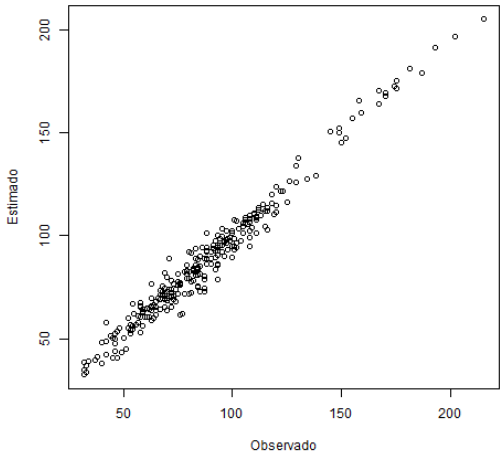


(a) Escenario 1 (b) Escenario 2

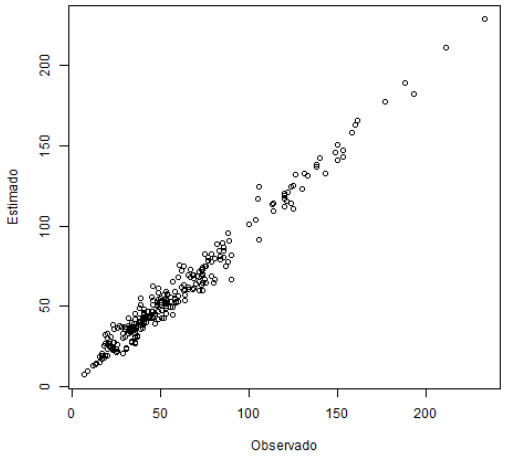
(c) Escenario 3

Figura 4.9: IC al 95 % de los efectos aleatorios espaciales en cada escenario

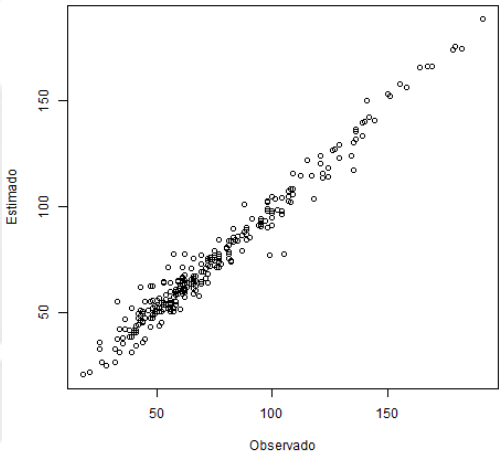
Por otro lado, se procede a verificar si los conglomerados detectados por el modelo son similares a los conglomerados originales que definieron en un inicio, es decir, si la agrupación de conglomerados es precisa con la agrupación inicial. A continuación en el cuadro 4.3 podemos ver los porcentajes de detección de conglomerados.



(a) Escenario 1



(b) Escenario 2



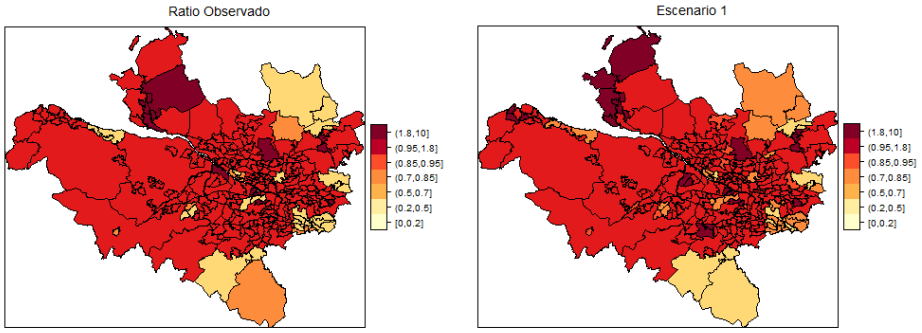
(c) Escenario 3

Figura 4.10: Gráfico de dispersión de los valores simulados vs estimaciones de la variable de estudio Y en cada escenario

	Tabla de resultados
	Precisión
Escenario 1	51 %
Escenario 2	53 %
Escenario 3	52 %

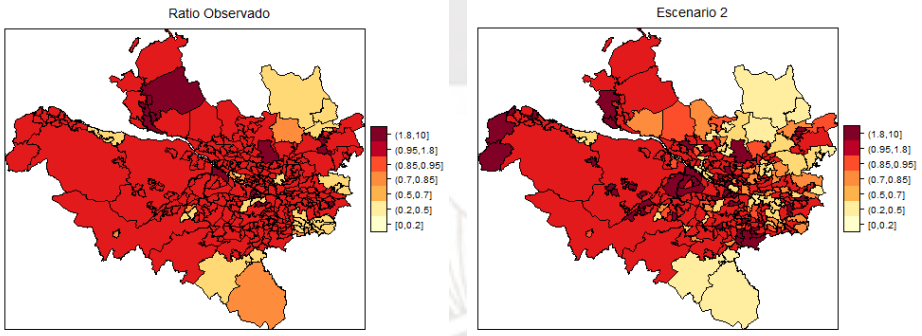
Cuadro 4.3: Tabla resumen de resultados, porcentajes de detección de conglomerados

En conclusión, se produjo una buena recuperación de los valores de los parámetros e hiperparámetros a priori. Ello se suma a que las estimaciones de los efectos espaciales son similares a los efectos espaciales simulados, que se pudo obtener estimaciones de Y que se ajustan con buenos resultados a los datos simulados y que todo lo mencionado anteriormente se realizó en tiempos reducidos de procesamiento.



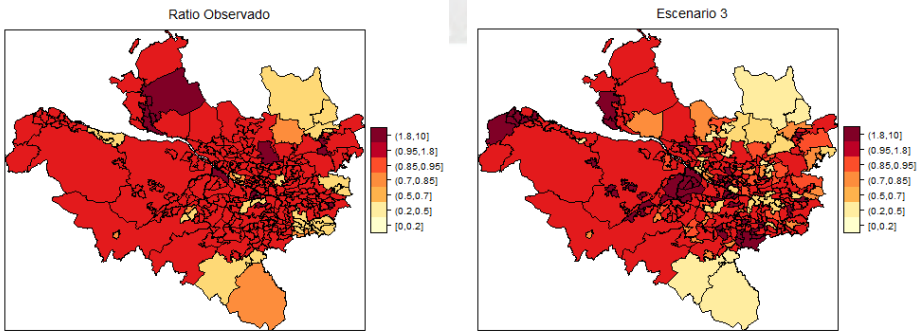
(a) Gráfico de valores simulados (b) Escenario 1

Figura 4.11: Comparativo de los valores simulados vs valores obtenidos en el Escenario 1



(a) Gráfico de valores simulados (b) Escenario 2

Figura 4.12: Comparativo de los valores simulados vs valores obtenidos en el Escenario 2



(a) Gráfico de valores simulados (b) Escenario 3

Figura 4.13: Comparativo de los valores simulados vs valores obtenidos en el Escenario 3

Capítulo 5

Aplicación

En el este capítulo se presentan los resultados de la aplicación del modelo de Poisson, considerando prioris condicionales autoregresivas (CAR) en datos de provincias presentado en el Capítulo 4. Además se realizará una comparación entre los valores estimados comparados a los datos observados. En la presente aplicación se tiene como objetivo identificar distritos con niveles de riesgo de impago para empresas del Perú usando estadística bayesiana (modelos CAR) y aprovechar estos niveles de riesgo para la detección de conglomerados que permitan gestionar y tener un mejor control del riesgo de crédito. Por último, se busca también con la información obtenida identificar factores relacionados al nivel de incumplimiento en las empresas del Perú.

5.1. Riesgo de crédito

5.1.1. Definición e importancia

El cumplimiento de las obligaciones financieras tienen un rol importante para las empresas en el Perú independiente del sector al que pertenezcan, de ahí su importancia de aplicar una correcta gestión de riesgo de crédito proactiva para evitar problemas futuros relacionados con el incumplimiento de pagos (PD).

El riesgo de crédito de acuerdo con la normativa de la superintendencia de banca, seguros y AFP (SBS), es la posibilidad de ocurrencia de eventos que impacten negativamente sobre los objetivos de la empresa o situación financiera (Nro 272-2017).

Uno de los principales conceptos relacionas al riesgo de crédito es la morosidad, que es el problema principal que sufren ciertas entidades de distinto tamaño (Goodhart y Schoenmaker, 1993). Niveles elevados de morosidad en un cartera se convierten en un problema relevante que llega a comprometer la viabilidad en el largo plazo de la entidad para luego afectar a su propio sistema. Como consecuencia, los altos niveles de morosidad de créditos conllevan a problemas de liquidez, los cuales si no son respaldados con planes de contingencia pasan a convertirse en problemas de solvencia, determinando probablemente la liquidación de la empresa (Freixas et al., 1994).

Mencionado los conceptos anteriores, resulta importante que las empresas en el Perú cuenten con una adecuada gestión de crédito. Las metodologías administrativas mediante el uso de herramientas (modelos de gestión del riesgo de crédito) permiten reducir el riesgo o mantenerlo

en medidas aceptables para asegurar un correcto funcionamiento en las empresas.

5.1.2. Tasa de incumplimiento (PD) de empresas del Perú en el sistema financiero

El riesgo de crédito en empresas del Perú resulta muy importante pues es sistémico por naturaleza, es decir, su deterioro involucra la totalidad de la economía nacional debido a su importancia como agentes dinámicos en la economía. Por ello, no sólo un adecuado control de niveles de morosidad beneficia de forma directa a las empresas, sino que también generan una estabilidad macroeconómica del país.

En el Perú, la tasa de incumplimiento (PD) en el sistema financiero para los períodos del 2018 y 2019 se viene encontrando en niveles constantes de PD (ver figura 5.1) producto de la inclusión de mecanismos que permitan ayudar a la gestión y aceptación del riesgo.

