UNIVERSIDAD NACIONAL DE COLOMBIA

Implementación Computacional de Modelos de Procesos Espaciales para Análisis de Redes Sociales
Computational Implementation of Spatial Process Models for Social Network Analysis

Jesús David Solano Velásquez

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Ciudad, Colombia
2022

Implementación Computacional de Modelos de Procesos Espaciales para Análisis de Redes Sociales

Jesús David Solano Velásquez

Trabajo de grado presentado como requisito parcial para optar al título de:
Maestría en Ciencias - Estadística

Director:
Doctor en estadística y matemáticas aplicada Juan Camilo Sosa Martinez

Línea de Investigación:
Análisis de Redes Sociales
Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C., Colombia
Septiembre 1 del 2022

A mis padres por todo su sacrificio y amor.

Agradecimientos

A la Universidad Nacional de Colombia, mi segundo hogar, por disponer de todos los recursos necesarios para hacer de mi formación académica la experiencia más grata y significante de toda mi vida. A mi director Juan Camilo Sosa Martinez, Ph.D. en Estadística y matemática aplicada y profesor asistente de la Universidad Nacional de Colombia, por toda su dedicación, disposición, paciencia y excelentes cualidades pedagógicas. Su aporte para mi formación es inmensurable. A Sandra Pastrán, MSc en estadística y directora del departamento de Estadística en Ipsos Napoleón Franco, nada de esto sería posible sin su apoyo y comprensión. A Juan Camilo Bernal Castillo, candidato a MSc en Estadística, por su amistad incondicional durante más de 6 años. Finalmente, a todos mis profesores, quienes no solo han contribuido a mi formación como estadístico sino también como ser humano.

Resumen

El modelamiento estadístico de las redes permite identificar su distribución de probabilidad, imputar datos faltantes y realizar predicciones sobre la formación de enlaces. Los modelos latentes abordan el modelamiento desde una perspectiva marginal, incorporan dependencias no condicionales por medio de efectos aleatorios. Un caso particular de los modelos latentes es el modelo basado en procesos espaciales completamente Bayesiano que soluciona los problemas de sobreajuste del modelo de espacio latente de distancia. En este documento se realiza la implementación computacional del modelo y se realiza un estudio de sus bondades de ajuste y bondades de predicción a través de redes sintéticas y reales. El modelo tiene buenas cualidades para la replicación de las estadísticas observadas en la red y la estimación de la superficie latente. Sin embargo, el poder predictivo, medido a través del área bajo la curva (AUC por sus siglas en inglés) no supera el valor de 0.7. También se presenta una forma alternativa de ajustar el modelo usando el algoritmo de caso-control. El modelo basado en la log-verosimilitud estimada tiene una buena calidad de bondad de ajuste. **Palabras clave: Cadenas de Markov, Monte Carlo, Bayesiana, Redes, Modelamiento Estadístico**.

Abstract

Statistical modeling of networks makes it possible to identify their probability distribution, impute missing data and make predictions about link formation. Latent models approach modeling from a marginal perspective, incorporating non-conditional dependencies through random effects. A particular case of latent models is the fully Bayesian spatial process-based model that solves the overfitting problems of the latent distance space model. In this paper the computational implementation of the model is performed and a study of its goodness of fit and goodness of prediction through synthetic and real networks is carried out. The model has good qualities for the replication of the statistics observed in the network and the estimation of the latent surface. However, the predictive power, as measured by the area under the curve (AUC), does not exceed 0.7. An alternative way of fitting the model using the case-control algorithm is also presented. The model based on the estimated log-likelihood has a good good goodness-of-fit quality.

Keywords: Markov Chains, Monte Carlo, Bayesian, Networks, Statistical Modelling

Índice general

Agradecimientos vii

Resumen ix

Lista de figuras xiii

Lista de tablas 1

1. Introducción 2

2. Antecedentes 3

2.1. Conceptos básicos . . . 4

2.1.1. Representación de una red . . . 5

2.1.2. Características de los vértices . . . 7

2.1.3. Características de la red . . . 9

2.2. Metodología Bayesiana . . . 10

2.2.1. Estimación . . . 10

2.2.2. Predicción . . . 11

2.2.3. Calidad de Ajuste . . . 11

2.2.4. Validación cruzada . . . 12

3. Metodología 14

3.1. Modelamiento . . . 14

3.1.1. Modelos de Factor Latente . . . 14

4. Estudio de Simulación 21

4.1. Primer Conjunto de Datos Sintético . . . 21

4.1.1. Análisis descriptivo de la red . . . 21

4.1.2. Modelamiento . . . 22

4.1.3. Bondad de Ajuste . . . 22

4.1.4. Predicción . . . 23

4.2. Segundo Conjunto de Datos Simulados . . . 24

4.2.1. Análisis descriptivo de la red . . . 24

4.2.2. Modelamiento . . . 26

4.2.3. Bondad de Ajuste . . . 26
4.2.4. Predicción . . . 27
4.3. Consolidación de Resultados . . . 27

5. Aplicación en Datos Reales 30
5.1. Zachary . . . 30
5.1.1. Análisis Descriptivo de la Red . . . 30
5.1.2. Modelamiento . . . 32
5.1.3. Bondad de Ajuste . . . 32
5.1.4. Predicción . . . 33
5.2. Lazega . . . 34
5.2.1. Análisis Descriptivo de la Red . . . 35
5.2.2. Modelamiento . . . 35
5.2.3. Bondad de Ajuste . . . 37
5.2.4. Predicción . . . 38
5.3. Datos de conflicto internacional . . . 38
5.3.1. Análisis Descriptivo de la Red . . . 40
5.3.2. Modelamiento . . . 43
5.3.3. Bondad de Ajuste . . . 43
5.3.4. Predicción . . . 44
5.4. Datos de Krackhardt . . . 45
5.4.1. Análisis descriptivo de la red . . . 45
5.4.2. Modelamiento . . . 47
5.4.3. Bondad de Ajuste . . . 48
5.4.4. Predicción . . . 48
5.5. Consolidación de los Resultados . . . 49

6. Aproximación Casos y Controles de la Verosimilitud 51
6.1. Cálculo de $n_{i,h}$ . . . 52
6.2. Algoritmo . . . 53
6.3. Estudio de Simulación y Aplicación en Datos de la Literatura . . . 53
6.3.1. Primer Conjunto de Datos Simulados . . . 53
6.3.2. Datos de Zachary . . . 54

7. Discusión 58

A. Anexo: Estudio de convergencia 60

B. Anexo: Algoritmo 69

C. Anexo: Notación 71

Referencias **73**

Lista de Figuras

2.1. A la izquierda la visualización de una red dirigida. A la derecha la visualización de una red no dirigida 4
2.2. Red de los datos de las familias Florentinas 5
2.3. Matriz de píxeles de los datos de las familias Florentinas. 7
2.4. Grado de los vértices de los datos de las familias Florentinas vs el promedio del grado de sus vecinos. 8
2.5. Curva ROC para el modelo de clasificación binaria. 13

3.1. Visualización de un espacio social bidimensional y su red. Vértices con posiciones cercanas en el Espacio Social son proclives a tener una mayor probabilidad de estar relacionados. La función $g(\cdot)$ corresponde a la función de enlace logit. 15
3.2. Representación DAG del modelo de Espacio Latente. 16
3.3. Espacio latente estimado para las familias Florentinas. 17
3.4. Representación DAG del modelo de Proceso Socio-Espacio Latente. 20

4.1. Red de los datos del primer caso de simulación. 21
4.3. Superficie estimada de los datos del primer caso de simulación. 22
4.4. Estadísticas para la evaluación de bondad de ajuste del modelo de los datos del primer caso de simulación. 23
4.5. Curvas ROC del procedimiento de validación cruzada de los datos del primer caso de simulación. 24
4.2. Superficie de los datos del primer caso de simulación. 25
4.6. Red de los datos del segundo caso de simulación. 25
4.7. Superficie de los datos del segundo caso de simulación. 26
4.8. Superficie estimada de los datos del segundo caso de simulación. 27
4.9. Estadísticas para la evaluación de bondad de ajuste del modelo de los datos del segundo caso de simulación. 28
4.10. Curvas ROC del procedimiento de validación cruzada de los datos del segundo caso de simulación. 29

5.1. Red de los datos de Zachary. 30
5.2. Distribución del grado de los vértices de los datos de Zachary. 31

5.3. Partición vía métodos jerárquicos de la red de los datos de Zachary. 31
5.4. Superficie estimada de los datos de Zachary. 32
5.5. Estadísticas para la evaluación de bondad de ajuste del modelo de los datos de Zachary. 33
5.6. Curvas ROC del procedimiento de validación cruzada de los datos de Zachary. 34
5.7. Red de los datos de Lazega. 35
5.8. Partición vía métodos jerárquicos de la red de los datos de Lazega. 36
5.9. Superficie estimada 36
5.10. Estadísticas para la evaluación de bondad de ajuste de los datos de Lazega. 37
5.11. Curvas ROC del procedimiento de validación cruzada de los datos de Lazega. 38
5.12. Histograma del logaritmo de la población de los datos de conflicto. 39
5.13. Histograma del puntaje político por nación de los datos de conflicto. 40
5.14. Red de los datos de conflicto. 40
5.15. Distribución del grado de los vértices de los datos de conflicto. 41
5.16. Grado de los vértices de los datos de conflicto vs el promedio del grado de sus vecinos. 41
5.17. Gran componente de la red de los datos de conflicto. 42
5.18. Partición vía métodos jerárquicos de la gran componente de la red de los datos de conflicto. 42
5.19. Superficie estimada de los datos de conflicto. 43
5.20. Estadísticas para la evaluación de bondad de ajuste del modelo de los datos de conflicto. 44
5.21. Curvas ROC del procedimiento de validación cruzada de los datos de conflicto 45
5.22. Red de los datos de Krackhardt. 46
5.23. Grado de los vértices de los datos de Krackhardt vs el promedio del grado de sus vecinos. 46
5.24. Partición vía métodos jerárquicos de la red de los datos de Krackhardt. . . 47
5.25. Superficie estimada de los datos de Krackhardt. 47
5.26. Estadísticas para la evaluación de bondad de ajuste del modelo de los datos Krackhardt. 48
5.27. Curvas ROC del procedimiento de validación cruzada de los datos de Krackhardt. 49

6.1. Superficie estimada de los datos del primer caso de simulación para el modelo inspirado en la metodología de casos y controles. 54
6.2. Estadísticas para la evaluación de bondad de ajuste del modelo usando la metodología de casos y controles para los datos del primer caso de simulación. 55
6.3. Superficie estimada de los datos Zachary para el modelo inspirado en la metodología de casos y controles. 56

6.4. Estadísticas para la evaluación de bondad de ajuste del modelo usando la metodología de casos y controles para los datos de Zachary. 57

A.1. Convergencia del algoritmo MCMC de los datos del primer caso de simulación. 61
A.2. Convergencia del algoritmo MCMC de los datos del segundo caso de simulación. 62
A.3. Convergencia del algoritmo MCMC de los datos de Zachary. 63
A.4. Convergencia del algoritmo MCMC de los datos de Lazega. 64
A.5. Convergencia del algoritmo MCMC de los datos de conflicto. 65
A.6. Convergencia del algoritmo MCMC de los datos de Krackhardt. 66
A.7. Convergencia del algoritmo MCMC modificado de los datos del primer conjunto de simulación. 67
A.8. Convergencia del algoritmo MCMC modificado de los datos de Zachary. . . 68

Lista de Tablas

4.1. AUC estimados para los modelos de datos sintéticos. 27

5.1. AUC estimados para los modelos de datos reales. 50

1. Introducción

La estructura de contacto derivada de una interacción definida entre un conjunto de individuos se denomina red. Estudiar los fenómenos vistos como una red de contactos es clave para diversas áreas del conocimiento. En el contexto epidemiológico permite tomar acciones tanto preventivas como correctivas para afrontar el brote de alguna enfermedad, en marketing para determinar estrategias de venta conociendo los patrones de compra con base en la navegación web de las personas, en biología permite evidenciar patrones en la interacción entre aves y plantas para estudiar los comportamientos de polinización y/o dispersión de semillas, solo por mencionar algunos ejemplos.

La estadística es determinante para el entendimiento, explicación y predicción de estas estructuras, y los últimos desarrollos tecnológicos hacen indispensable la computación para hacer estadística. Aunque la computación es un facilitador para el desarrollo de los estudios, no deja de ser un desafío traducir la teoría para una correcta ejecución por la máquina.

Este trabajo desarrolla la implementación computacional usando código R y Rcpp de un modelo de espacio latente, basado en procesos espaciales completamente Bayesiano a partir del trabajo preliminar de Linkletter (2007). Este modelo se propone para solucionar los problemas de sobreajuste del modelo de espacio latente de distancia (e.g., Hoff et al. 2002). Se ejemplifican las medidas para el análisis descriptivo de redes a través de un conjunto de datos de juguete. Se realiza un resumen de los conceptos Bayesianos necesarios para el estudio. Se hace el desarrollo del modelo, una explicación detallada del algoritmo MCMC y un estudio exhaustivo de las bondades predictivas y de bondad ajuste, apoyado en el análisis topológico de las redes y las respectivas probabilidades de interacción, por medio de redes tanto reales como sintéticas. Los desarrollos computacionales se ponen a disposición del lector a través de un repositorio creado en GitHub, el cual se encuentra referenciado en el capitulo 3. Finalmente, se implementa una forma alternativa de ajustar el modelo usando el algoritmo de caso-control inspirado en el trabajo de Raftery et al. (2012) y se realiza una evaluación de las bondades de calidad de ajuste.

2. Antecedentes

Un modelo estadístico es una colección de distribuciones de probabilidad indexadas por un vector de parámetros desconocido $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. El modelamiento estadístico de las redes permite identificar su distribución de probabilidad, imputar datos faltantes y realizar predicciones sobre la formación de enlaces. Los métodos tradicionales no son útiles para el modelamiento de redes debido a las estructuras de dependencia entre las observaciones, se han presentado alternativas que tengan en cuenta esta característica.

Entre las primeras propuestas se encuentran los Modelos de Grafos Aleatorios (*Random Graph Models*, Gilbert 1959) y la familia de Modelos de Grafos Aleatorios Generalizados (*Generalized Random Graph Models*). Los Modelos de Configuración (*Configuration Models*, Bender and Canfield 1978), los Modelos de Mundo Pequeño (*Small-World Models*, Watts and Strogatz 1998) y los Modelos de Fijación Preferencial (*Preferential Attachment Models* Barabási and Albert 1999) son algunos ejemplos de Modelos de Grafos Aleatorios Generalizados.

Los Modelos Exponenciales de Grafos Aleatorios (e.g., Frank and Strauss 1986), ERGMs (por sus siglas en inglés), se proponen como una extensión directa de los modelos lineales generalizados clásicos. Frank and Strauss (1986) proponen modelos con una estructura de Markov que provee formas de dependencia entre los enlaces. Los ERGMs permiten modelar las redes en función tanto de variables endógenas (características estructurales de la red como la transitividad, por ejemplo) como exógenas (características propias del sistema como los atributos de los vértices, por ejemplo).

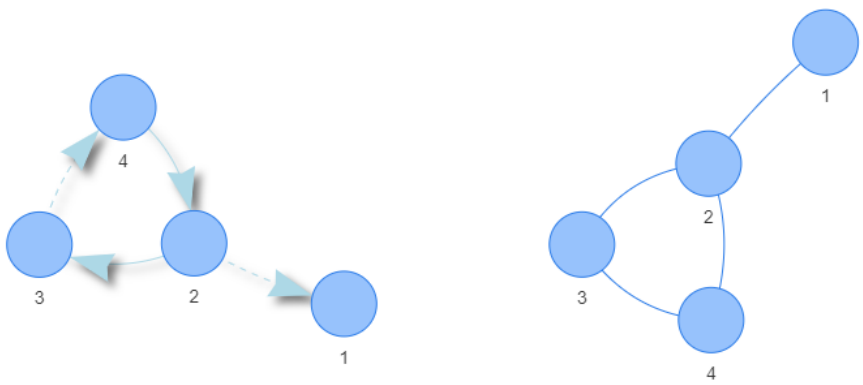
Los modelos latentes (Hoff et al. 2002) son una alternativa a los ERGMs, abordan el modelamiento desde una perspectiva marginal, incorporan dependencias no condicionales por medio de efectos aleatorios. Los modelos de espacio latente asumen que cada individuo tiene una posición no observada (latente) en un espacio euclidiano d-dimensional (denominado Espacio Social). Otros modelos de espacio latente populares son los Modelos Factoriales (Hoff 2009) y los Modelos Bilineales (e.g., Hoff 2005).

Debido a su flexibilidad e interpretabilidad, los modelos de espacio latente han sido empleados y extendidos en una gran variedad de aplicaciones. Entre otras aplicaciones se distinguen: Configuración en Redes Sociales: Un Modelo de Medida (e.g., Schweinberger and Snijders 2003), Modelos para Redes Múltiples, (e.g., Salter-Townshend and McCormick 2017, Durante et al. 2018, Wang et al. 2019, Sosa and Betancourt 2022), Modelos para Estructuras Sociales Cognitivas (e.g., Sewell 2019, Sosa and Rodríguez 2021), y Modelos

para Redes Dinámicas (e.g., Sewell and Chen 2015; Kim 2018). Para una revisión de la literatura más exhaustiva se recomienda al lector Sosa and Buitrago (2021).
Desde el punto de vista computacional, en la literatura estadística hay disponibles algunas iniciativas para el ajuste de los modelos de redes. En particular, en R, la librería `ergm` presenta un conjunto de funciones para el modelamiento de los ERGMs y la librería `latentnet` (e.g., Handcock and Krivitsky 2008) implementa el ajuste de modelos de espacio latente. En el libro *Statistical Analysis of Network Data with* `R` (Kolaczyk and Csárdi 2020) se pueden estudiar ejemplos de uso.

2.1. Conceptos básicos

Una red se define como un objeto matemático notado por $G = \{E, V\}$ con E un conjunto de enlaces y V el conjunto de vértices. El tamaño de la red, corresponde al número de enlaces que lo componen y el orden al número de vértices.
Un enlace es una representación de una relación binaria entre dos pares de vértices i y j, notado por $e = (i, j)$. Una red se denomina **no dirigida** si y solo si $(i, j) = (j, i)$, caso contrario, la red se llama **dirigida**. La Figura 2.1 ilustra esta definición.



(a) Red dirigida (b) Red no dirigida

Figura 2.1.: A la izquierda la visualización de una red dirigida. A la derecha la visualización de una red no dirigida

En las redes dirigidas la relación no es simétrica. Considere la relación *"ser cliente"*, dados dos pares de empresas i y j, si j es cliente de i habrá un enlace hacía i proveniente de j. Nótese que si j *"es cliente de"* i no implica que i *"sea cliente de"* j.
Cuando el sentido de la relación es simétrico, e.g., la relación *"trabaja con"*, se establece un enlace entre los respectivos pares de vértices tal como se muestra en el panel b de la Figura 2.1.

Se presentan algunas medidas descriptivas en el análisis de redes no dirigidas. La extensión a redes dirigidas, es inmediata. Para ejemplificar el uso de estas medidas, se utiliza el conjunto de datos correspondiente a los matrimonios que se produjeron entre 16 familias Florentinas en el Siglo L (Padgett and Ansell 1993).

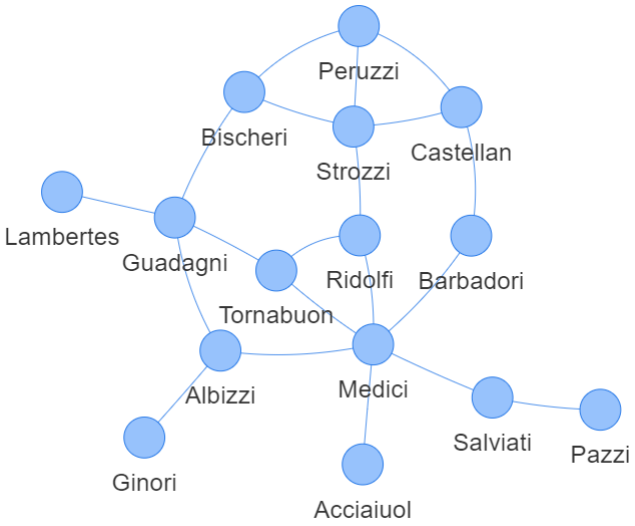


Figura 2.2.: Red de los datos de las familias Florentinas

2.1.1. Representación de una red

A continuación se presentan algunas formas de representación alternativas de una red diferentes a la diagramación de la red con vértices y enlaces.

Matriz de Adyacencia

Se define la **matriz de adyacencia** de una red como la matriz cuadrada $\mathbf{Y} = [y_{i,j}]$ que representa sus relaciones. Se define la entrada $y_{i,j}$ como

$$y_{i,j} = \begin{cases} 1 & \text{si } (i,j) \in E \\ 0 & \text{en otro caso} \end{cases}$$

Para redes no dirigidas la matriz de adyacencia es simétrica. Este objeto es fundamental para el modelamiento, dado que se busca asociar a la matriz de adyacencia una distribución de probabilidad $p(\mathbf{Y} \mid \boldsymbol{\theta})$ indexada por un parámetro desconocido $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Por ejemplo, la matriz de adyacencia para el conjunto de datos de la familia Florentina es:

$$\mathbf{Y} = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0
\end{pmatrix}$$

Nótese que en la diagonal principal todas las entradas corresponden a 0, los cuales se denominan ceros estructurales, dado que las redes que se están considerando no admiten relaciones reflexivas. La cantidad de entradas diferentes en una red no dirigida es $n(n-1)/2$ mientras que en una red dirigida es $n(n-1)$.

Finalmente, para las redes no dirigidas no es necesario almacenar computacionalmente toda la matriz dado que la porción superior de la matriz es igual a la porción inferior, es decir, $y_{i,j} = y_{j,i}$ para $1 \leq i < j \leq n$.

Matriz de píxeles

La **matriz de píxeles** es una alternativa gráfica para la matriz de adyacencia donde la relación entre el vértice i y el vértice j es representada como una celda de color. Siguiendo con la red de familias Florentinas se presenta la Figura 2.3.

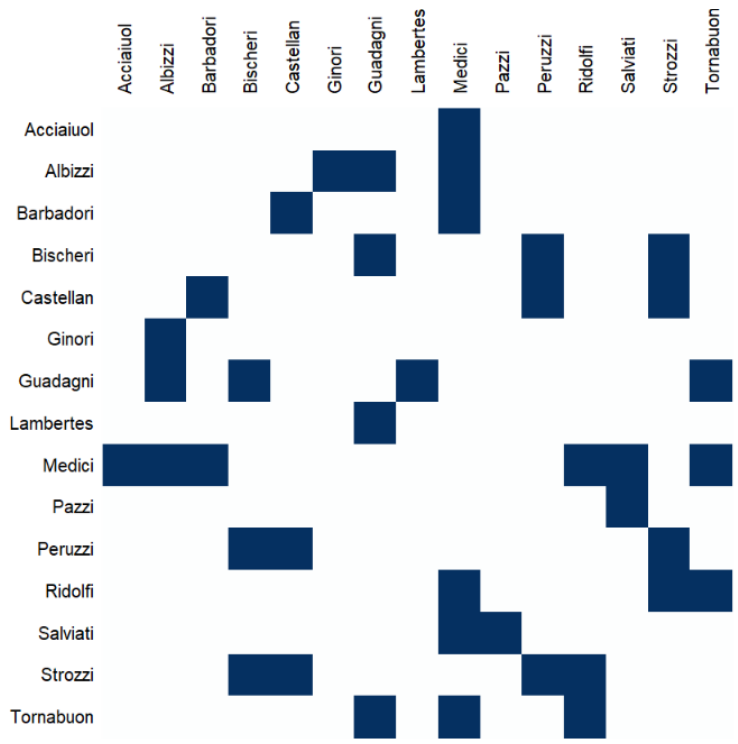


Figura 2.3.: Matriz de píxeles de los datos de las familias Florentinas.

Las relaciones están representadas como los píxeles en azul, los píxeles en blanco indican la ausencia de relación entre un par de vértices. Por ejemplo, de la matriz de píxeles en la Figura 2.3 se puede concluir que existe una relación entre Albizi y Ginori, pero no hay relación de Albizi con Acciaiuol.

2.1.2. Características de los vértices

Las medidas de centralidad determinan la importancia de un vértice. En este estudio las medidas de centralidad implementadas son: el **grado**, la **centralidad por cercanía** y la **centralidad por intermediación**.
Se define $j \in V$ como un **vecino** de $i \in V$ si $(i, j) \in E$. El grado de un vértice i es el número de vecinos que tiene. Un **camino** es una secuencia continua de vértices que permite conectar dos pares de vértices en la red. La **distancia geodésica** es la longitud del camino más corto entre dos vértices.

Centralidad por cercanía

La centralidad por cercanía del vértice $i \in V$ es una función de la sumatoria de las distancias del vértice con respecto al total de vértices. Indica la importancia del vértice i en función de su cercanía a los demás vértices y se define como:

$$C_{\mathsf{C}}(i) = \frac{1}{\sum_{j \in V} \mathsf{dist}(j,i)}$$

donde $\mathsf{dist}(j,i)$ la distancia geodésica entre los vértices $\{i,j\} \in V$. Esta medida es normalizada para la comparación entre redes.

Centralidad por intermediación

La importancia de intermediación del vértice $i \in V$ esta basada en la posición del vértice entre otros pares de vértices y se define como:

$$C_{\mathsf{B}}(i) = \sum_{k \neq l \neq i \in V} \frac{\sigma(k,l \mid i)}{\sigma(k,l)}$$

donde $\sigma(k,l \mid i)$ el número de caminos más cortos entre k y l que pasan a través de i y $\sigma(k,l)$ el número de caminos más cortos entre k y l en la red.

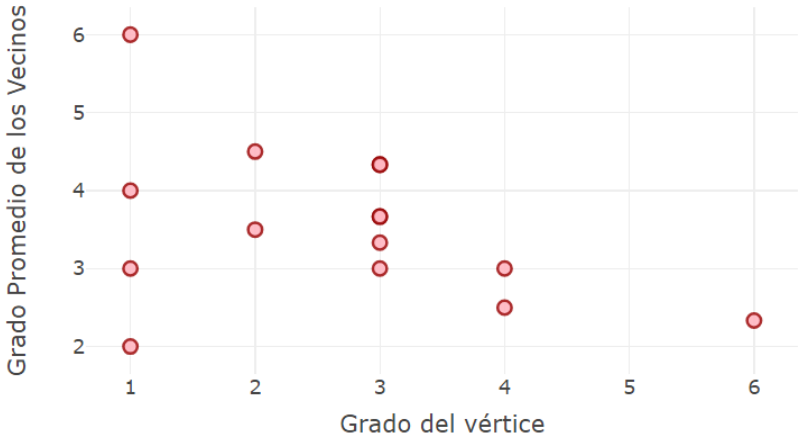


Figura 2.4.: Grado de los vértices de los datos de las familias Florentinas vs el promedio del grado de sus vecinos.

La Figura 2.4 muestra que vértices de grado alto se relacionan con vértices de grado bajo. Vértices con grados bajos se relacionan de manera indistinta con vértices de grados altos y bajos. Las familias más centrales por cercanía son Medici, Ridolfi y Albizzi. Las familias más centrales por intermediación son Medici, Guadagni y Albizzi. Finalmente, las familias con poder adquisitivo alto se relacionan con familias de poder adquisitivo bajo. Si el lector esta interesado en más detalles de aplicación se recomienda dirigirse a Kolaczyk and Csárdi (2020).

2.1.3. Características de la red

El estudio de la **cohesión** mide como los vértices (o subconjuntos de vértices) se mantienen juntos de acuerdo con la relación que define los enlaces en la red.
La **conectividad** de una red se estudia a través de su densidad o evaluando su resiliencia, i.e., el impacto sobre el número de enlaces en la red al remover un vértice.
Las medidas de **asortatividad** son variaciones de un coeficiente de correlación. Determinan la manera en que se agrupan los vértices con características similares (i.e., homofilia).