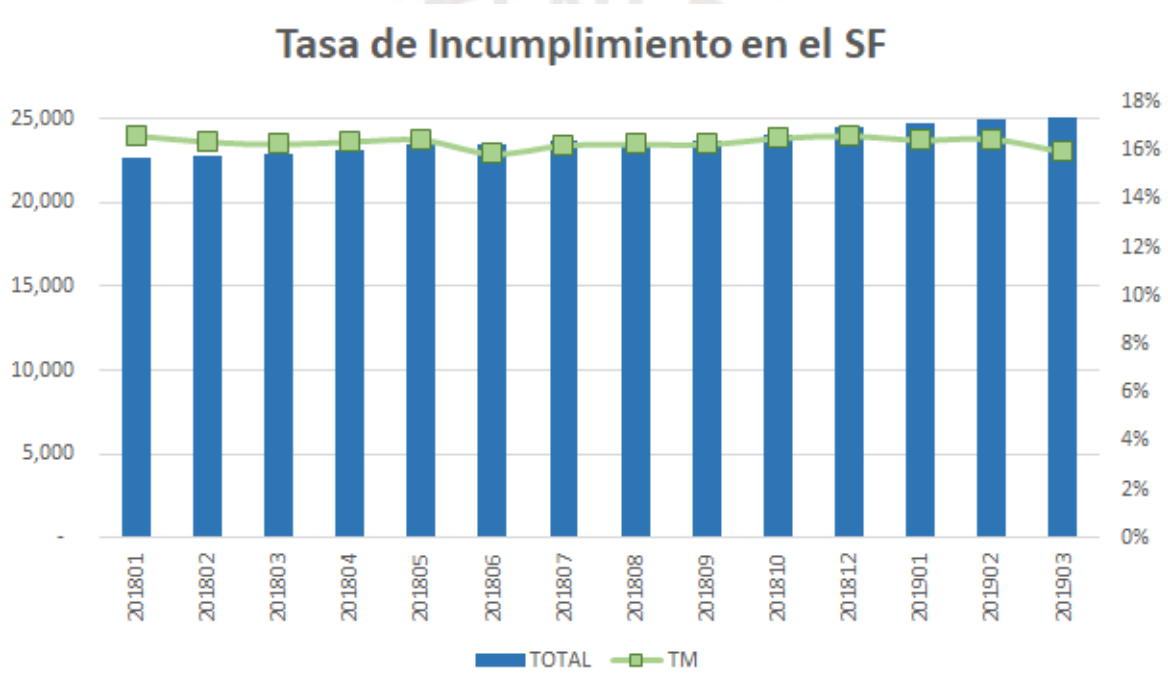


Figura 5.1: Tasa de incumplimiento en el sistema financiero - Empresas del Perú (201801 - 201903)

Finalmente, resulta relevante mencionar la importancia de contar con la precisión de modelos a utilizar para la medición del riesgo de crédito para un adecuada evaluación y mitigación del mismo.

En conclusión, se debe atender que la gestión de riesgo de crédito de forma integral que comprende diferentes procedimientos relacionados al comportamiento de pago permitiendo tener una mejor evaluación del riesgo permitiendo que las empresas peruanas pueden cumplir con sus objetivos comerciales, logrando de esta forma realizar un crecimiento sostenido gracias a una adecuada gestión que ayuda a la mitigación del riesgo.

5.2. Análisis exploratorio de los datos

Los datos provienen de los datos administrativos de los créditos a empresas con las entidades financieras registradas en el Reporte Consolidado de Créditos (RCC). Esta información es consolidada mes a mes por la Superintendencia de Banca, Seguros y AFP (en adelante SBS). La información del RCC corresponde al saldo de crédito de cada empresa por entidad bancaria en los periodos del 2018 y 2019. Se considera como variable dependiente la tasa de incumplimiento (en adelante PD), cuya definición es el ratio entre en número de empresas con más de 60 días de atraso en los próximos 12 meses y total de empresas, y como covariables a la antigüedad de las empresas, estado del registro único del contribuyente (RUC), clasificación industrial internacional uniforme de todas las actividades económicas (CIIU), tipo de persona entidad y días de atraso en el sistema financiero.

Cada observación para el caso de estudio corresponde a los créditos que se otorgaron a las empresas en cada uno de los departamentos, provincias y distritos del Perú. Se muestra acontinuación un resumen de PD para las empresas del Perú a nivel departamental (ver figura 5.2) donde podemos observar regiones con mayor PD (recordar que tenemos como PD de referencia del gráfico 5.1, que representa la PD del sistema financiero), con ayuda del gráfico podemos observar que principalmente las regiones del norte (Tumbes, Piura, Lambayeque, Loreto y Amazonas), regiones del sur (Puno), además de algunas regiones del centro como Junín y Ayacucho (ver cuadro 5.1).

Por otro lado, es importante mencionar que también existen departamentos con PD menor a la PD del sistema financiero, principalmente encontramos departamentos en las regiones al sur este (Madre de Dios y Cusco), regiones del sur (Moquegua), regiones del centro (Huánuco) y Lima.

Todo lo mencionado anteriormente lo resumimos en el cuadro 5.2

	Tasa de Incump (PD)
Tumbes	24.8 %
Ayacucho	23.4 %
Piura	21.8 %
Lambayeque	21.3 %
Puno	20.8 %
Junin	20.7 %
Loreto	20.6 %
Amazonas	20.6 %

Cuadro 5.1: Tabla resumen de departamentos del Perú con mayor PD

5.2.1. Medidas de asociación espacial

En esta sección se describe la correlación espacial encontrada a partir de estadísticos que nos resumen la autocorrelación espacial de la variable tasa de incumplimiento entre las provincias del Perú, estos son el índice de Moran y la C de Geary (Ripley, 1981).

Los estadísticos de correlación mencionados anteriormente, miden la dependencia espacial, la cual supone la existencia de correlación entre provincias próximas entre sí en un conjunto de datos.

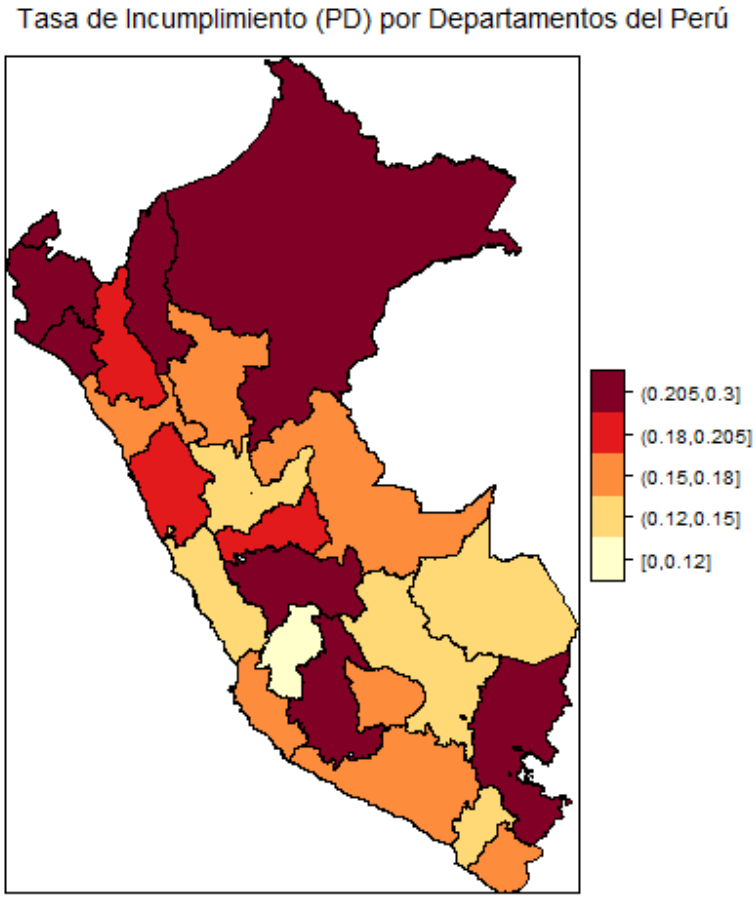


Figura 5.2: Tasa de incumplimiento por Departamentos - Empresas del Perú (201801 - 201903)

Fuente: Elaboración propia

	Tasa de Incump (PD)
Apurímac	15.2 %
La Libertad	15.1 %
Lima	15.0 %
Moquegua	15.0 %
Huánuco	14.8 %
Madre de Dios	14.7 %
Cusco	14.5 %
Huancavelica	10.0 %

Cuadro 5.2: Tabla resumen de departamentos del Perú con menor PD

Con el test de Moran se busca contrastar la hipótesis nula de distribución espacial aleatoria, comparando los valores de cada unidad de provincia (provincias del Perú). Los resultados encontrados (ver tabla 5.3) se confirma evidencia estadística para afirmar que la variable de estudio (tasa de incumplimiento) posee correlación espacial a un nivel de significancia del 5 %, dado que el valor-p es inferior a 0.05 rechazamos la hipótesis nula.

Tabla: Test I de Moran	
I-Moran	p-value
0.067	0.045

Cuadro 5.3: Test I de Moran

El segundo estadístico mencionado es el C de Geary, teniendo los siguientes resultados (ver tabla 5.4), como se sabe el valor de C nunca es negativo, además valores pequeños (entre 0 y 1) indican asociación espacial positiva (Cliff y Ord, 1973), evidenciando existencia de correlación entre observaciones próximas entre sí en un conjunto de datos.

Tabla: Test C de Geary	
C de Geary	varianza
0.979	0.003

Cuadro 5.4: Test C de Geary

Como último estadístico tenemos al Getis-Ord (para más detalle revisar el anexo B) que nos evidencia rastros de presencia de altos o bajos niveles de conglomeración, es decir, si $G > E$ existe altos niveles de agrupación mientras que si $G < E$ evidencia bajos nivel de agrupación en el conjunto de datos.

Tabla: Test Getis-Ord	
G estadístico	Expectation
2.75e-02	2.72e-02

Cuadro 5.5: Test Getis-Ord

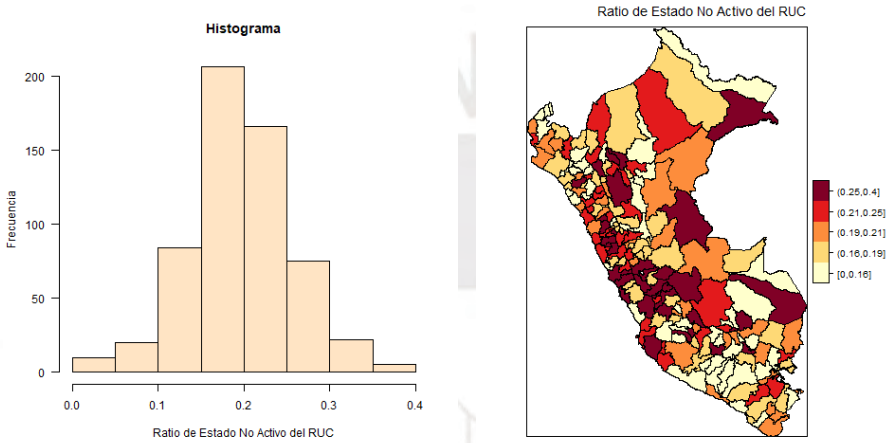
Los resultados del test Getis-Ord (ver tabla 5.5), nos muestra la existencia suficiente de evidencia estadísticamente significativa para afirmar que la variable de estudio (tasa de incumplimiento) posee niveles relevantes de agrupación.