Densidad

"La densidad de una red corresponde a la frecuencia de las relaciones observadas con respecto a las relaciones potenciales" (Kolaczyk and Csárdi 2020, p.55) y se define como:

$$\mathsf{d}(G) = \frac{\mid E \mid}{\mid V \mid (\mid V \mid -1)/2}$$

esta medida tiene rango en el intervalo $[0, 1]$ y aplica para redes sin bucles y simples. Valores cercanos a 1 indican que la red es propensa a ser un clique, i.e., para un par de vértices $\{i, j\} \in V$ siempre existe un enlace $(i, j) \in E$.

Coeficiente de transitividad

El coeficiente de transitividad permite estudiar el nivel de agrupamiento en los datos. Se puede interpretar como la probabilidad de que haya un enlace entre dos vértices adyacentes ($i \in V$ y $j \in V$ son adyacentes si existe un enlace entre ellos) y se define como:

$$\mathsf{c}(G) = \frac{3\tau_{\triangle}(G)}{\tau_3(G)}$$

donde $\tau_{\triangle}(G)$ es el número de triángulos (cliques de tres vértices) en G y $\tau_3(G)$ el número de triplas conectadas en G (i.e., sub-red de tres vértices con dos enlaces. En general una sub-red G_s de G esta conectada si para cualquier par de vértices $\{i, j\} \in V_s$ existe un camino que los conecta).

Asortatividad

Asume valores en $[-1, 1]$. Valores cercanos a 1 indican que vértices de características similares se agrupan entre sí. Valores cercanos a 0 indican que los enlaces se crean de manera aleatoria. Para una característica cuantitativa el coeficiente de asortatividad es:

$$r = \frac{\sum_{x,y} f_{xy} - f_{x+} f_{+y}}{\sigma_x \sigma_y}$$

Continuando con el ejemplo de las familias Florentinas, la densidad es 0.1905, i.e., se presentó el 19.05 % de todos los enlaces posibles en la red. La transitividad es 0.1915, i.e., la probabilidad de que haya una enlace entre un par de vértices adyacentes es baja. Finalmente, la asortatividad es -0.3024 tomando el poder adquisitivo de cada familia como característica cuantitativa de interés.

2.2. Metodología Bayesiana

Sea $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta$ un vector de parámetros que determina la estructura estocástica del vector de observaciones $\boldsymbol{y} = (y_1, \cdots, y_n)$. Siguiendo el teorema de Bayes se tiene:

$$\mathsf{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{\mathsf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})\mathsf{p}(\boldsymbol{\theta})}{\int \mathsf{p}(\boldsymbol{\theta}^*)\mathsf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}^*)\mathsf{d}\boldsymbol{\theta}^*}$$

donde $\mathsf{p}(\boldsymbol{y} \mid \boldsymbol{\theta})$ es la función de verosimilitud de $\boldsymbol{y}$ dado $\boldsymbol{\theta}$, $\mathsf{p}(\boldsymbol{\theta})$ corresponde al conocimiento previo sobre el comportamiento de los parámetros y recibe el nombre de distribución previa de $\boldsymbol{\theta}$, $\mathsf{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$ representa el conocimiento sobre el vector de parámetros actualizado por la información muestral y recibe el nombre de distribución posterior de $\boldsymbol{\theta}$.

2.2.1. Estimación

La estimación puntual de un parámetro o una función de un parámetro de interés se puede obtener mediante la media posterior de $\boldsymbol{\theta}$. Por ejemplo, un estimación puntual de una función $g(\boldsymbol{\theta})$ se define como:

$$\mathsf{E}(g(\boldsymbol{\theta} \mid \boldsymbol{y})) = \int_{\Theta} g(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathsf{d}\boldsymbol{\theta} \tag{2.1}$$

En la práctica se emplean métodos de Cadenas de Monte Carlo para aproximar empíricamente el estimador puntual presentado en la Ecuación (2.1). Por ejemplo, para $\boldsymbol{\theta}_1^{(1)}, \ldots, \boldsymbol{\theta}_1^{(B)}$ muestreados de la distribución posterior $\mathsf{p}(\boldsymbol{\theta}|\boldsymbol{y})$, la aproximación de la media posterior en la Ecuación (2.1) se define como:

$$\mathsf{E}(g(\boldsymbol{\theta} \mid \boldsymbol{y})) \approx \frac{1}{B} \sum_{b=1}^{B} g(\boldsymbol{\theta}^{(b)})$$

Realizaciones de la distribución posterior se utilizan para hacer inferencias sobre $\boldsymbol{\theta}$. Cuando la distribución posterior $\mathsf{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$ no sigue una forma estándar se emplean otro tipo de herramientas para obtener las realizaciones, por ejemplo, los algoritmos de Cadenas de Markov Monte Carlo (MCMC, por sus siglas en inglés).
El algoritmo de Gibbs es un método MCMC ampliamente utilizado en la práctica. Se define $\boldsymbol{\theta}_{-\ell}^{(b)}$ como:

$$\boldsymbol{\theta}_{-\ell}^{(b)} = (\theta_1^{(b)}, \ldots, \theta_{\ell-1}^{(b)}, \theta_{\ell+1}^{(b-1)}, \ldots, \theta_p^{(b-1)})$$

dado un vector de valores iniciales y $\mathsf{p}(\theta_\ell \mid \boldsymbol{\theta}_{-\ell}^{(b)}, \boldsymbol{y})$ con forma estándar, se pueden obtener realizaciones de la distribución posterior iterando cada componente de $\boldsymbol{\theta}$ de su distribución condicional completa. Para para $b = 1, \ldots, B$ se muestrea $\theta_\ell^{(b)}$ con base en la siguiente expresión

$$\theta_\ell^{(b)} \sim \mathsf{p}(\theta_\ell \mid \boldsymbol{\theta}_{-\ell}^{(b)}, \boldsymbol{y})$$

El algoritmo de Metropolis-Hastings (e.g., Chib and Greenberg 1995) se utiliza cuando no es posible simular directamente de una distribución probabilística asociada con θ_ℓ (e.g., la distribución condicional completa de θ_ℓ).
Considerando $\mathsf{q}(\cdot \mid \cdot)$ una función de distribución de probabilidad simétrica, se propone un nuevo valor θ_ℓ^* para θ_ℓ obtenido a partir de $\mathsf{q}(\theta_\ell^* \mid \theta_\ell^{(b-1)})$. El criterio de aceptación esta dado por:

$$r(\theta_\ell) = \frac{\mathsf{p}(\theta_\ell^* \mid \boldsymbol{\theta}_{-\ell}^{(b)}, \boldsymbol{y})\mathsf{q}(\theta_\ell^* \mid \theta_\ell^{(b-1)})}{\mathsf{p}(\theta_\ell^{(b-1)} \mid \boldsymbol{\theta}_{-\ell}^{(b)}, \boldsymbol{y})\mathsf{q}(\theta_\ell^{(b-1)} \mid \theta_\ell^*)} \tag{2.2}$$

Con base en la Ecuación (2.2), y dado un valor $u \sim \mathsf{U}(0, 1)$ se establece,

$$\theta_\ell^{(b)} = \begin{cases} \theta_\ell^* \text{ si } \; u < r(\theta_\ell) \\ \theta_\ell^{(b-1)} \text{en otro caso.} \end{cases}$$

2.2.2. Predicción

Para un nuevo vector de covariables $\boldsymbol{x}^*$, la distribución de predicción $\mathsf{p}(y^* \mid \boldsymbol{y})$ se define como en la Ecuación (2.3). En la práctica, estimaciones puntuales, e.g., $\mathsf{E}(y^* \mid \boldsymbol{y})$ (o por intervalo) se obtienen con base a las realizaciones de la distribución posterior reemplazando el **vector de nuevas covariables** $\boldsymbol{x}^*$ en cada una de las iteraciones del algoritmo MCMC del proceso de estimación.

$$\mathsf{p}(y^* \mid \boldsymbol{y}) = \int_\Theta \mathsf{p}(y^* \mid \boldsymbol{\theta})\mathsf{p}(\boldsymbol{\theta} \mid \boldsymbol{y})\mathsf{d}\boldsymbol{\theta} \tag{2.3}$$

2.2.3. Calidad de Ajuste

La distribución de predicción en la Ecuación (2.3) evaluada en las covariables originales permite obtener realizaciones $\tilde{\boldsymbol{y}}$ y su distribución posterior se define como:

$$\mathsf{p}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}) = \int_{\Theta} \mathsf{p}(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}) \mathsf{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathsf{d}\boldsymbol{\theta}$$

Para la evaluación de la calidad de ajuste se estima la distribución de una estadística de interés de la red (i.e., $T(\boldsymbol{y})$) a partir de las realizaciones obtenidas de reemplazar el vector de covariables observado en cada una de las iteraciones del algoritmo MCMC en el proceso de estimación. Si la estadística observada $T(\boldsymbol{y})$ es un valor atípico en relación con la distribución estimada se concluye que el modelo no reproduce adecuadamente el parámetro.

2.2.4. Validación cruzada

Considere L particiones de igual tamaño obtenidas de una permutación de los $\frac{n(n+1)}{2}$ elementos únicos fuera de la diagonal de la matriz de adyacencia $\mathbf{Y}$. Los elementos de la partición i, con $i \in \{1, \ldots, L\}$, se asumen como valores no observados. Dado un vector de valores iniciales para el vector de parámetros $\boldsymbol{\theta}$ el algoritmo de validación cruzada se define como:

1. Dado $\boldsymbol{\theta}^{(b)}$ imputar los valores faltantes a partir de la distribución muestral. $\mathsf{p}(\boldsymbol{y} \mid \boldsymbol{\theta}^{(b)})$ y obtener $\hat{\boldsymbol{y}}^{(b)}$.

2. Actualizar $\boldsymbol{\theta}^{(b+1)}$ a través del algoritmo MCMC.

3. Repetir los pasos $1 - 2$ hasta que se complete un número de iteraciones B.

4. Calcular la proporción $prob_k$ de unos estimados para la observación faltante k de la partición i, con $k \in 1, \ldots, \frac{n(n+1)}{2L}$

5. Determinar el auc_i con base en el vector $(prob_1, \ldots, prob_{ni})$ y los valores observados.

La estimación puntual del AUC del modelo se define como:

$$\hat{\text{AUC}} = \sum_{i}^{L} \frac{\text{auc}_i}{L}$$

El AUC asume valores entre [0, 1]. Valores de AUC cercanos a 1 indican que el modelo tiene un poder predictivo alto. Valores de AUC cercanos a 0.5 indican que el modelo no tiene capacidad de distinguir entre clases. Valores cercanos a 0 indican que el modelo predice las categorías de manera inversa.
Para definir el área bajo la curva (AUC, por sus siglas en inglés), necesariamente hay que hablar de la curva ROC (Receiver Characteristic Operator curve) porque es sobre esa curva sobre la que se calcula el área.

La curva de probabilidades ROC gráfica la tasa de verdaderos positivos (TPR) contra la tasa de falsos positivos (FPR) para diferentes puntos de corte para las probabilidades estimadas que permiten distinguir entre clases en un modelo de respuesta binaria. En este estudio se discrimina entre la presencia o ausencia de un enlace. El AUC es usado como una medida de resumen de la curva ROC e indica qué tan bueno es el modelo para distinguir entre clases. De manera ilustrativa, suponga un modelo de clasificación binaria con curva ROC como en la Figura 2.5,

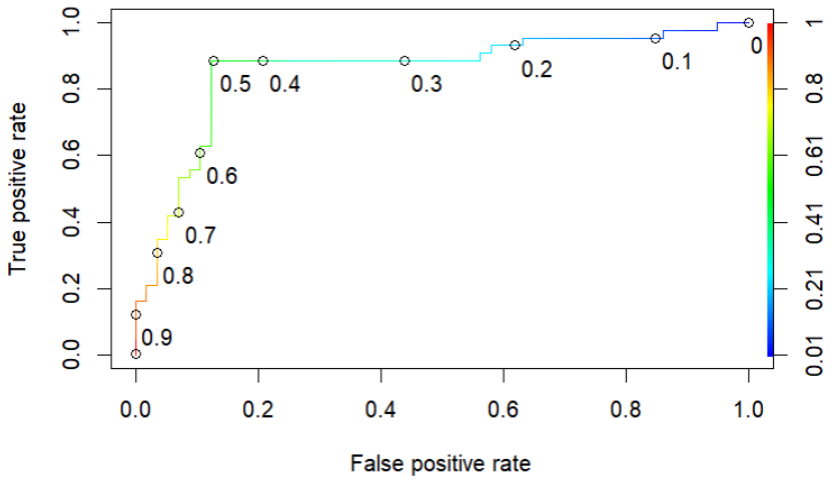


Figura 2.5.: Curva ROC para el modelo de clasificación binaria.

El AUC estimado para este modelo es de 0.86. Para un modelo con ajuste perfecto se tendrá un TPR igual a 1 cuando FPR es igual a 0 para todos los puntos de corte, en ese caso la curva ROC es un cuadrado de lado 1. Para más detalles se recomienda al lector dirigirse a Gonçalves et al. (2014).

3. Metodología

3.1. Modelamiento

El modelamiento estadístico de redes permite asociar una distribución de probabilidad indexada por un vector de parámetros desconocido $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ a la matriz de adyacencia $\mathbf{Y}$, lo que posibilita capturar patrones importantes de la red como la transitividad y la homofilia. Dado que el supuesto de independencia entre las observaciones no se cumple, técnicas tradicionales de modelamiento como los modelos lineales generalizados no son adecuadas. Por tal motivo, entre otros, se proponen los modelos ERGMs y los modelos de factor latente con el fin de incluir la estructura de dependencia entre las observaciones en el modelamiento.

3.1.1. Modelos de Factor Latente

Asumiendo un total de n individuos, los modelos de factor latente modelan las probabilidades de interacción $\pi_{i,j} = \mathsf{Pr}(y_{i,j} = 1 \mid \beta_0, \boldsymbol{\beta}, \zeta_{i,j})$, con $i < j$ y $i, j = 1, \ldots, n$. Se utiliza solo una parte de la matriz de adyacencia ($i < j$) ya que para este trabajo solo es de interés el estudio de redes no dirigidas. En caso contrario se deberá utilizar toda la matriz ($i \neq j$) exceptuando la diagonal principal (porque no se admiten relaciones reflexivas dado que no se está estudiando los multigrafos). La distribución muestral se nota como:

$$\mathsf{p}(\boldsymbol{y} \mid \boldsymbol{\pi}) \ = \prod_{i<j} \pi_{i,j}^{y_{i,j}} (1 - \pi_{i,j})^{1-y_{i,j}} \tag{3.1}$$

donde $\eta_{i,j} = \mathsf{logit}(\pi_{i,j}) = \beta_0 + \boldsymbol{x}_{i,j}\boldsymbol{\beta}^T + \zeta_{i,j}$ corresponde al predictor lineal, $\zeta_{i,j}$ es el error y captura las dependencias entre los $y_{i,j}$, $\boldsymbol{x}_{i,j} = (\mid x_{i,1} - x_{j,1} \mid, \ldots, \mid x_{i,p} - x_{j,p} \mid)$, $\boldsymbol{\pi} = \{\pi_{i,j}\}_{i<j=1}^{n}$ y $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ un vector de parámetros de tamaño p.
La función logit en la Ecuación (3.2) es la función de enlace como en un modelo lineal generalizado clásico. Si el analista lo prefiere también podría utilizar otra función de enlace como la función probit.

$$\mathsf{logit}(\pi_{i,j}) = \mathsf{log}(\frac{\pi_{i,j}}{1 - \pi_{i,j}}) \tag{3.2}$$

La forma de distribución muestral indica que los $y_{i,j}$ son condicionalmente independientes dadas las probabilidades de interacción $\pi_{i,j}$. Sin embargo, no implica que los $y_{i,j}$ para

$i, j = 1, \ldots, n$ sean independientes marginalmente (sin condicionar) como lo asumen los $GLMs$ clásicos.

Modelos de Espacio Latente

Con el fin de capturar información no contenida en la formulación de la Ecuación (3.1) Hoff et al. 2002 postula la existencia de un espacio *euclidiano* social d-dimensional denominado Espacio Social (usualmente se considera $d = 2$ para que la visualización y caracterización del sistema relacional sea sencillo de interpretar, aunque espacios dimensionales superiores también son posibles), cuyas características latentes (no observadas) son notadas por $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,d}) \in \mathbf{R}^d$, para $i = 1, \ldots, n$. El modelo de factor latente se refórmula con

$$\zeta_{i,j} = - \mid\mid \boldsymbol{z}_i - \boldsymbol{z}_j \mid\mid \tag{3.3}$$

donde $\mid\mid \cdot \mid\mid$ es la norma euclidiana. Valores altos implican que los dos vértices se encuentran distantes en el Espacio Social. El impacto sobre la probabilidad de que se presente un enlace entre dos vértices $i, j \in V$ es incluido a través del signo menos en frente de la norma de la Ecuación (3.3). Para ilustrar se presenta la Figura 3.1.

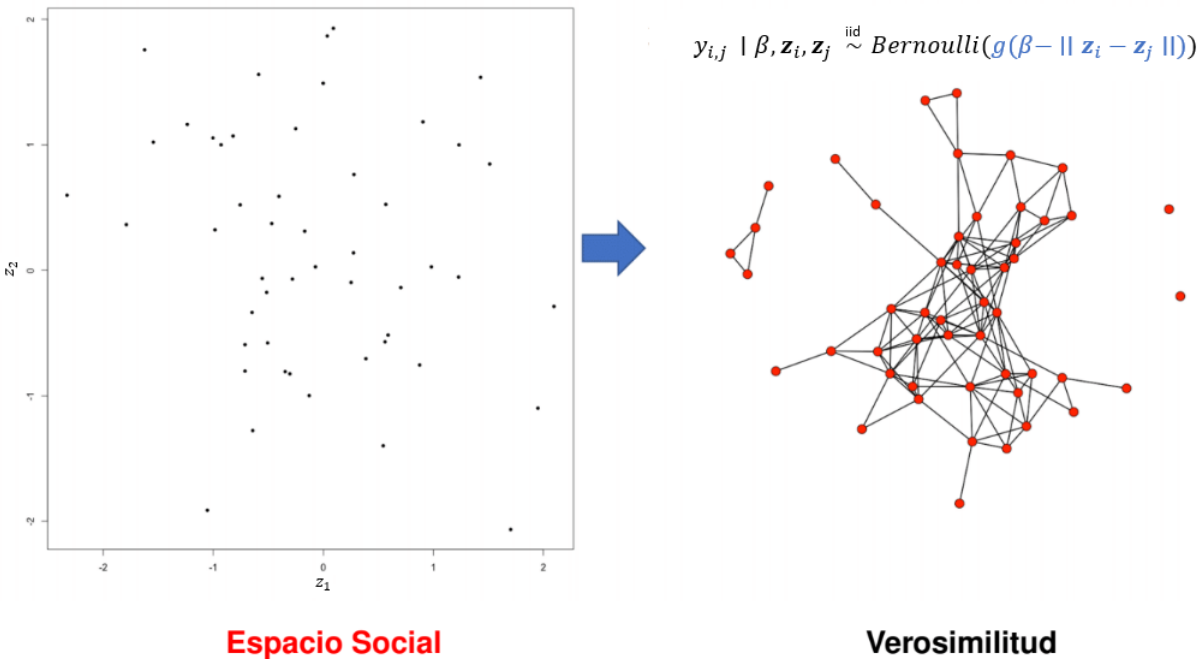


Figura 3.1.: Visualización de un espacio social bidimensional y su red. Vértices con posiciones cercanas en el Espacio Social son proclives a tener una mayor probabilidad de estar relacionados. La función $g(\cdot)$ corresponde a la función de enlace logit.

Esta configuración induce al modelamiento de la homofilia y la transitividad, definidos respectivamente por $\boldsymbol{x}_{i,j}$ y $\zeta_{i,j}$. Considerando que la función de enlace logit es monótona creciente y que $\zeta_{i,j}$ es negativo, vértices que estén cerca en el espacio latente tendrán mayor probabilidad de estar conectados.
La estimación es realizada desde un enfoque Bayesiano, asignando distribuciones previas para cada uno de los parámetros β_h con $h = 1, \ldots, p$ y para cada uno de los vectores $\boldsymbol{z}_i$, con $i = 1, \ldots, n$. Siendo $\mathbf{I}_d$ la matriz identidad de dimensión d, dentro de las escogencias comunes para las distribuciones previas se tiene:

$$\beta_h \sim \mathsf{N}(0, \psi_\beta) \quad h = 1, \ldots, p \quad \text{y} \quad \boldsymbol{z}_i \sim \mathsf{N}_d(\mathbf{0}_d, \psi_z \mathbf{I}_d) \; i = 1, \ldots, n$$

con ψ_β y ψ_z constantes. Dado que las distribuciones condicionales completas no tienen una forma estándar, la estimación se lleva a cabo usando Metropolis-Hastings. La estimación también se puede llevar a cabo utilizando otras técnicas además de MCMC. Por ejemplo, usando algoritmos variacionales. Una estimación de $\pi_{i,j}$ es

$$\hat{\pi}_{i,j} = \mathsf{E}(\pi_{i,j} \mid \boldsymbol{y}) = \frac{1}{B} \sum_{b=1}^{B} \frac{\exp\{\eta_{i,j}^{(b)}\}}{1 + \exp\{\eta_{i,j}^{(b)}\}}$$

con

$$\eta_{i,j}^{(b)} = \beta_0^{(b)} + \boldsymbol{x}_{\boldsymbol{i,j}}^{\boldsymbol{T}} \boldsymbol{\beta}^{(b)} - || \, \boldsymbol{z}_i^{(b)} - \boldsymbol{z}_j^{(b)} \, || \tag{3.4}$$

siendo $\beta_0^{(b)}$, $\boldsymbol{\beta}^{(b)}$ y $\mathbf{Z}^{(b)} = (\boldsymbol{z}_1^{(b)}, \ldots, \boldsymbol{z}_n^{(b)})^T$ muestras de la distribución posterior, para $b = 1, \ldots, B$. La Figura 3.2 muestra la representación DAG (Grafo acíclico dirigido) del modelo.

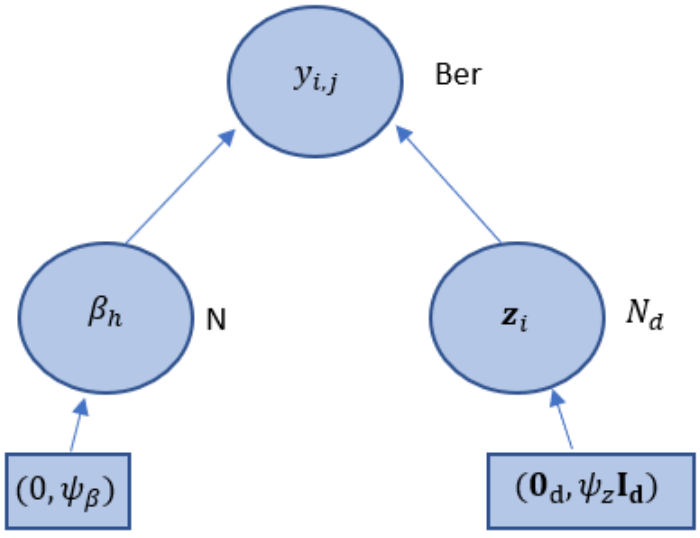


Figura 3.2.: Representación DAG del modelo de Espacio Latente.

Para ejemplificar, se ajusta el modelo sobre el conjunto de familias Florentinas. Por fines prácticos el modelo es ajustado únicamente con intercepto. El modelo final se presenta en la Ecuación (3.5) donde $\boldsymbol{z}_i \in \mathbf{R}^2$ para $i = 1, \ldots, n$. La estimación puntual del intercepto β_0 es $\hat{\beta}_0 = 4.9736$. La Figura 3.3 presenta la estimación para $\mathbf{Z}$.

$$\eta_{i,j} = \mathsf{logit}(\pi_{i,j}) = \beta_0 - || \boldsymbol{z}_i - \boldsymbol{z}_j || \tag{3.5}$$

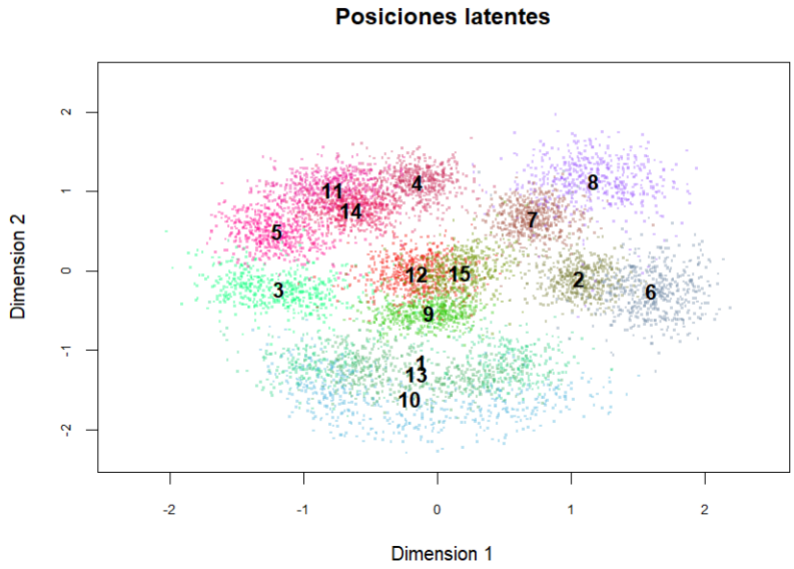


Figura 3.3.: Espacio latente estimado para las familias Florentinas.

Modelo de Proceso Socio-Espacial Latente

La formulación de la Ecuación (3.4) podría presentar problemas de sobreajuste [1] (Linkletter 2007, p. 38) dado que no está orientada a predecir probabilidades marginales de relacionarse para vértices fuera de la muestra observada. Además, ignora que la relación entre los atributos y la probabilidad de conexión puede ser compleja y cambiar de acuerdo a la región del espacio de covariables, por lo tanto, $\boldsymbol{x}_{i,j}$ no será suficiente para algunos casos. Linkletter (2007) propone flexibilizar el modelo y realizar un ajuste más suave de la relación entre $\mathbf{X}$ y $\mathbf{Y}$ a través de,

$$\eta_{i,j} = \mu - | z_i - z_j |$$

donde $z_i = z(\boldsymbol{x}_i)$ es una función de valor real latente (no observada) del vector de covariables $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})$ que recibe el nombre de modelo latente socio-espacial (LSSP, por sus

[1] Error de modelamiento en el que una función se corresponde únicamente con el conjunto de datos con el que fue estimada.