Se tiene un total de siete covariables recolectadas por la Superintendencia Nacional de Aduanas y de Administración Tributaria (SUNAT), además de información del Reporte Consolidado de Créditos (RCC). En la tabla 5.6 se muestra la estructura de datos que se ha utilizado.
Por otro lado, de las covariables mencionadas anteriormente citamos algunas adicionales como la covariable ratio de estado del RUC no activo, la cual hace referencia al porcentajes de empresas con el RUC no activo, según la Superintendencia Nacional de Aduanas y de Administración Tributaria (SUNAT). Como podemos observar en la figura 5.3(a), el histograma de esta covariable muestra una distribución asimétrica a la izquierda, lo cual implica que son pocas las provincias con empresas que poseen su RUC no activo (posible informalidad). Mientras que el mapa de la figura 5.3(b) nos permite ver qué provincias del Perú tienen mayores ratios de RUC no activo, como son las provincias de Cusco, Huaral y Satipo.

Otra de las covariables relevantes es la antigüedad del RUC de las empresas, la cual hace referencia a los años de antigüedad de las empresas en el Perú, según fuente Superintenden-

Tabla de variables			
Fuente	Variable	Descripción	Tipo
SUNAT	HABIDO	% de empresas ubicadas en su local de negocio	Continua
	NO HABIDO	% de empresas no ubicadas en su local de negocio	Continua
	ACTIVO	% de empresas con RUC activo	Continua
	NO ACTIVO	% de empresas con RUC no activo	Continua
	ANTIGÜEDAD	Experiencia empresarial medida en años	Continua
RCC	FLAG PE	% de empresas medianas	Continua
	FLAG ME	% de empresas pequeñas	Continua
	CT60 MALOS	♯ de empresas malas(días de atraso mayor a 60 días)	Continua
	CT60 TOTAL	♯ total de empresas	Continua

Cuadro 5.6: Tabla resumen de variables analizadas para el modelo propuesto



(a) Gráfico de histograma (b) Ratio de RUC no activo

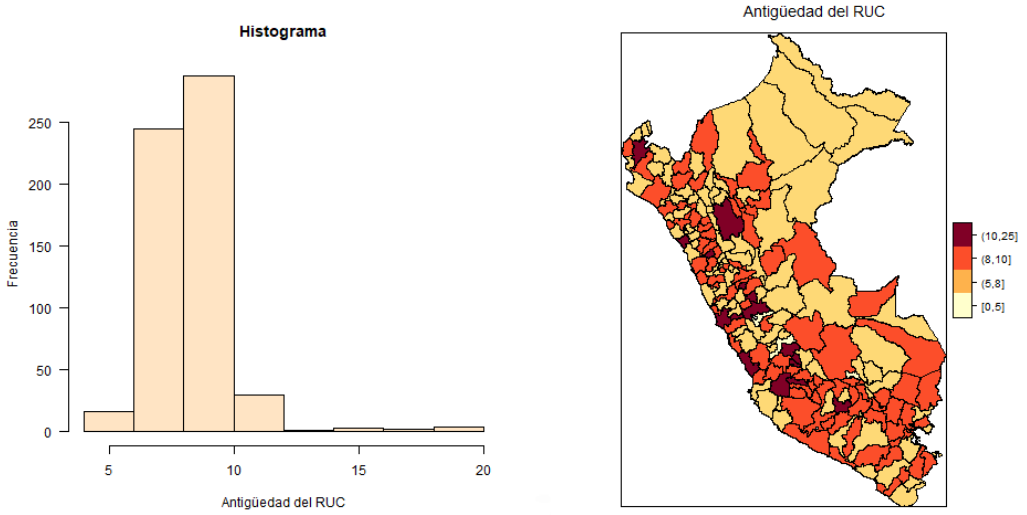
Figura 5.3: Gráficos de la covariable ratio de estado del RUC no activo

cia Nacional de Aduanas y de Administración Tributaria (SUNAT). Como podemos observar en la figura 5.4(a) el histograma de esta covariable presenta una distribución asimétrica a la derecha, lo cual indica que son pocas las provincias con antigüedad de RUC mayor a 10 años. Mientras que el mapa en la figura 5.4(b) nos permite ver qué provincias del Perú tienen empresas con mayor experiencia empresarial, como por ejemplo Cañete, Antabamba, Ambo, Trujillo y Arequipa.

Por otro lado, en el siguiente gráfico (ver figura 5.5) podemos ver la matriz de correlaciones de la variable respuesta TASA (tasa de incumplimiento o PD) con el resto de covariables, a su vez podemos ver problemas de correlación entre ciertas covariables, por lo que se pasarán a retirarlas teniendo en cuenta también la interpretación en relación con la variable de estudio.

5.3. Resultados

En esta parte del trabajo se presenta los resultados de las estimaciónes de las dos propuestas de modelos mencionadas en el capítulo 3, posteriormente se compara ambas propuestas de modelos para ver que modelo es mejor mediante el criterio de la raíz del error cuadrático medio (RSME), criterios de información como la devianza y Watanabe (DIC y WAIC) y el



(a) Gráfico de histograma (b) Antigüedad de RUC

Figura 5.4: Gráficos de la covariable antigüedad de RUC

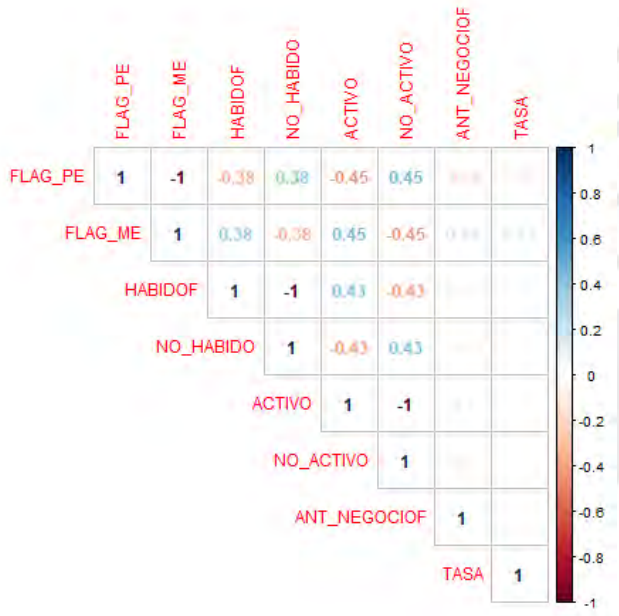


Figura 5.5: Matriz de correlaciones

logaritmo de pseudo-verosimilitud marginal (LPML).

5.3.1. Modelo de Poisson para datos de provincias

Esta primera propuesta consiste en diseñar un modelo de Poisson log-linealizado, donde el parámetro tasa de incumplimiento (PD por sus siglas en inglés) o nivel de morosidad es representado por covariables y un conjunto de efectos aleatorios, los cuales toman en cuenta la autocorrelación espacial, y son modelados generalmente por una priori condicional auto-

regresiva (CAR).

Dicho lo anterior, en el siguiente cuadro (ver tabla 5.7) se muestra un resumen de los valores estimados por el primer modelo propuesto (modelo Poisson - CAR), en el cuadro podemos obervar las estimaciones a posteriori de la media y la desviación estándar, además de ello se puede observar que los coeficientes de regresión $\beta_1, \beta_2, \beta_3, \beta_4$ de las covariables FLAG PE, ACTIVO, NO HABIDO y ANTIGÜEDAD, son significativos.

	Tabla de resultados			
	Parámetro	Media	Desv. est.	95 % IC
Poisson CAR	β_0	-0.972	0.570	(-1.585 , -0.343)
	β_1	-2.253	0.341	(-2.622 , -1.879)
	β_2	1.137	0.495	(0.588 , 1.667)
	β_3	-0.173	0.307	(-0.509 , -0.162)
	β_4	0.027	0.010	(0.016 , 0.0338)
	τ_ϕ	0.010	0.05	(0.005 , 0.031)

Cuadro 5.7: Tabla resumen de resultados, entre ellos los valores de la media, desviación estándar e IC al 95 % de los parámetros encontrados en la primera propuesta (modelo Poisson - CAR))

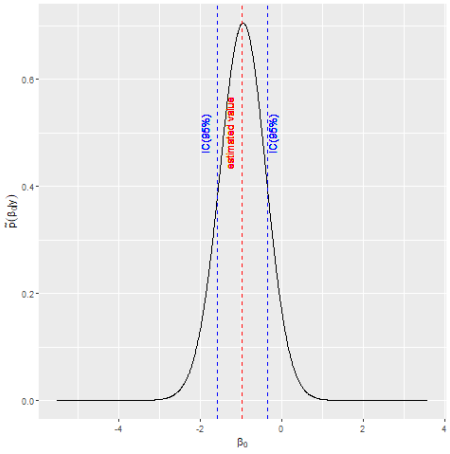
Dado los valores de los coeficientes de las covariables podemos dar una interpretación en relación a la variable de estudio (Tasa de incumplimiento o PD), es decir, si exponenciamos la interpretación se lee como el riesgo relativo como por ejemplo la variable ACTIVO (% de empresas con RUC activo) ante el aumento de una unidad del ratio RUC activo supone una disminución de PD en un 16 %=1-exp(-0.173), esta lectura tiene el mismo sentido negativo para los parámetros β_1 y β_3 (FLAG PE y ACTIVO) con respecto a la tasa de incumplimiento (PD). Por otro lado, los coeficientes positivos β_2 y β_4 (NO HABIDO y ANTIGÜEDAD) muestran una relación positiva en relación a la tasa de incumplimiento (PD).

En las siguiente gráfico (ver figuras 5.6 y 5.7) se puede observar las distribuciones marginales a posteriori de cada parámetro e hiperparámetro, donde las líneas de color azul hacen referencia los intervalos de credibilidad al 95 % y la línea de color rojo a referencia a la media de los valores estimados. Con respecto a los efectos espaciales, en el siguiente gráfico (ver figura 5.8) se observa la estimación de la media a posteriori y los intervalos de credibilidad al 95 % de los efectos espaciales de las 196 provincias del Perú.