siglas en inglés) y μ es el log-ratio promedio de conexión para dos vértices con misma puntuación $LSSP$. Aquellos vértices con valores similares de z_i tienen más probabilidad de estar conectados. El valor absoluto de $z_i - z_j$ induce transitividad, que es un aspecto fundamental en el modelamiento de redes sociales.
Para ejemplificar, suponga que se desea modelar la relación entre la edad y el vínculo de amistad. Dados cuatro vértices con edades $x_1 = 5$, $x_2 = 10$, $x_3 = 30$ y $x_4 = 25$. Usualmente personas de 30 años tienen amigos de 25 años, caso contrario, no es usual que niños de 10 años tengan amigos de 5 años. Dependiendo de los valores que tome la variable edad, la misma diferencia absoluta de las edades entre dos personas tiene diferentes impactos sobre la probabilidad en los vínculos de amistad.
Para este modelo la distancia entre los z_i se mide a través del valor absoluto dado que son escalares. Caso contrario al modelo de Factor Latente, donde los $\boldsymbol{z}_i$ son vectores, de ahí que para calcular su distancia a otro vector $\boldsymbol{z}_j$ sea necesario el uso de la norma euclidiana.
La estimación Bayesiana de los parámetros se hace por medio de Metroplis-Hastings. La distribución previa para μ se puede asignar de manera conjugada por medio de una distribución Normal. La distribución previa para z_i es un proceso gaussiano. Dado un conjunto finito de observaciones se define $\mathbf{Z}^{(b)} = z(\boldsymbol{X}) = (z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_n))^T$ con distribución normal multivariada.
Dado un proceso ruido blanco [2] discretizado con espacio de estados $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$, restringido a una grilla $\mathsf{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m)$ obtenida a través de un diseño de hipercubo latino [3] con $\boldsymbol{w}_r \in \boldsymbol{\chi}$ para $r = 1, \ldots, m$, siendo $\boldsymbol{\chi}$ el soporte de $z(\boldsymbol{x})$. La siguiente expresión corresponde a una aproximación para el proceso gaussiano $z(\boldsymbol{x})$

$$z(\boldsymbol{x}) = \sum_{r=1}^{m} \alpha_r k(\boldsymbol{x} - \boldsymbol{w}_r) \tag{3.6}$$

con $k(\cdot)$ un kernel gaussiano de dimensión p estructurado independientemente. Asumiendo que el ancho de banda del kernel varia en la dirección de cada covariable se define el vector de parámetros $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)$ y $k(\cdot)$ se puede parametrizar como,

$$k(\boldsymbol{x}_i - \boldsymbol{w}_r) = \prod_{h=1}^{p} \rho_h^{(w_{r,h} - x_{i,h})^2}$$

con $w_{r,h}$ y $x_{i,h}$ los elementos en la posición h de $\boldsymbol{w}_r$ y $\boldsymbol{x}_i$ respectivamente, $\rho = e^{-\frac{1}{2\sigma_h^2}}$ y σ_h es la desviación estándar del kernel en la dirección h.
Finalmente, la distribución previa para los coeficientes de la expansión dada en la Ecuación (3.6) es

[2] Proceso estocástico media cero, de varianza constante y con todas sus variables independientes.
[3] Método estadístico para generar una muestra casi aleatoria de valores de los parámetros de una distribución conjunta dada.

$$\boldsymbol{\alpha} \sim \mathsf{N}_m(\mathbf{0}_m, \mathbf{I}_m) \tag{3.7}$$

donde $\mathbf{I}_m$ la matriz identidad de orden m. Además, la distribución previa de ρ_h y μ es,

$$\rho_h \sim \mathsf{U}[0,1] \qquad \text{para} \quad h = 1, \ldots, p \quad \text{y} \quad \mu \sim \mathsf{N}(0, \psi_\mu). \tag{3.8}$$

Dada la distribución previa de la Ecuación (3.7) para $\boldsymbol{\alpha}$ y asumiendo que el proceso para z es un proceso gaussiano media cero, su estructura de covarianza está dada por,

$$\mathsf{Cov}(z(\boldsymbol{x}_i), z(\boldsymbol{x}_j)) = \sum_{r=1}^{m} k(\boldsymbol{x}_i - \boldsymbol{w}_r) k(\boldsymbol{x}_j - \boldsymbol{w}_r)$$

por tanto,

$$\hat{z}(\boldsymbol{x}_0) = \mathsf{E}(z(\boldsymbol{x}_0) \mid \boldsymbol{y}) = \frac{1}{B} \sum_{b=1}^{B} \sum_{r=1}^{m} \alpha_r^{(b)} k^{(b)}(\boldsymbol{x}_0 - \boldsymbol{w}_r)$$

donde $z_{\boldsymbol{x}_0}^{(b)} = \sum_{r=1}^{m} \alpha_r^{(b)} k^{(b)}(\boldsymbol{x}_0 - \boldsymbol{w}_r)$, con $(\mu^{(b)}, \boldsymbol{\alpha}^{(b)}, \boldsymbol{\rho}^{(b)})$ las realizaciones obtenidas vía MCMC a partir de la distribución posterior,

$$\mathsf{p}(\mu, \boldsymbol{\alpha}, \boldsymbol{\rho} \mid \boldsymbol{y}) = \mathsf{p}(\boldsymbol{y} \mid \mu, \boldsymbol{\alpha}, \boldsymbol{\rho}) \mathsf{p}(\mu) \mathsf{p}(\boldsymbol{\alpha}) \mathsf{p}(\boldsymbol{\rho}) \tag{3.9}$$

y

$$k^{(b)}(\boldsymbol{x}_0 - \boldsymbol{w}_r) = \prod_{h=1}^{p} {\rho_h^{(b)}}^{(w_{r,h} - x_{i,h})^2}.$$

con μ un escala, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ un vector de tamaño y $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)$ un vector de tamaño p. La Figura 3.4 muestra la representación DAG del modelo.

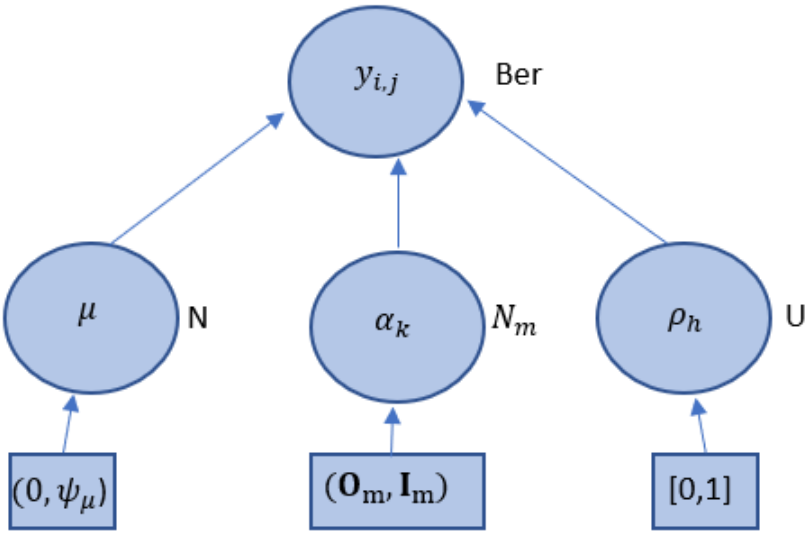


Figura 3.4.: Representación DAG del modelo de Proceso Socio-Espacio Latente.

Finalmente, una estimación para las probabilidades de interacción $\pi_{i,j}$ de dos vértices i, j con vectores de covariables $\boldsymbol{x}_i$ y $\boldsymbol{x}_j$ se puede calcular como:

$$\hat{\pi}_{i,j} = \frac{1}{B} \sum_{b=1}^{B} \frac{\exp\{\eta_{i,j}^{(b)}\}}{1 + \exp\{\eta_{i,j}^{(b)}\}}$$

donde

$$\eta_{i,j}^{(b)} = \mu^{(b)} - \mid z^{(b)}(\boldsymbol{x}_i) - z^{(b)}(\boldsymbol{x}_j) \mid \quad \forall \ b = 1, \ldots, B.$$

En el siguiente repositorio se encuentra toda la implementación ejemplificada y documentada para poder replicar el método de estimación, predicción y estudio de poder predictivo del modelo.

4. Estudio de Simulación

4.1. Primer Conjunto de Datos Sintético

Las Figuras 4.1 y 4.2 muestran la representación gráfica de vértices y enlaces para una red con log-ratio promedio de conexión igual a $\mu = -0.5$, de orden $n = 40$ y superficie dada por $g(x, y) = 1.5x^2\exp\{x^2\}$.

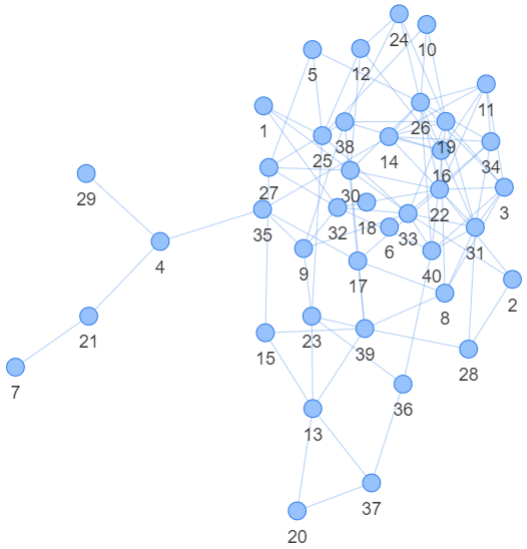


Figura 4.1.: Red de los datos del primer caso de simulación.

4.1.1. Análisis descriptivo de la red

Los vértices 26, 31 y 33 tienen los valores más altos de centralidad por cercanía. Los vértices 35, 4 y 39 son los más centrales por intermediación. El promedio de la distancia geodésica es 2.1098. La distancia geodésica más grande es de 7. Existe un efecto de agrupamiento bajo con un valor de la transitividad igual a 0.2138. La densidad de la red es igual a 0.1231.

4.1.2. Modelamiento

En adelante, las especificaciones de calentamiento, réplicas y salto se conservan iguales a los presentados a continuación a menos de que se especifique lo contrario.
Dado un período de calentamiento de 50000, un número de réplicas Monte Carlo de 10000, un salto sistemático (adelgazamiento) de 10 iteraciones, y una matriz de covariables $\mathbf{X} \in [0,1] \times [0,1]$ de dimensión 40×2 simulada a partir de una distribución uniforme continua en el intervalo $[0,1]$, se presentan los resultados de estimación del modelo.
Los paneles a,b,c,d, e y f de la Figura A.1 permiten estudiar la convergencia de los diferentes parámetros de interés, las cadenas realizan una correcta exploración del espacio de estados. Con base es los paneles g y h se realiza el estudio de convergencia para z_i, para ambos casos se presenta un comportamiento simétrico al rededor del valor teórico.
La Figura 4.3 presenta la superficie estimada que conserva bastante similitud con la superficie presentada en la Figura 4.2. Lo anterior permite concluir un buen desempeño en la estimación de los parámetros involucrados en el modelo.

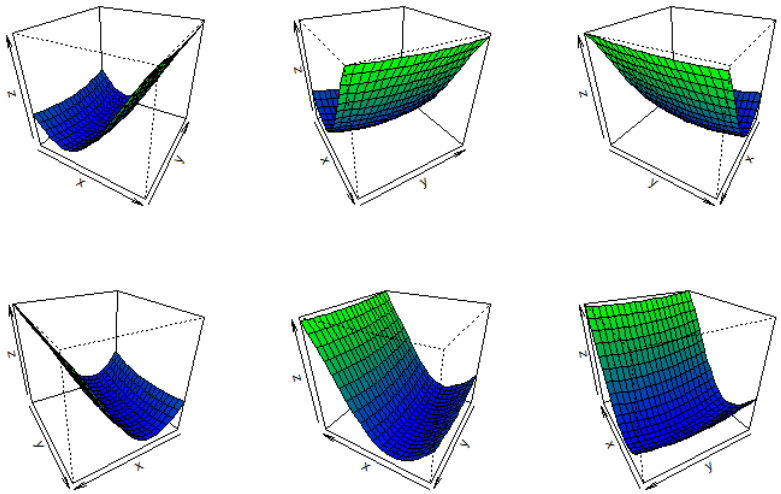
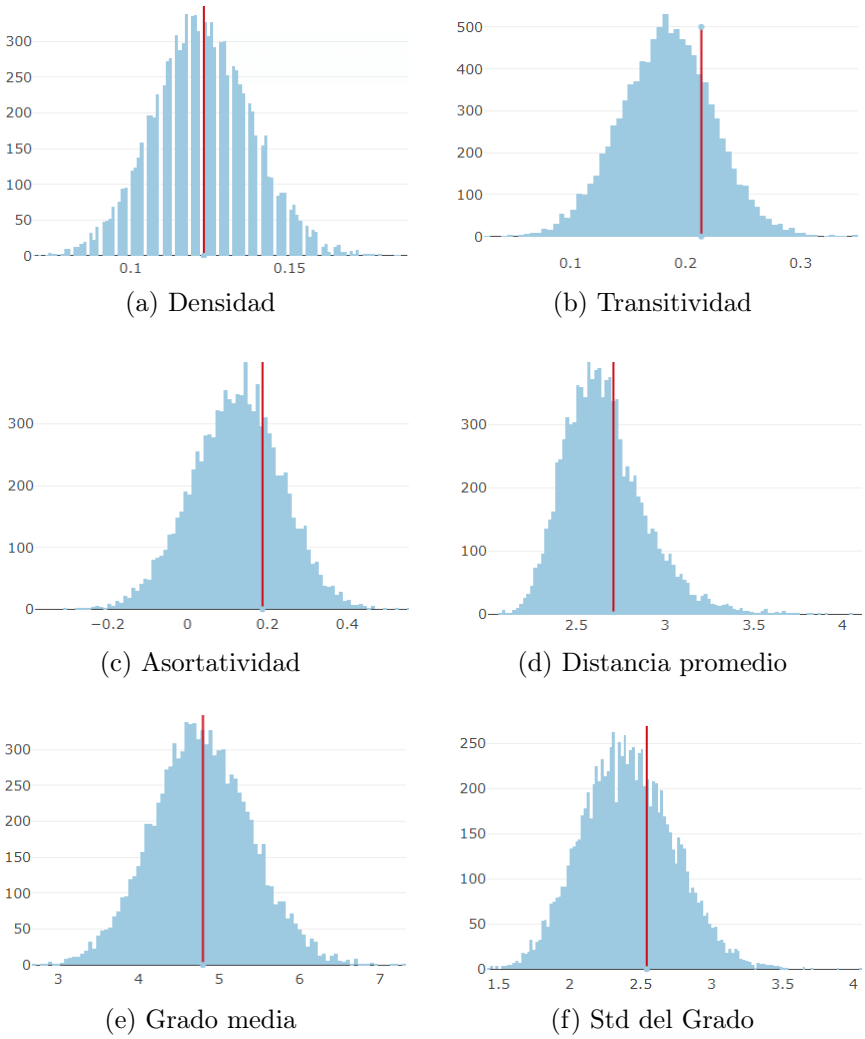


Figura 4.3.: Superficie estimada de los datos del primer caso de simulación.

4.1.3. Bondad de Ajuste

Las estadísticas para estudiar corresponden a la densidad, la transitividad, la asortatividad, la distancia promedio, el grado promedio de los vértices y su desviación estándar. La Figura 4.4 permite concluir que el modelo es capaz de replicar adecuadamente las estadísticas de interés, los valores predictivos posteriores pertenecen al dominio de la distribución estimada de las estadísticas.



(a) Densidad (b) Transitividad

(c) Asortatividad (d) Distancia promedio

(e) Grado media (f) Std del Grado

Figura 4.4.: Estadísticas para la evaluación de bondad de ajuste del modelo de los datos del primer caso de simulación.

4.1.4. Predicción

La Figura 4.5 presenta las curvas ROC para cada grupo del procedimiento de validación cruzada. Se presenta un desempeño bueno de la calidad predictiva del modelo dado que los valores se encuentran entre 0.7 y 0.9. El vector de AUC de los grupos es igual a $(0.73, 0.73, 0.84, 0.73, 0.76)$. El AUC estimado del modelo es igual a 0.76.

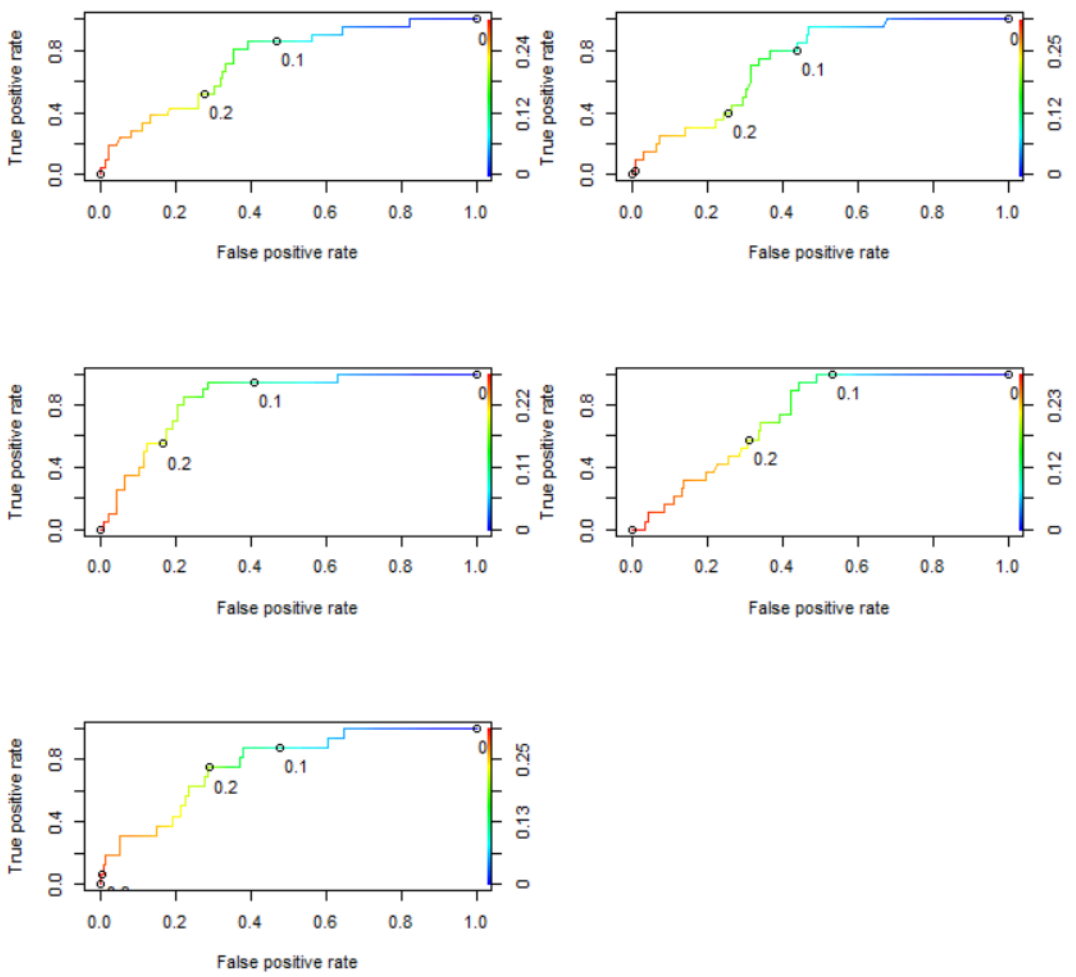


Figura 4.5.: Curvas ROC del procedimiento de validación cruzada de los datos del primer caso de simulación.

4.2. Segundo Conjunto de Datos Simulados

Las Figuras 4.6 y 4.7 muestran la representación gráfica de vértices y enlaces para una red con log-ratio promedio de conexión igual a $\mu = -0.7$, de orden $n = 50$ y superficie dada por $g(x, y) = 1.5\exp\{x^2\}\sin((x + y)^2)$.

4.2.1. Análisis descriptivo de la red

El vértice con mayor grado, y con valor más alto de centralidad por cercanía y por intermediación es el vértice 34. El vértice 12 tiene los dos valores siguientes más altos para el grado y para la cercanía por intermediación. En promedio el camino más corto es de

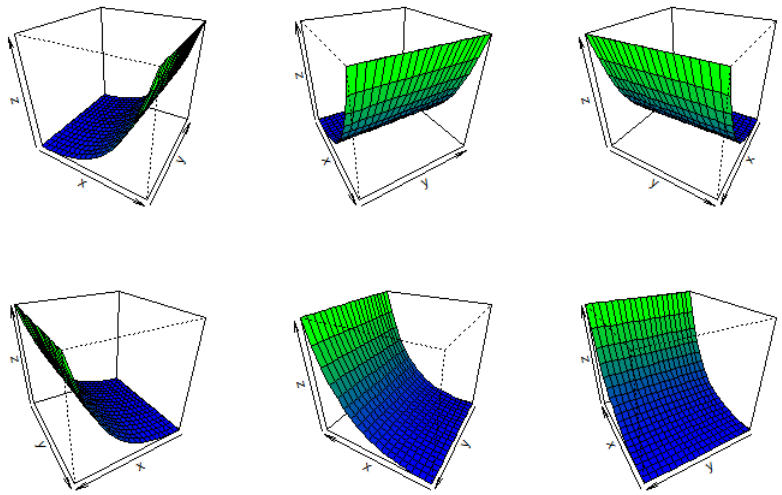


Figura 4.2.: Superficie de los datos del primer caso de simulación.

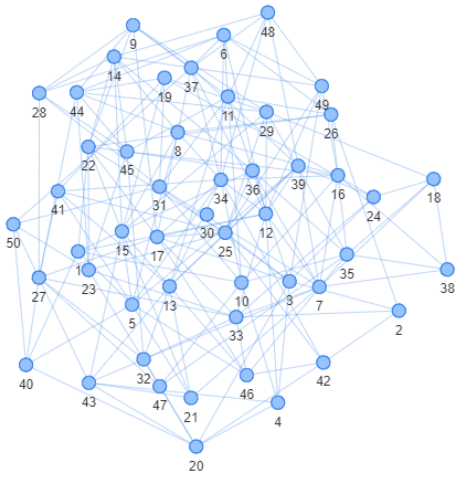


Figura 4.6.: Red de los datos del segundo caso de simulación.

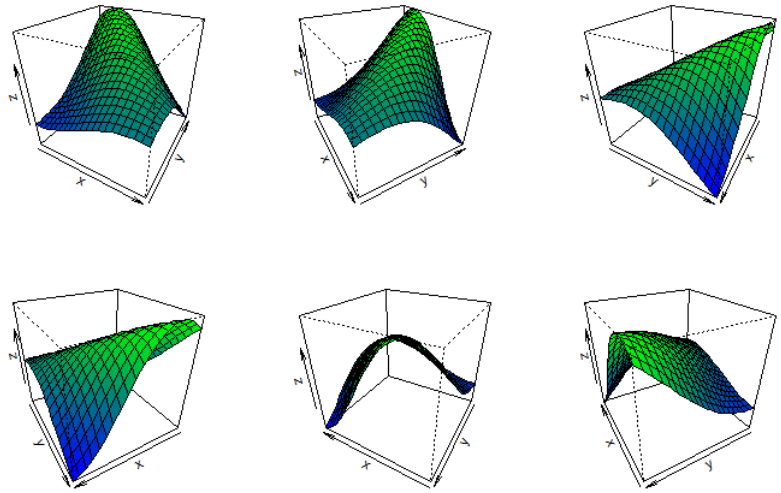


Figura 4.7.: Superficie de los datos del segundo caso de simulación.

2.1054. El diámetro de la es red es 4. La transitividad es igual a 0.1543 y la densidad es de 0.1539.

4.2.2. Modelamiento

Los paneles a, b, c, d, e y f de la Figura A.2 permiten estudiar la convergencia de los diferentes parámetros de interés, las cadenas realizan una correcta exploración del espacio de estados. Con base es los paneles g y h se realiza el estudio de convergencia para z_i, para ambos casos se presenta un comportamiento simétrico al rededor del valor teórico.

En la Figura 4.8 se presenta la superficie estimada. El modelo tiene un desempeño regular de estimación de la superficie. Se observan dificultades para estimar las partes bajas de la superficie con respecto a la Figura 4.7.

4.2.3. Bondad de Ajuste

Las estadísticas para estudiar corresponden a la densidad, la transitividad, la asortatividad, la distancia promedio, el grado promedio de los vértices y su desviación estándar. La Figura 4.9 permite concluir que el modelo es capaz de replicar adecuadamente las estadísticas de interés, los valores predictivos posteriores pertenecen al dominio de la distribución estimada de las estadísticas.

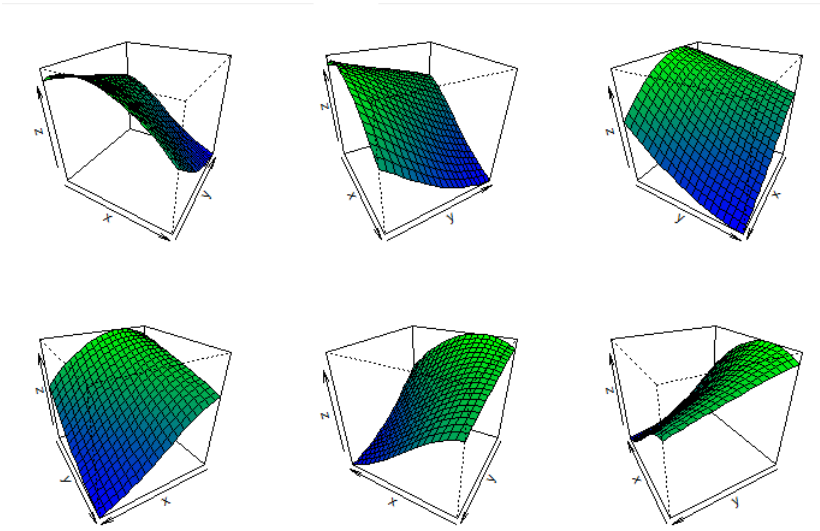


Figura 4.8.: Superficie estimada de los datos del segundo caso de simulación.

4.2.4. Predicción

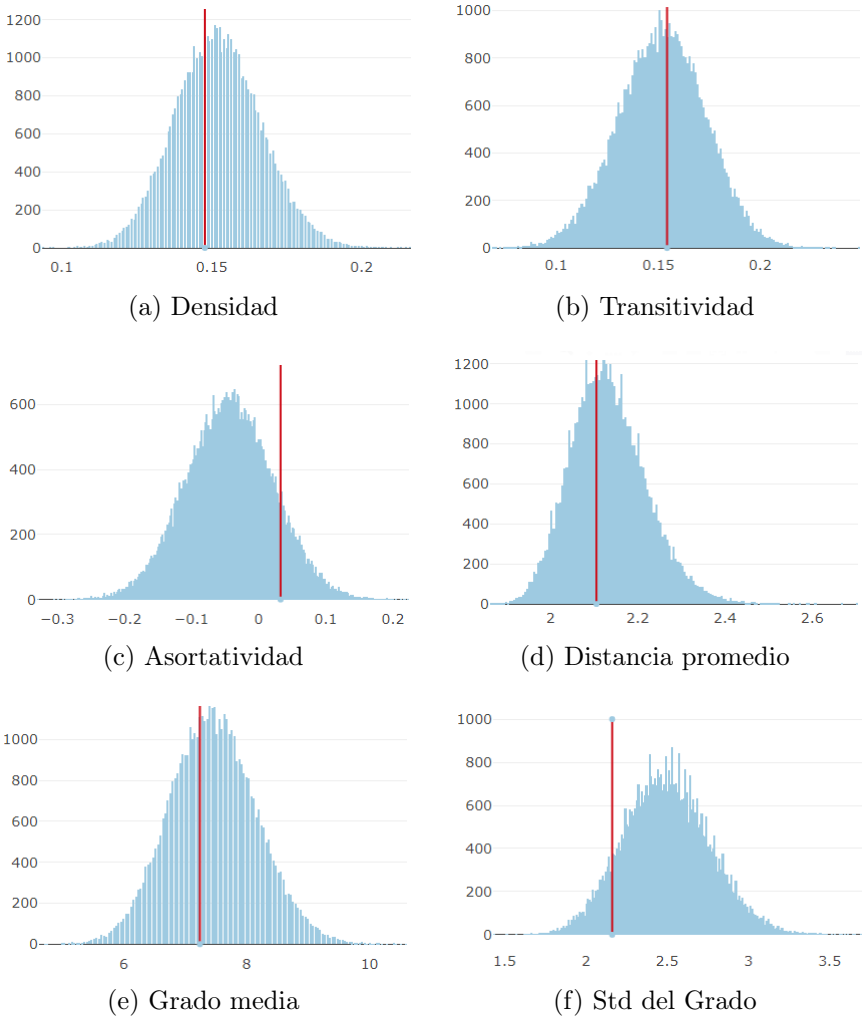
La Figura 4.10 presenta las curvas ROC para cada grupo del procedimiento de validación cruzada. Se presenta un desempeño regular de la calidad predictiva del modelo dado que los valores se encuentran entre 0.5 y 0.6. El vector de AUC de los grupos es igual a $(0.59, 0.54, 0.52, 0.59, 0.54)$. El AUC estimado del modelo es igual a 0.56.