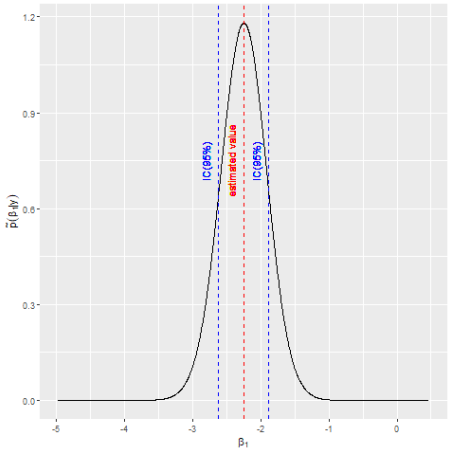
A continuación, con la primera propuesta del modelo Poisson - CAR, se realizó la estimación de la tasa de incumplimiento (PD) para cada provincia del Perú, para esta primera propuesta se calculó la raíz del error cuadrático medio (RMSE). Este indicador se define la siguiente manera:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{196}(y_i - \hat{y}_i)^2}{196}}, \tag{5.1}$$

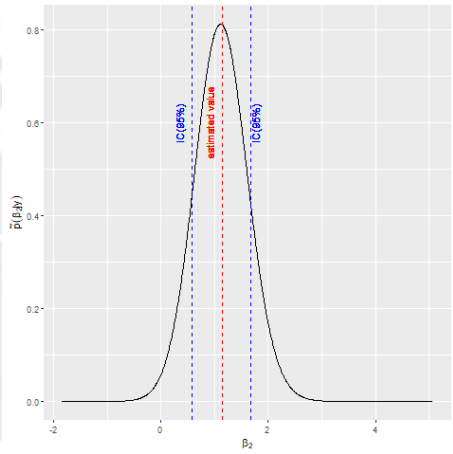
donde y_i es la tasa de incumplimiento (PD) para la i-ésima provincia.



(a) Intercerpto



(b) Variable Flag PE (% de empresas medianas)



(c) Variable No Habido (empresas no ubicadas en su local de negocio)

Figura 5.6: Gráfico de las distribuciones marginales a posteriori de los coeficientes β_0,β_1 y β_2 para el modelo Poisson - CAR

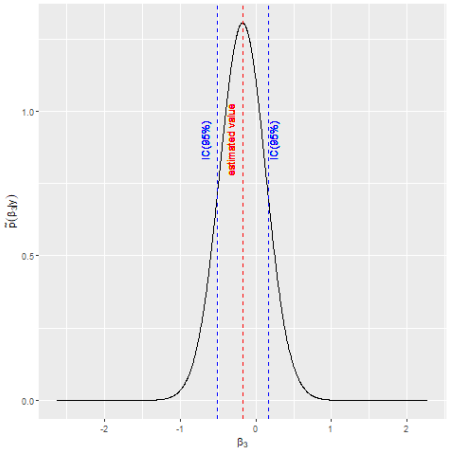
De los resultados obtenidos de esta primera propuesta de modelo, se obtuvo los siguientes resultados, mostrados en la siguiente tabla:

	Tabla de resultados		
	RMSE	DIC	WAIC
Poisson - CAR	0.05	1486.10	1518.83

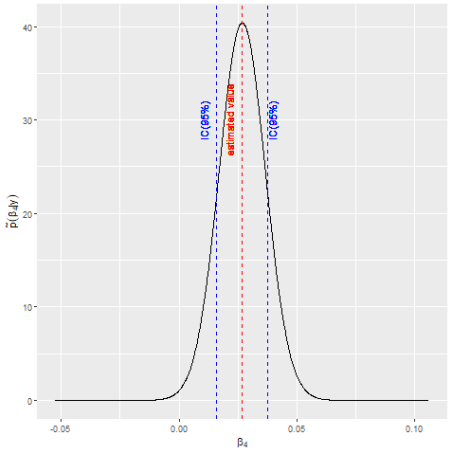
Cuadro 5.8: Tabla resumen de resultados - RSME, DIC y WAIC del modelo Poisson - CAR

Finalmente, en los siguientes gráficos se muestran mapas que bosquejan las estimaciones de la media a posteriori de la variable de estudio (tasa de incumplimiento), y también se logra comparar con el uso del digrama de dispersión los valores reales vs los valores estimados del número de empresas morosas en cada provincia del Perú a través del modelo Poisson - CAR.

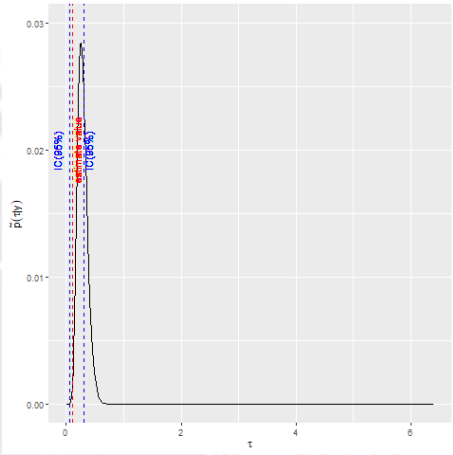
Por otro lado, tenemos que el gráfico de diagrama de dispersión (ver figura 5.10) se observa las estimaciones de la empresas que incumplieron con sus pagos o riesgo de crédito (días



(a) Variable Estado activo del RUC



(b) Variable Antigüedad del RUC



(c) Hiperparámetro de precisión τ_ϕ

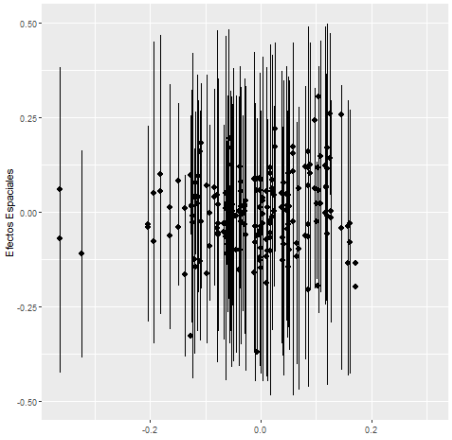
Figura 5.7: Gráfico de las distribuciones marginales a posteriori de los coeficientes β_3, β_4 e hiperparámetro de precisión τ_ϕ, para el modelo Poisson - CAR

de atraso mayor a 60) se ajustan a los valores reales (ver figura 5.10(b)), sin embargo al calcular las estimaciones para los valores de tasas de incumplimiento (PD), solo para valores por debajo de 0.2 se ajustan mejor con el modelo que se propone.

5.3.2. Modelo de Poisson por conglomerados para datos de provincias

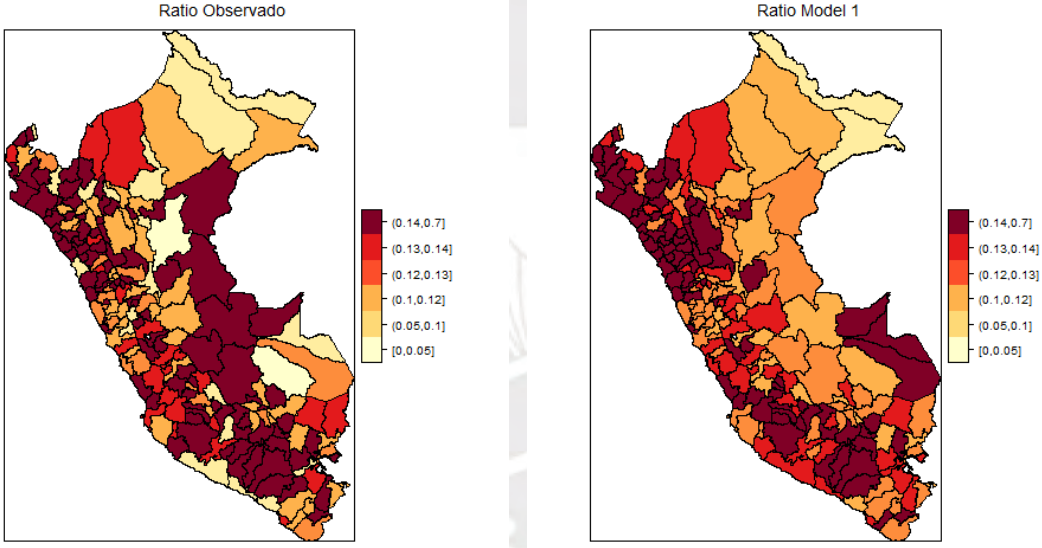
Esta segunda propuesta consta de dos pasos. En la primera etapa se realiza la generación de conglomerados por los criterios de agrupación aglomerativa jerárquica (entre los criterios tenemos enlace simple, enlace promedio, enlace completo y el método Ward), los cuales se definieron en el sección 2.1, y que consisten en juntar aquellas provincias del Perú que son similares mientras separa aquellas provincias que son diferentes de acuerdo a la variable de estudio (tasa de incumplimiento - PD).

En la segunda etapa, tomando en cuenta los estudios de Anderson et al. (2014), se aplicará un algoritmo de conglomerados a priori a los datos de empresas expuestas a la morosidad, pues es probable que se exhiban patrones espaciales, por ende, esta segunda propuesta de modelo se resume en diseñar un modelo que se ajuste a la variable de estudio , PD, una



(a) Efectos aleatorios espaciales de las 196 provincias del Perú

Figura 5.8: Intervalos de credibilidad al 95 % de los efectos aleatorios espaciales del modelo Poisson - CAR



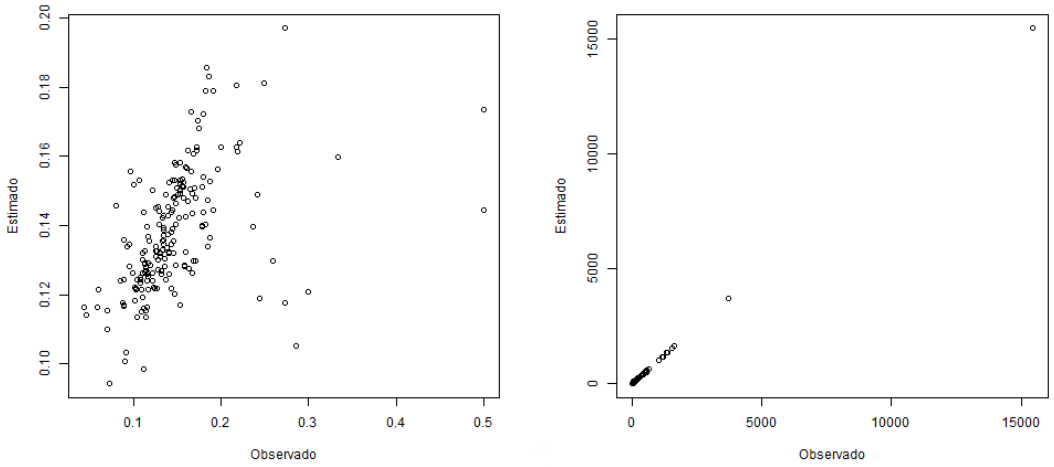
(a) PD - Provincias del Perú (b) Valores estimados con el modelo Poisson - CAR

Figura 5.9: Mapa comparativo de PD reales vs PD estimadas por el modelo Poisson - CAR

vez definido el conglomerado al que pertenece la provincia del Perú, es decir, calcularemos el nivel de riesgo , PD, con ayuda de covariables y un conjunto de efectos aleatorios, los cuales toman en cuenta la autocorrelación espacial, todo lo mencionado anteriormente haciendo uso del modelo de Poisson considerando prioris condicionales autoregresivas (CAR).