4.3. Consolidación de Resultados

Para ambos casos la bondad del ajuste es buena, por lo tanto, las inferencias sobre la topología de la red son confiables. En términos predictivos los resultados son muy sensibles a la superficie. Superficies con especificaciones funcionales muy complejas son más difíciles predecir por modelo. La tabla 4.1 presenta el resumen de las medidas de bondad predictivo del modelo para los dos conjuntos de datos sintéticos.

Modelo	Orden	Tamaño	AUC
Modelo 1	40	96	0.76
Modelo 2	50	181	0.56

Tabla 4.1.: AUC estimados para los modelos de datos sintéticos.



(a) Densidad (b) Transitividad

(c) Asortatividad (d) Distancia promedio

(e) Grado media (f) Std del Grado

Figura 4.9.: Estadísticas para la evaluación de bondad de ajuste del modelo de los datos del segundo caso de simulación.

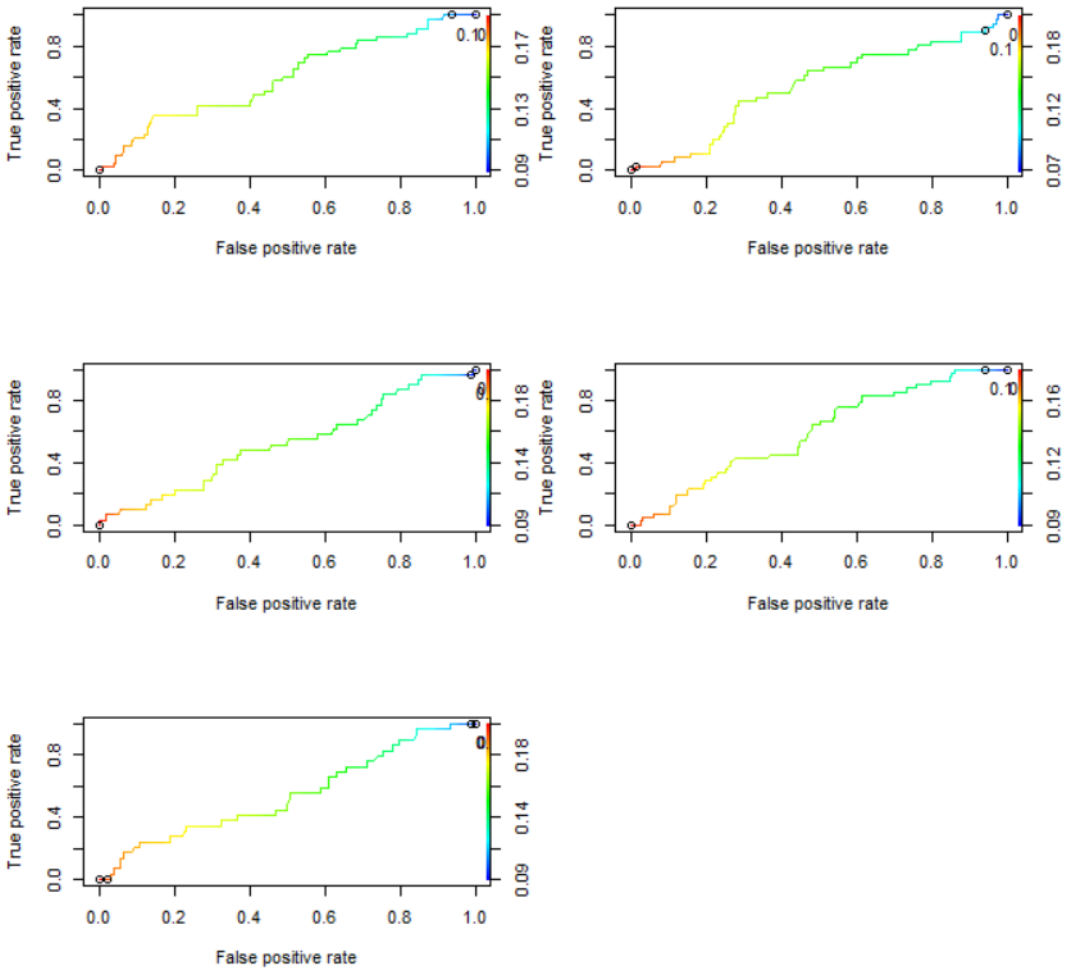


Figura 4.10.: Curvas ROC del procedimiento de validación cruzada de los datos del segundo caso de simulación.

5. Aplicación en Datos Reales

5.1. Zachary

Esta red es el resultado de 2 años de estudio de las interacciones sociales entre 34 individuos de un club de karate durante la década de los 70 (Zachary 1977). Esta red ha sido ampliamente estudiada en la literatura (e.g., Kolaczyk and Csárdi 2020). Los enlaces entre dos vértices indican interacción social. John A. (34) y Mr Hi (1) corresponden al instructor y administrador del club. Las covariables empleadas corresponden a la centralidad cercanía y centralidad por intermediación de los vértices.

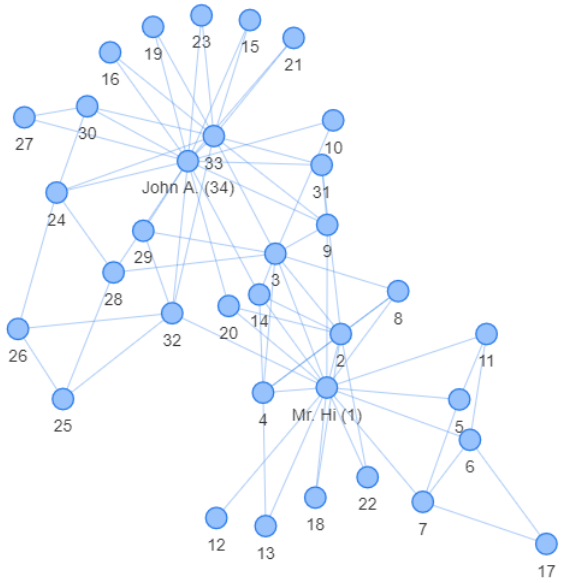


Figura 5.1.: Red de los datos de Zachary.

5.1.1. Análisis Descriptivo de la Red

La Figura 5.2 permite ver que la mayoría de los individuos tienen grado alrededor de 5. Vértices con grado superior a 10 corresponden a John A (34), Mr Hi (1) y el vértice 33, que son los más centrales en términos de cercanía e intermediación.

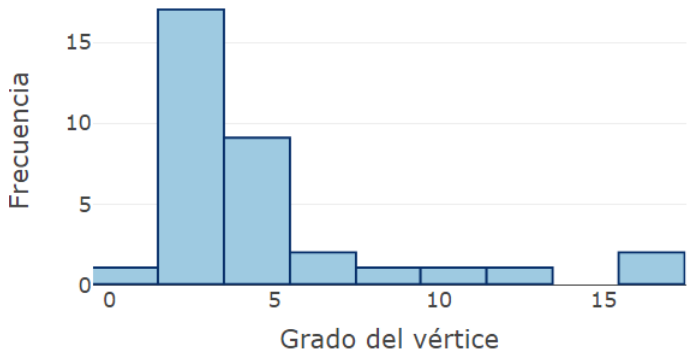


Figura 5.2.: Distribución del grado de los vértices de los datos de Zachary.

La red tiene un coeficiente de transitividad de 0.256. Se presenta el 14 % de todas las posibles conexiones y el coeficiente de asortatividad es igual a -0.476. Las covariables empleadas para el modelamiento incluyen una binaria del club al que pertenece cada individuo y las medidas de centralidad estudiadas.

En la Figura 5.3 se realiza una partición de la red. El primer grupo está definido mayoritariamente por individuos que interactúan con John A (34). Análogamente se tiene para Mr Hi (1). El tercer grupo está definido por individuos cuyas interacciones están en los dos grupos mencionados previamente.

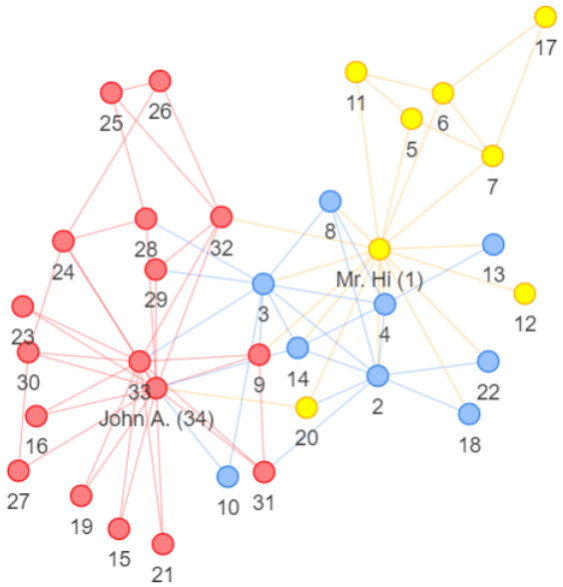


Figura 5.3.: Partición vía métodos jerárquicos de la red de los datos de Zachary.

5.1.2. Modelamiento

Los paneles a, b, c, d, e y f de la Figura A.3 permiten estudiar la convergencia de los diferentes parámetros de interés, las cadenas realizan una correcta exploración del espacio de estados. Con base es los paneles g y h se realiza el estudio de convergencia para z_i, para ambos casos se presenta un comportamiento simétrico al rededor del valor teórico. La Figura 5.4 presenta la superficie estimada para el conjunto de datos.

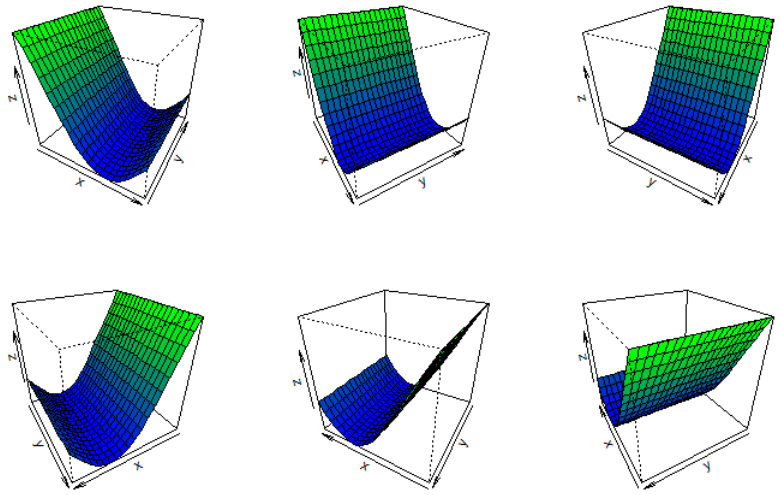
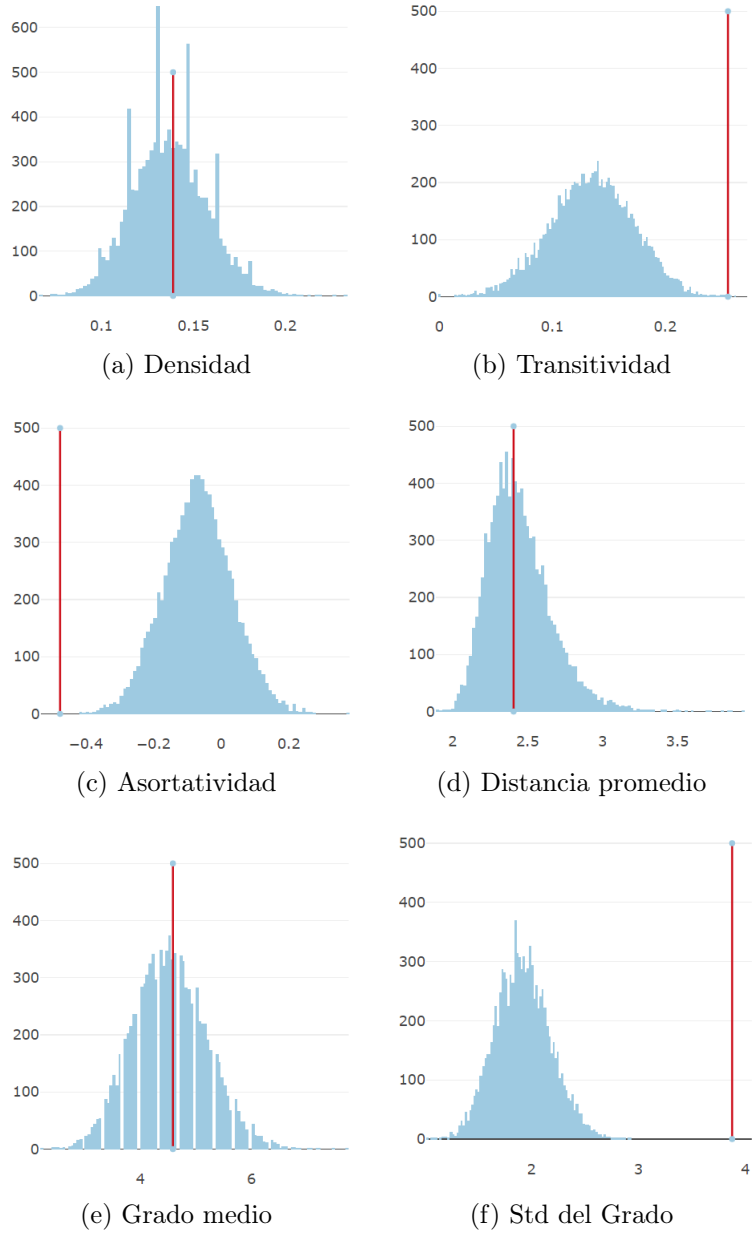


Figura 5.4.: Superficie estimada de los datos de Zachary.

5.1.3. Bondad de Ajuste

Las estadísticas para estudiar corresponden a la densidad, la transitividad, la asortatividad, la distancia promedio, el grado promedio de los vértices y su desviación estándar. La Figura 5.5 presenta la distribución de las estadísticas de interés. El modelo es capaz de replicar adecuadamente la densidad, la distancia promedio y el grado promedio, caso contrario a la asortatividad, la transitividad y la desviación estándar del grado. En conclusión, el agrupamiento y la correlación no son capturados de forma correcta por el modelo, caso contrario a la centralidad.