1. Conglomerados a priori usando agrupación aglomerativa jerárquica

En esta sección se crean conglomerados usando criterios de agrupación aglomerativa jerárquica, entre los criterios tenemos enlace simple, enlace promedio, enlace completo y el método ward. Se utilizaron las variables mencionadas en la sección 5.2.1 (ver tabla 5.6), es importante mencionar la diferencia entre los métodos de conglomerados, pues tomaremos co-



(a) Valores estimados de PD

(b) Valores estimados ♯ de empresas con riesgo de crédito

Figura 5.10: Comparativo de los valores reales vs valores estimados con el modelo Poisson - CAR

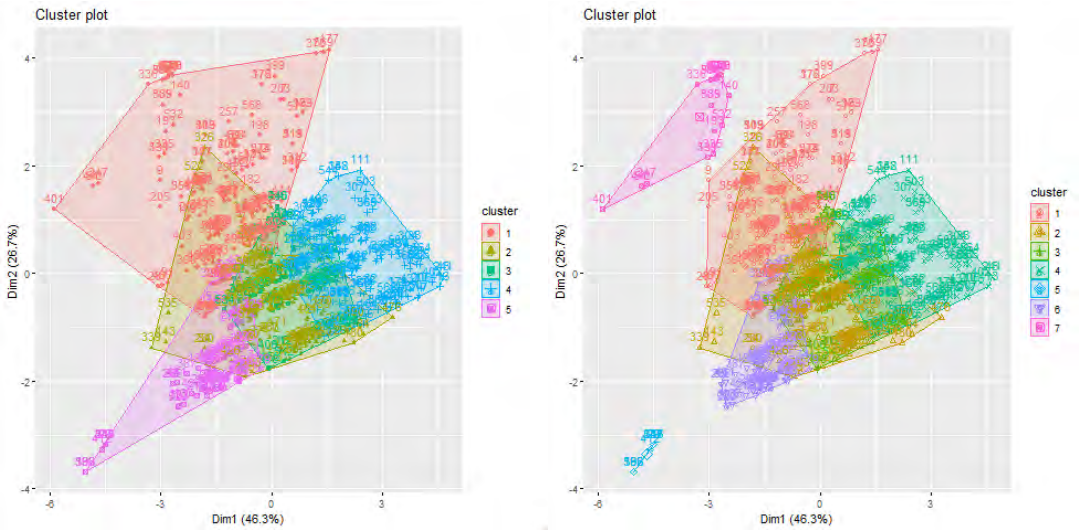
mo medida de comparación el coeficiente de aglomeración, que mide la cantidad de estructura de agrupamiento encontrada , donde los valores más cercanos a 1 sugieren una estructura de agrupación fuerte.
Evaluados los distintos métodos (mencionados en el capítulo 2 sección 2.1), en el siguiente cuadro (ver tabla 5.9) se puede ver la diferencias de los coeficientes de aglomeración, donde se aprecia que el método de Ward nos evidencia un mejor coeficiente de aglomeración, parte de este resultado se debe a que el método mencionado anteriormente junto con el método del enlace medio son los menos sensibles a outliers.

	Tabla de resultados			
	E. simple	E. completo	E. medio	E. Ward
C. aglomeración	0.9457	0.9772	0.9607	0.9934

Cuadro 5.9: Tabla resumen - Comparación de coeficientes de aglomeración

Definido lo anterior, el siguiente paso sería escoger el número de conglomerados que permita diferenciar los conglomerados de provincias del Perú, se usaron conglomerados con grupos de dos hasta 7 grupos, además se incorporó como regla limitante que cada grupo contenga como mínimo un porcentaje mayor al 10 % del total de empresas en el Perú de forma que los conglomerados sean materialmente significantes en cada período de análisis, a continuación en el siguiente gráfico (ver figura 5.11) se muestra la vista de provincias del Perú considerando cinco y siete conglomerados (ver figuras 5.11(a) y 5.11(b) respectivamente), en general no se ve diferencias relevantes, a excepción de la materialidad de los grupos.

Dados los resultados de los conglomerados se procede a utilizar un criterio adicional según el riesgo de crédito empresarial (RCE). Este criterio consiste en asociar provincias con PD similar, y de esta forma reducir el número de conglomerados con participación menor al 10 %. En la tabla 5.10 se observa que, considerando el método Ward, los grupos finales son cinco.



(a) Gráfica considerando 5 grupos (b) Gráfica considerando 5 grupos

Figura 5.11: Comparativo de conglomerados considerando 5 y 7 grupos respectivamente

Por otro lado, en la tabla 5.11 podemos observar la descripción de los cinco grupos finales bajo el método Ward y el criterio RCE.

Tabla: Criterio de Ward - 5 grupos		
Grupo	PD	Participación (%)
1	16.0 %	27 %
2	14.0 %	24 %
3	13.0 %	12 %
4	14.0 %	24 %
5	13.5 %	23 %

Cuadro 5.10: Conglomerados bajo el criterio Ward.

Tabla: Criterio de RCE + Ward		
Grupo	PD	Participación(%)
1	15.0 %	24 %
2	14.5 %	24 %
3	13.0 %	14 %
4	14.0 %	24 %
5	16.0 %	14 %

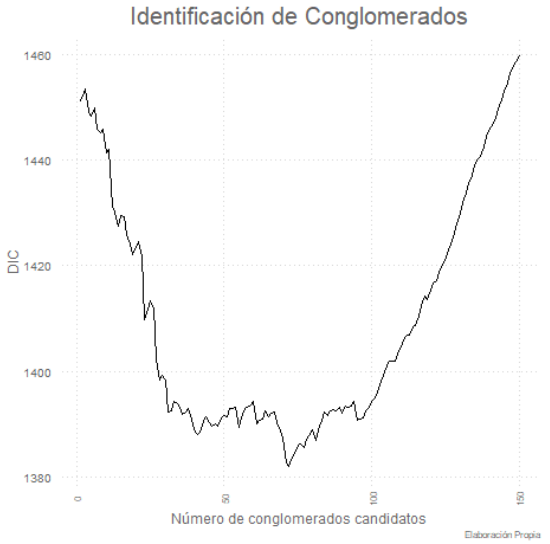
Cuadro 5.11: Conglomerados bajo el criterio RCE y de Ward.

2. Conglomerados para datos de provincias

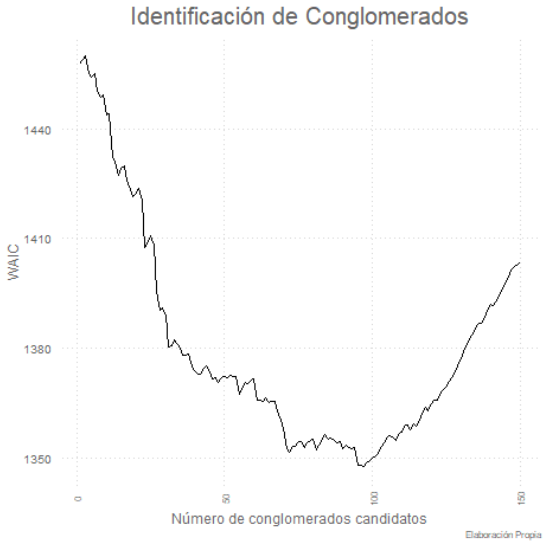
Etapa 1 : Generación de conglomerado usando agrupación aglomerativa jerárquica

Esta primera etapa, consiste en utilizar los conceptos mencionados en la sección 5.3.2. Considerando estos conceptos se utiliza un algoritmo de conglomerados a priori a la data de empresas del Perú expuestas a la morosidad. El algoritmo en mención consiste en juntar aquellas provincias del Perú que son similares mientras separamos aquellas provincias que

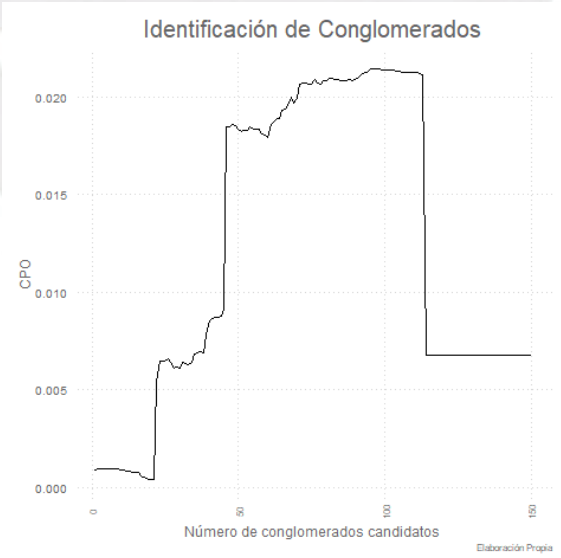
son diferentes considerando la magnitud de la variable de estudio PD y respetando la estructura de contiguidad espacial de las provincias. Este proceso es beneficioso para identificar conglomerados con similar PD (riesgo de morosidad). Posteriormente, se selecciona el mejor número de conglomerado de un conjunto de $m = 150$ candidatos, la elección del número de conglomerados se realizó usando criterios de selección de modelos. Esta segunda etapa, consiste en el uso del modelo bayesiano poisson loglinealizado que se ajusta a la PD de cada número candidato de conglomerados para los datos de estudio, eligiendo como el mejor modelo aquel que minimiza el DIC, WAIC y maximiza el LPML. En particular, en las figuras 5.12(a) y 5.12(b), se observa que el número óptimo de conglomerados es $k = 71$.



(a) Selección del número óptimo de conglomerados mediante menor DIC



(b) Selección del número óptimo de conglomerados mediante menor WAIC



(c) Selección del número óptimo de conglomerados considerando mayor LPML

Figura 5.12: Gráfico resumen de indicadores estadísticos para la selección del número óptimo de conglomerados

Etapa 2 : Estimación del modelo usando conglomerados a priori

Después de seleccionar el número de conglomerados de los 150 candidatos usando el modelo bayesiano poisson loglinealizado que se ajusta a tasa de incumplimiento, se muestra continuación los resultados de la segunda propuesta de modelo, es decir, de la estimación de la PD aplicando del algoritmo de conglomerados a priori a la data de empresas del Perú expuestas al riesgo de crédito. Además se calculó la raíz cuadrada del error cuadrátido medio (RMSE) para este modelo. Los resultados se muestran en la tabla 5.12.

Si se compara el modelo Poisson - CAR con el modelo Poisson-CAR-conglomerado, se concluye que el segundo modelo se ajusta mejor a los datos, justificando la necesidad de incluir efectos fijos para los conglomerados.

	Tabla de resultados		
	RMSE	DIC	WAIC
Poisson-CAR conglomerado	0.014	1382.05	1351.54

Cuadro 5.12: RSME, DIC y WAIC del modelo Poisson - CAR incluyendo conglomerados a priori

Se revisaron los parámetros de esta segunda propuesta de modelo los cuales son resumidos en la tabla 5.13. Se muestra las estimaciones a posteriori de la media y la desviación estándar, además de ello se puede observar que los coeficientes de las covariables son significativos.

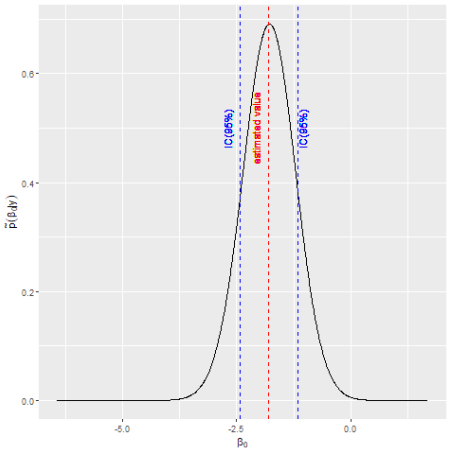
	Tabla de resultados			
	Parámetro	Media	Desv. est.	95 % IC
Poisson - CAR conglomerados	β_0	-1.794	0.651	(-2.422 , -1.156)
	β_1	-0.724	0.351	(-1.094 , -0.363)
	β_2	0.579	0.575	(0.005 , 1.147)
	β_3	-0.236	0.247	(-0.597 , -0.124)
	β_4	0.024	0.010	(0.016 , 0.034)
	τ_ϕ	0.010	0.004	(0.004 , 0.015)

Cuadro 5.13: Tabla resumen de media, desviación estándar e IC al 95 % de los parámetros e hiperparámetros encontrados en el modelo Poisson - CAR incluyendo conglomerados.

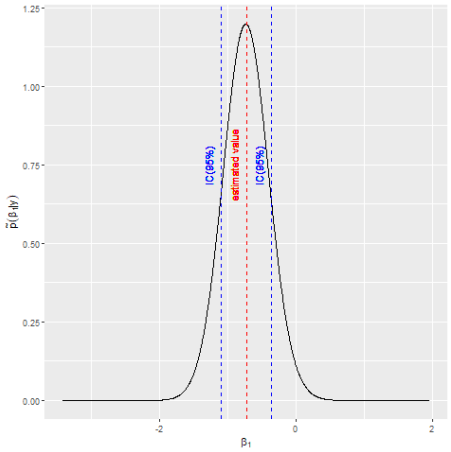
Dados los valores de los coeficientes de regresión podemos realizar una interpretación en relación a la variable de estudio PD al exponenciarlas, por ejemplo ante el aumento de una unidad del ratio RUC activo se obtiene una disminución de 21 %=1-exp(-0.236) del PD. Esta lectura tiene el mismo sentido negativo para los parámetros β_1 y β_3 (FLAG PE y ACTIVO) con respecto a la PD. Por otro lado, los coeficientes positivos β_2 y β_4 (NO HABIDO y ANTIGÜEDAD) muestran una relación positiva en relación a la PD.
Por otro lado, tenemos que α es pequeño, pero también significativo, el IC no contiene al 0. Además tener un τ_ϕ pequeño indica que el modelo captura la variabilidad espacial.

En el siguiente gráfico (ver figuras 5.13 y 5.14) se puede observar las distribuciones marginales a posteriori de cada parámetro e hiperparámetro, donde las líneas de color azul hacen

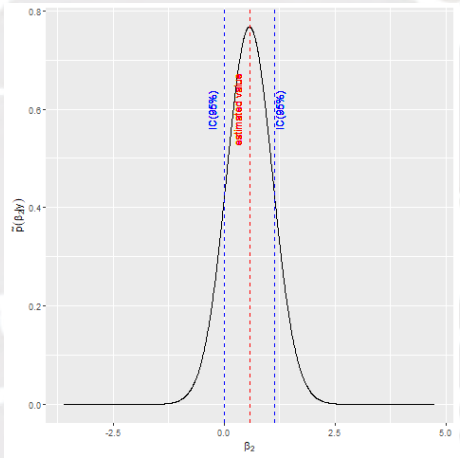
referencia los intervalos de credibilidad al 95 % y la línea de color rojo hace referencia a la media de los valores estimados.



(a) Intercerpto



(b) Variable Flag PE (% de empresas medianas)

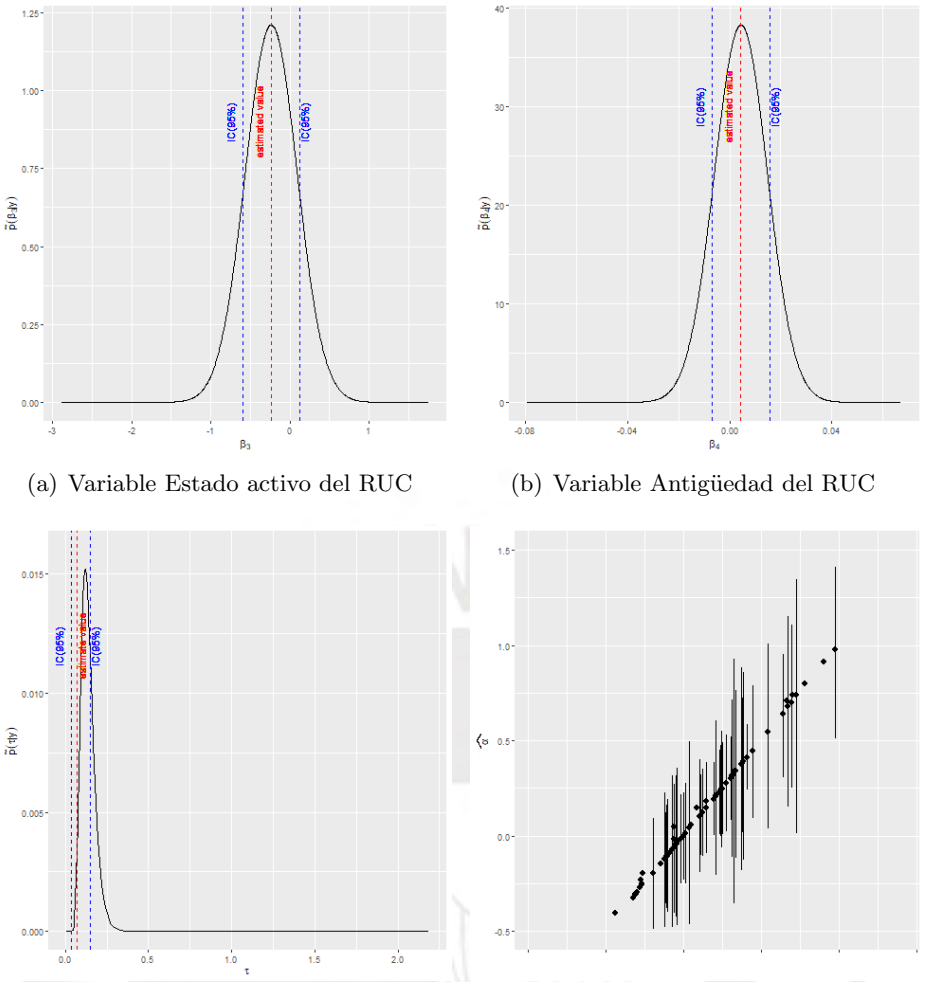


(c) Variable No Habido (empresas no ubicadas en su local de negocio)

Figura 5.13: Gráfico de las distribuciones marginales a posteriori de los coeficientes β_0,β_1 y β_2 del modelo Poisson - CAR con conglomerados

Con respecto a los efectos espaciales, en el siguiente gráfico (ver figura 5.15) se observa los intervalos de credibilidad de los efectos espaciales de las 196 provincias del Perú al 95 % y la estimación de la media a posteriori del efecto espacial, representada por el punto negro en el gráfico.

Posteriormente, se muestra en los siguientes gráficos el mapa que bosqueja las estimaciones de la media a posteriori de la variable de estudio PD. Además, con ayuda del diagrama de dispersión, se logra comparar los valores reales contra los valores estimados del número de empresas morosas en cada provincia del Perú a través del modelo Poisson - CAR con conglomerados apriori. En el primer gráfico (ver figura 5.16(a)) se observa las tasas de incumplimiento reales, evidenciando la diferencia de los niveles de PD por cada provincia del



(a) Variable Estado activo del RUC

(b) Variable Antigüedad del RUC

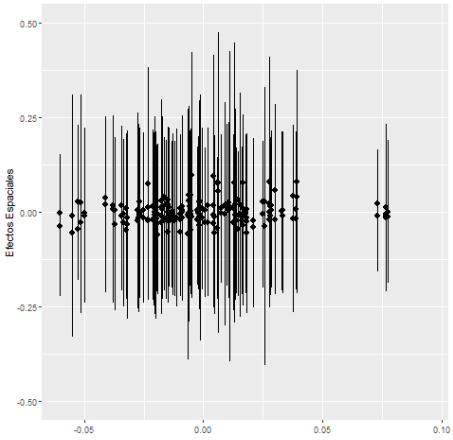
(c) Parámetro de precisión τ_ϕ

(d) Parámetros de conglomerados α_j

Figura 5.14: Gráfico de las distribuciones marginales a posteriori de los coeficientes β_3, β_4 e hiperparámetros τ_ϕ y α_j del modelo Poisson - CAR con conglomerados.

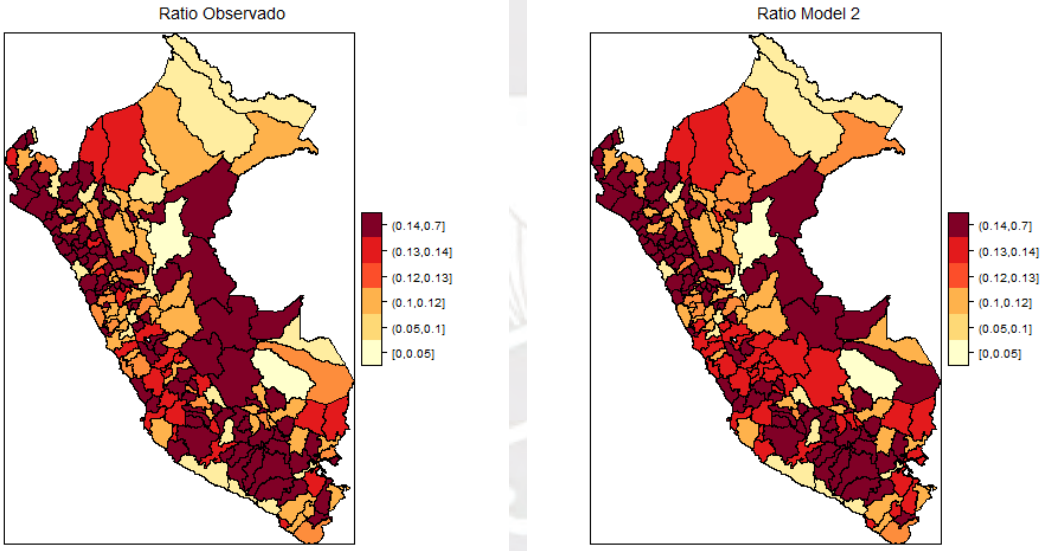
Perú, donde las provincias de mayor PD están diferenciadas por los matices de color rojo, mientras que las provincias de menores tasas tienen matices del color amarillo. El mapa de la derecha (ver figura 5.16(b)) muestra los resultados que ha logrado el modelo propuesto, el cual ha considerado la información de los conglomerados a priori y la autocorrelación espacial entre las provincias permitiendo que las estimaciones se ajusten mejor a los valores observados como podemos ver en el mapa.

Por otro lado, tenemos que el gráfico de diagrama de dispersión (ver figura 5.17) se observa las estimaciones de la empresas que incumplieron con sus pagos (empresas con días de atraso mayor a 60) se ajustan a los valores reales (ver figura 5.17(b)), también tenemos las estimaciones para los valores de tasas de incumplimiento (PD), verificando que el modelo propuesto se ajusta a los valores reales de la variable tasa de incumplimiento (PD).
Finalmente, cruzaremos los conglomerados finales de esta segunda etapa con los conglomerados encontrados en parte primera (ver tablas 5.10 y 5.11), esta validación de conglomerados se resumen en un indicador de detección de conglomerados, podemos ver los resultados en el



(a) Efectos aleatorios espaciales de las 196 provincias del Perú, del modelo Poisson - CAR con conglomerados

Figura 5.15: Intervalos de credibilidad al 95 % de los efectos aleatorios espaciales del modelo Poisson - CAR incluyendo conglomerados a priori



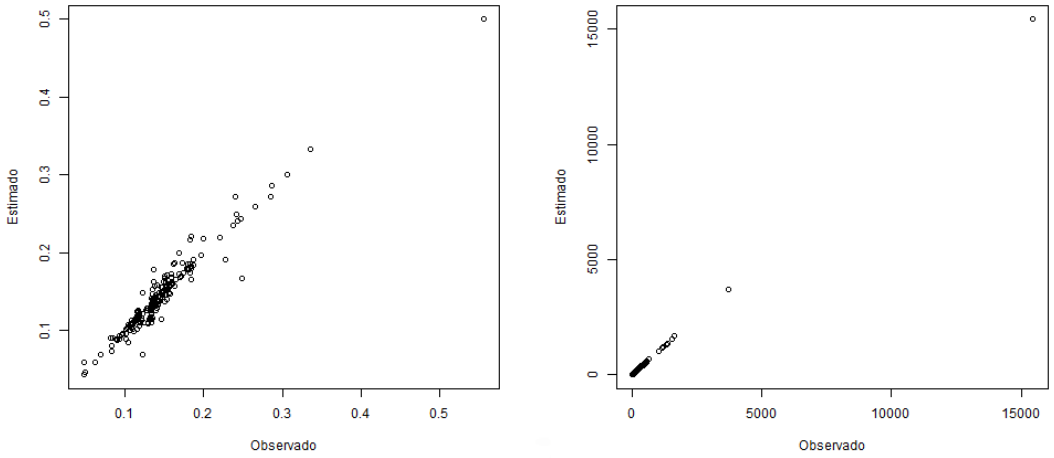
(a) PD - Provincias del Perú

(b) Valores estimados con el modelo Poisson - CAR

Figura 5.16: Mapa de las provincias del Perú - comparativo de los tasas de incumplimiento (PD) reales vs estimadas por el modelo Poisson - CAR incluyendo conglomerados a priori

siguiente cuadro (ver tabla 5.14)

Por otro lado, dado que el mejor modelo es el que incluye conglomerados y nuestro objetivo es poder realizar una mejor gestión y detectar aquellas provincias de alto y bajo riesgo, procedemos a reducir el número de conglomerados encontrado anteriormente. Esta reducción a seis conglomerados se observa en la figura 5.18. Además mostramos la composición de los seis conglomerados finales en la tabla 5.15.



(a) Valores estimados de PD

(b) Valores estimados ♯ de empresas con riesgo de crédito

Figura 5.17: Comparativo de los valores reales vs valores estimados con el modelo Poisson - CAR incluyendo conglomerados a priori

Conglomerado	Precisión
Ward 5 grupos	74 %
Criterio Emp + Ward	77 %

Cuadro 5.14: Tabla resumen de resultados, porcentajes de detección de conglomerados

En conclusión, la propuesta del modelo Poisson-CAR con conglomerados muestra mejores resultados. Este modelo incluye efectos fijos para los conglomerados y efectos aleatorios espaciales. Este modelo ajusta mejor la variable PD en las provincias del Perú. Identificando las provincias con PD de mayor nivel de riesgo (riesgo alto) como Huanca Sancos, Cañete, Condesuyos y Cusco, y las provincias con menor PD (riesgo bajo) como Viru, La Mar, Leoncio Prado y Ucayali.

	Tabla de resultados		
	PD	Composición	Nivel de Riesgo
Conglom. 6	25.1 %	13 %	Riesgo alto
Conglom. 5	16.3 %	18 %	Riesgo alto
Conglom. 4	14.6 %	10 %	Riesgo medio
Conglom. 3	13.7 %	29 %	Riesgo medio
Conglom. 2	11.6 %	17 %	Riesgo bajo
Conglom. 1	8.7 %	13 %	Riesgo bajo

Cuadro 5.15: Descripción de conglomerados finales

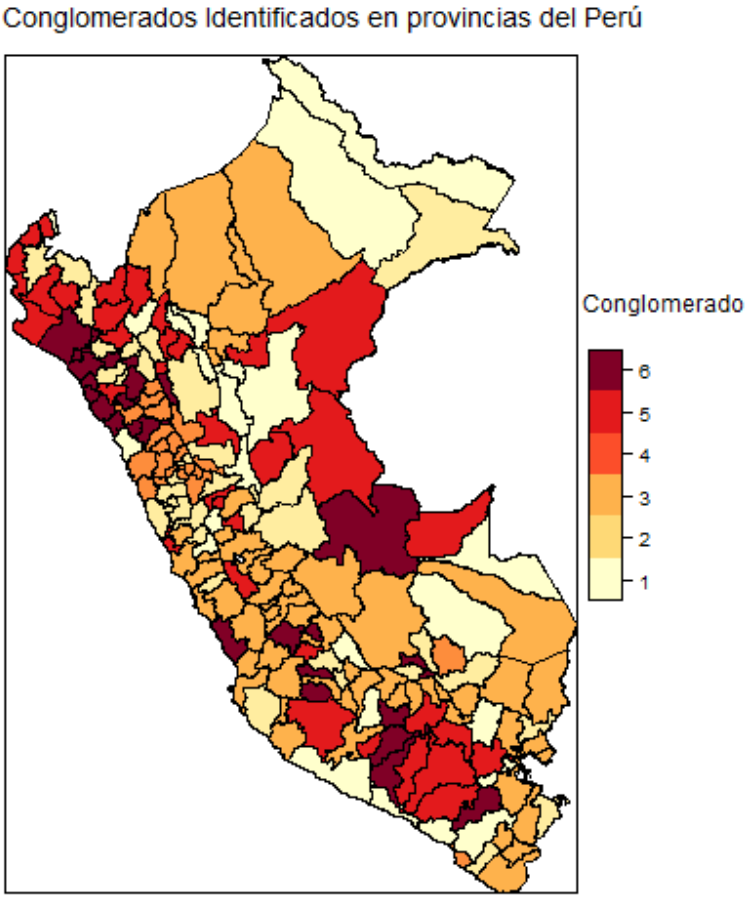


Figura 5.18: Conglomerados identificados para las provincias del Perú, formados por asociación del riesgo de morosidad, donde podemos diferenciar conglomerados de alto y bajo nivel de riesgo, representados por colores rojo y amarillo, respectivamente.

Capítulo 6

Conclusiones

6.1. Conclusiones

- Las aplicación de los modelos propuestos resultaron relevantes en el estudio de la variable respuesta, pues las estimaciones de las tasas de incumplimiento (PD) de las empresas en provincias del Perú se aproximaron a los datos reales. Con ello se pudo detectar en una primera instancia el número óptimo de conglomerados (72 grupos), posteriormente se redujo este número de conglomerados a 6 con el objetivo que permita realizar una mejor gestión y poner foco a provincias con altas tasas de incumplimiento, denominadas provincias con nivel de riesgo alto, y provincias con bajas tasas de incumplimiento, denominadas provincias con nivel de riesgo bajo.
  Por otro lado, es importante mencionar que se obtuvieron buenos resultados con ayuda de las covariables (RUC activo, antigüedad de la empresa, situación habida de la empresa y porcentaje de empresas pequeñas en las provincias del Perú) pues las covariables resultaron relevantes en los modelos propuestos.

- Finalmente, de la aplicación de ambos modelos propuestos, el modelo de Poisson con conglomerados y efectos espaciales se ajustó mejor a la variable respuesta, tasa de incumplimiento (PD).

6.2. Sugerencias para investigaciones futuras

- Se propone como investigación futura la aplicación de estos modelos a una forma más granular, por ejemplo, a nivel distrital teniendo como punto de partida la materialidad de información en cada distrito del Perú.

- También se plantea como investigación futura el uso de modelos estadísticos CAR que toman en consideración la variable tiempo, es decir, modelos estadísticos CAR espacio-temporales que puedan resultar relevantes para las estimaciones de la variable de estudio.

Apéndice A

Resultados teóricos

Derivación del modelo CAR intrínseco

El modelo parte de asumir las distribuciones condicionales:

$$Y_i|Y_j \sim N(\sum_j b_{ij}Y_j, \sigma_i^2),$$

para $j \neq i$ y $i = 1, \ldots, n$. A través del lema de Brook, estas distribuciones condicionales completas son compatibles, es decir la función de densidad conjunta es proporcional a una normal multivariada, tal que:

$$f(y_1, \ldots, y_n) \propto \exp\left(-\frac{1}{2}y^T D^{-1}(I-B)y\right),$$

donde

$$B = \begin{pmatrix} b_{11} & b_{12} & & & \\ b_{21} & b_{22} & b_{23} & & \\ & \ddots & \ddots & \ddots & \\ b_{n1} & & b_{n,n-1} & b_{nn} & \end{pmatrix}$$

$$D = \begin{pmatrix} \sigma_1^2 & 0 & & & \\ 0 & \sigma_2^2 & 0 & & \\ & \ddots & \ddots & \ddots & \\ 0 & & 0 & \sigma_n^2 & \end{pmatrix}.$$

Para que $D^{-1}(I-B)$ sea simétrica se debería cumplir que $\frac{b_{ij}}{w_{i+}} = \frac{b_{ji}}{w_{j+}}$, pero como $\sigma_i^2 \neq \sigma_j^2$ entonces B no es simétrica. Entonces para asegurar la simetría de B se asume que $b_{ij} = \frac{w_{ij}}{w_{i+}}$ y $\sigma_i^2 = \frac{\sigma^2}{w_{i+}}$, donde $w_{i+} = \sum_j w_{ij}$, es decir el número de vecinos de i. De esta forma

$$D^{-1}(I-B) = \begin{pmatrix} \frac{w_{1+}}{\sigma^2} & 0 & & & \\ 0 & \frac{w_{2+}}{\sigma^2} & 0 & & \\ & \ddots & \ddots & \ddots & \\ 0 & & 0 & \frac{w_{n+}}{\sigma^2} & \end{pmatrix} \begin{pmatrix} 1 & -\frac{w_{12}}{w_{1+}} & & & \\ -\frac{w_{21}}{w_{2+}} & 1 & -\frac{w_{23}}{w_{2+}} & & \\ & \ddots & \ddots & \ddots & \\ -\frac{w_{n1}}{w_{n+}} & & & -\frac{w_{n(n-1)}}{w_{n+}} & 1 \end{pmatrix}.$$

$$D^{-1}(I-B) = \begin{pmatrix} \frac{w_{1+}}{\sigma^2} & -\frac{w_{12}}{\sigma^2} & \cdots & & -\frac{w_{1n}}{\sigma^2} \\ -\frac{w_{21}}{\sigma^2} & \frac{w_{2+}}{\sigma^2} & -\frac{w_{23}}{\sigma^2} & \cdots & \\ & & \ddots & \ddots & \ddots \\ -\frac{w_{n1}}{\sigma^2} & & & -\frac{w_{n(n-1)}}{\sigma^2} & \frac{w_{n+}}{\sigma^2} \end{pmatrix}.$$

Luego, $D^{-1}(I-B)$ puede ser reescrita como $\frac{1}{\sigma^2}(W_1 - W)$, donde $W_1 = diag(w_{i+})$. Cabe resaltar que por el Teorema de Hammrsley-Clifford $f(y_1, \ldots, y_n)$ es única, pero impropia debido a la singularidad de $D^{-1}(I-B)$.

Apéndice B

Test Getis-Ord

El test de Getis-Ord mide la concentración de regiones, permitiendo detectar patrones de relaciones espaciales locales entre las regiones y sus vecinos. El funcionamiento de esta herramienta consiste en el estudio de una región alrededor de regiones vecinas, es decir, una región puede tener un valor alto pero es posible que estadísticamente no sea una área significativa. Para ser una región estadísticamemte significativa una región debe tener un valor alto y su vez estar rodeada por otras áreas con valores altos. El estadístico Getis-Ord viene dado por:

$$G = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{[n\sum_{j=1}^{n} w_{ij}^2 - (\sum_{j=1}^{n} w_{ij})]}{n-1}}}, \tag{B.1}$$

donde x_j es valor de estudio de la región j, w_{ij} es el peso espacial entre las regiones i y j, n es el número total de regiones de estudio, además tenemos que:

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2}.$$

Apéndice C

Código de aplicación

```
library(rgdal)
library(RColorBrewer)
library(classInt)
library(readxl")
library("dplyr")
library("stringr")
library(readr")
library("pillar")
library(MASS)
library(mclust)
library(INLABMA)
library(INLA)
library(MCMCpack)
library(clusterSim)
library(tictoc)

    library(raster)
s <- shapefile("C:/Users/Alex/mapa_provincias.shp")
require(spdep)
require(RANN)
```

C.0.1. Matriz adyacente para Perú

```
    nobs <- length(s)
    pe.nb<-poly2nb(s)
    coords <- coordinates(s) plot(s, border="grey") plot(pe.nb, coords, add=TRUE)
    which(card(pe.nb) == 0)

    aux=rep(NA,length(pe.nb))
for(i in 1:length(pe.nb))
aux[i]=length(pe.nb[[i]])

    mn.adj.mat = nb2mat(pe.nb, style="B")
```

Definiendo la matriz de vecindad
W <− mn.adj.mat
W[(W>0)] <− 1
Wstar <− -W
diag(Wstar) <− as.numeric(apply(W, 1, sum))

Definiendo el número de provincias
n = nro regiones
n< −196

Definiendo el número empresas en default para cada provincia
Y.real = prov$CT60_ MALOS

Definiendo el número empresas en cada provincia
E.real = prov$CT60_ TOTAL

phi.real <− log(Y.real/E.real)

Loglinealizando el ratio
logrisk <− log(Y.real/ E.real)

C.0.2. Etapa 1 - Aplicando evaluación de conglomerados

stage1 <− clustering.function(logrisk, W)

C.0.3. Etapa 2 - Seleccionando el mejor modelo usando INLA

C <− diag(apply(W,2,sum)) - W
max.cluster <− 150

C.0.4. Seleccionado las variables explicativas

x <− prov[,c(4:10)]
FLAG_PE <− prov[,c(4)]
HABIDOF <− prov[,c(6)]
ACTIVO <− prov[,c(8)]
ANT_NEGOCIOF <− prov[,c(10)]

Selección del mejor conglomerado
tic()
stage2_s1 <− model.selection_modelo(stage1, Y.real,E.real, C, max.cluster,
FLAG_PE, HABIDOF, ACTIVO, ANT_NEGOCIOF)
toc()

C.0.5. Guardando los criterios de información

DIC

DIC1 <- stage2_s1$model.final$dic$dic

WAIC

WAIC1 <- stage2_s1$model.final$waic$waic

LPML

LPML1 <- stage2_s1$model.final$lp$lp

Cálculo de la tasa de Default

risk <- stage2_s1$model.final$summary.fitted.values[,4]/E.real

C.0.6. Cálculo de RMSE

rmse <- sqrt(mean((risk-exp(phi.real))$\hat{2}$))

C.0.7. Porcentaje de Identificación

Rand <- comparing.Partitions(w4$kf,stage2$factor.clust,type=rand")

Bibliografía

Agresti (2015). Foundations of Linear and Generalized Linear Models , *Wiley Series in Probability and Statistics* .

Anderson, C., Lee, D. y Dean, N. (2014). Identifying clusters in Bayesian disease mapping, *School of Mathematics and Statistics* **15**.

Bayes, T. y Price, R. (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, MA. and F.R.S. , *Philosophical Transactions of the Royal Society of London* pp. 370–418.

Berger, J. (2006). The Case for Objective Bayesian Analysis , *Bayesian analysis* pp. 385–402.

Besag, J. (1974). Spatial and the statistical analysis of lattice systems , pp. 192–225.

Besag, J., York, J. y Mollie, A. (1991). Addendum a Bayesian image restoration, with two applications in spatial statistics , *Annals of the Institute of Statistical Mathematics* **43**: 1–20.

Brooks, S., Gelman, A., Jones, G. L. y Meng, X.-L. (2011). Handbook of Markov Chain Monte Carlo , *Chapman and Hall* .

Charras-Garrido, M., Abrial, D., Goer, J. D. y Dachian (2012). Classification method for disease risk mapping based on discrete hidden markov random fields, *Biostatistics* **13**(2): 241–255.

Cliff, A. D. y Ord, J. K. (1973). Spatial Autocorrelation, *Pion Ltd* .

Diggle, P., Rowlingson, B. y Su, T. (2005). Addendum a Point process methodology for on-line spatio-temporal disease surveillance , *Environmetrics* **16**: 423–434.

Freixas, X., Paya, J. D. H. y Inurrieta, A. (1994). Determinantes macroeconómicos de la morosidad bancaria: un modelo empírico para el caso español , **199**: 125–126.

Gilks, W. R., Richardson, S. y Spiegelhalter, D. J. (1996). Markov Chain Monte Carlo in Practice , *Chapman and Hall* .

Goodhart, C. y Schoenmaker, D. (1993). Institutional separation between supervisory and monetary agencies, *LSE Financial Markets Group, London* .

Hastie, T., Tibshirani, R. y Friedman, J. (2001). The Elements of Statistical Learning, *Springer* **14**.

Kauffman, L. y Rouseew, P. (1990). Finding groups in Data: An Introduction to cluster analysis , *New York: John Wiley* .

Kulldorff, M. (1997). Addendum a A spatial scan statistic. , *Communications in Statistics* **26**: 1481–1496.

Kulldorff, M. (1999). Addendum a Estimation of disease rates in small areas: a newmixed model for spatial dependence por Halloran, M. and Berry , *Chapter Statistical Models in Epidemiology, the Environment and Clinical Trials.* **14**: 135–178.

MacKay, D. J. C. (2003). Information Theory, Inference, and Learning Algorithms, *Cambridge University Press* .

Macnaughton Smith, P., Williams, W., Dale, M. y Mockett, L. (1965). Dissimilarity analysis: a new technique of hierarchical subdivision , *Nature* pp. 1034–1035.

Massart, D., P. F. y Kaufman, L. (1983). Non-hierarchical clustering with MASLOC , *The Journal of the Pattern Recognition Society* pp. 507–516.

Ripley, B. D. (1981). Spatial Statistics , *International Statistical Review / Revue Internationale de Statistique* **14**.

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

Rue, E. A. (2009). INLA: An Introduction , *Norwegian University of Science and Technology* .

Rue, H. y Held, L. (2005). Gaussian Markov Random Fields: Theory and Applications, *Volume 104 of Monographson Statistics and Applied Probability* .

Rue, H., Riebler, A., Sorbye, S. H., Illian, J. B., Simpson, D. P. y Lindgren, F. K. (2017). Bayesian Computing with INLA: A Review , *Annual Review of Statistics and Its Application* pp. 395–421.

Szekely, G. J. y Rizzo, M. L. (2005). Hierarchical clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method , *Journal of Classification* pp. 151–183.