(a) Densidad (b) Transitividad

(c) Asortatividad (d) Distancia promedio

(e) Grado medio (f) Std del Grado

Figura 5.5.: Estadísticas para la evaluación de bondad de ajuste del modelo de los datos de Zachary.

5.1.4. Predicción

La Figura 5.6 presenta las curvas ROC para cada grupo del procedimiento de validación cruzada. Se presenta un desempeño aceptable de la calidad predictiva del modelo dado que los valores se encuentran entre 0.6 y 0.7. El vector de AUC de los grupos es igual a

$(0.67, 0.64, 0.63, 0.66, 0.69)$. El AUC estimado del modelo es igual a 0.66.

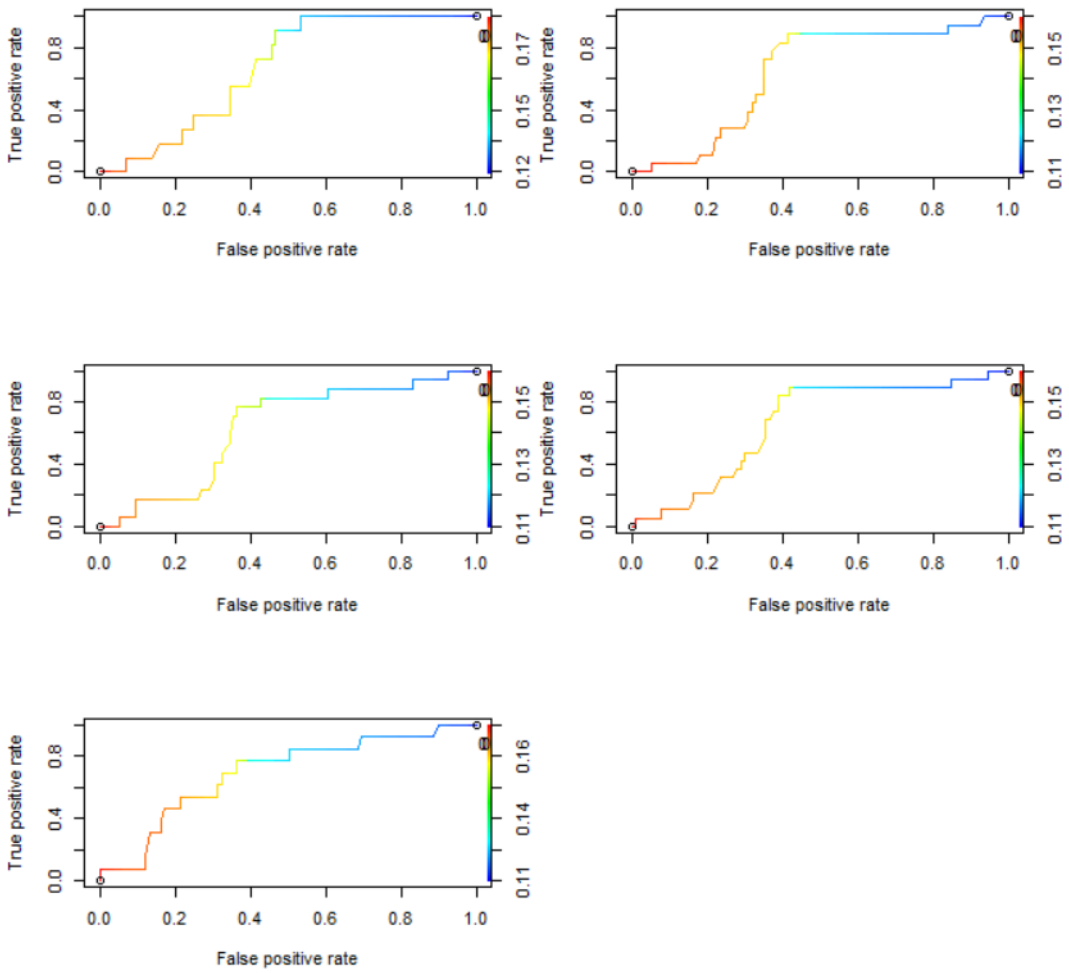


Figura 5.6.: Curvas ROC del procedimiento de validación cruzada de los datos de Zachary.

5.2. Lazega

El conjunto de datos (ampliamente estudiado en la literatura, e.g., Kolaczyk and Csárdi 2020) está compuesto por 71 abogados (socios y asociados) pertenecientes a una red de sociedades de derecho corporativo estudiada en un bufete corporativo de EEUU en 1988-1991. Los predictores son la antigüedad, el estado formal, la oficina en la que trabaja, el género, la facultad de derecho a la que asistió, las mediciones de desempeño individual, la edad, entre otros. Se utiliza la antigüedad y la edad re-escalados al intervalo $[0, 1]$ como predictores del modelo.

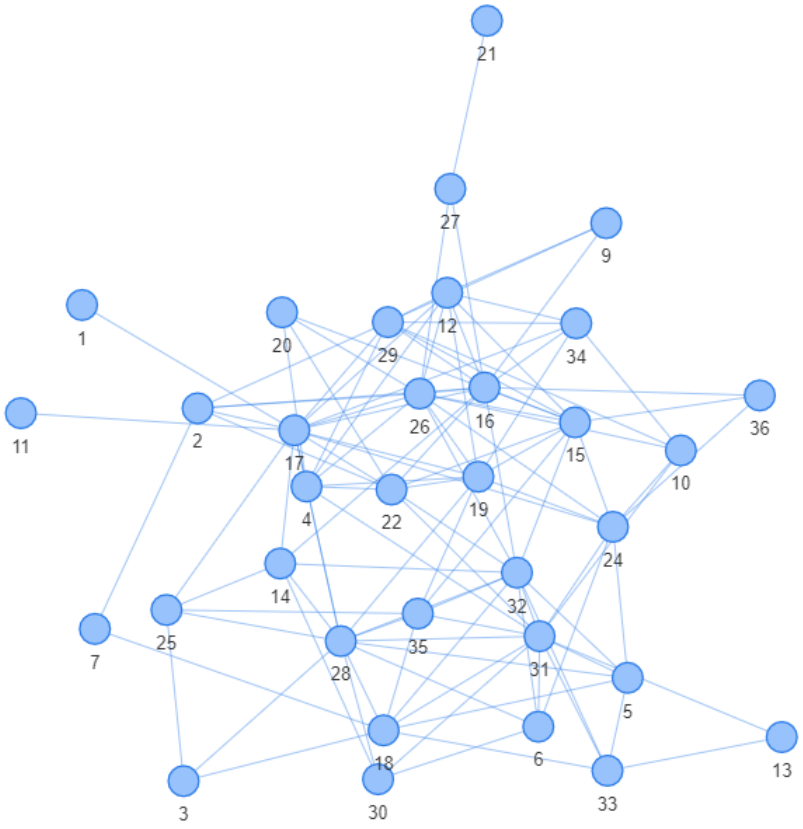


Figura 5.7.: Red de los datos de Lazega.

5.2.1. Análisis Descriptivo de la Red

Los vértices 16 y 15 tienen los valores más altos de centralidad. El coeficiente de transitividad es 0.389. El grafo particionado se presenta en la Figura 5.8. Se presenta el 21 % de todas las posibles conexiones y el coeficiente de asortatividad es igual a -0.1681.

5.2.2. Modelamiento

Los paneles a, b, c, d, e y f de la Figura A.4 permiten estudiar la convergencia de los diferentes parámetros de interés, las cadenas realizan una correcta exploración del espacio de estados. Con base es los paneles g y h se realiza el estudio de convergencia para z_i, para ambos casos se presenta un comportamiento simétrico al rededor del valor teórico. La Figura 5.9 presenta la superficie estimada para el conjunto de datos.

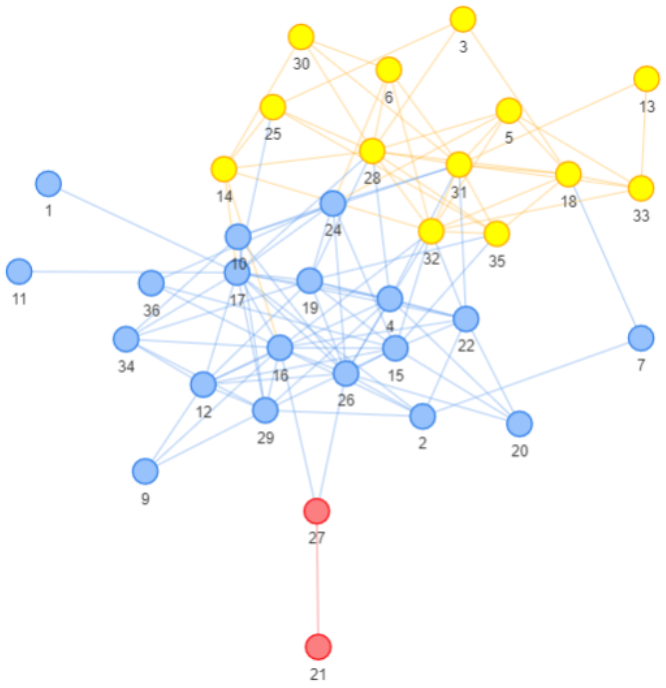


Figura 5.8.: Partición vía métodos jerárquicos de la red de los datos de Lazega.

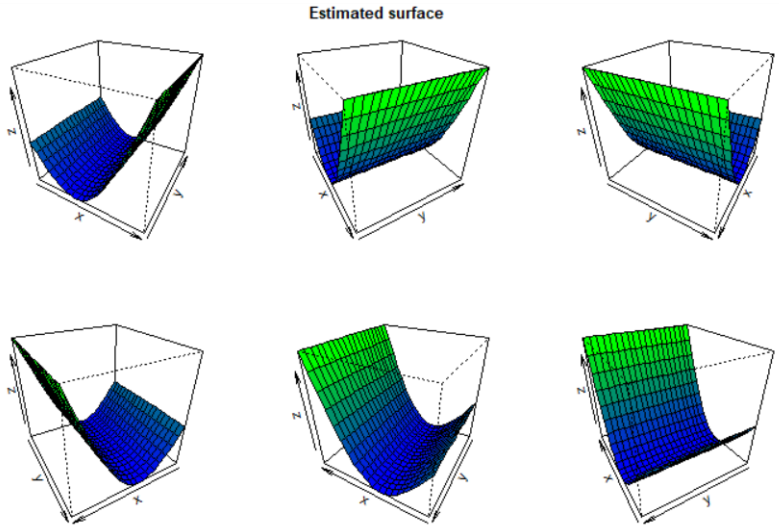
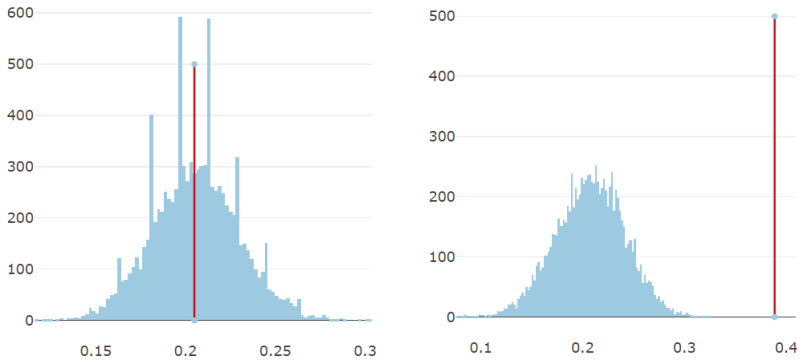


Figura 5.9.: Superficie estimada

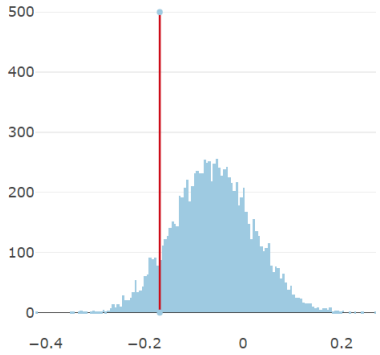
5.2.3. Bondad de Ajuste

Las estadísticas para estudiar corresponden a la densidad, la transitividad, la asortatividad, la distancia promedio, el grado promedio de los vértices y su desviación estándar. La Figura 5.10 muestra que el modelo replica satisfactoriamente todas las estadística de interés de la red a excepción de la transitividad y la desviación estándar del grado.

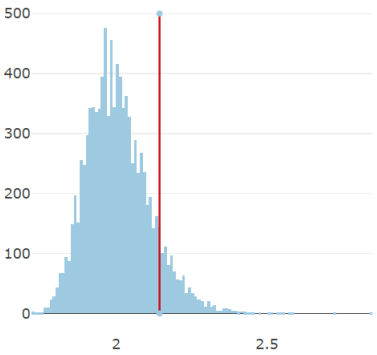


(a) Densidad

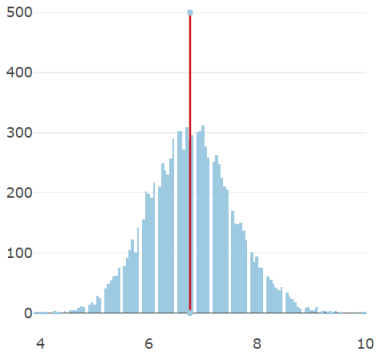
(b) Transitividad



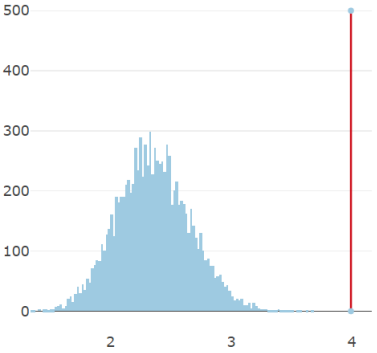
(c) Asortatividad



(d) Distancia promedio



(e) Grado medio



(f) Std del Grado

Figura 5.10.: Estadísticas para la evaluación de bondad de ajuste de los datos de Lazega.

5.2.4. Predicción

La Figura5.11 presenta las curvas ROC para cada grupo del procedimiento de validación cruzada. Se presenta un desempeño regular de la calidad predictiva del modelo dado que los valores se encuentran entre 0.5 y 0.7. El vector de AUC de los grupos es igual a $(0.49, 0.70, 0.50, 0.50, 0.66)$. El AUC estimado del modelo es igual a 0.57.

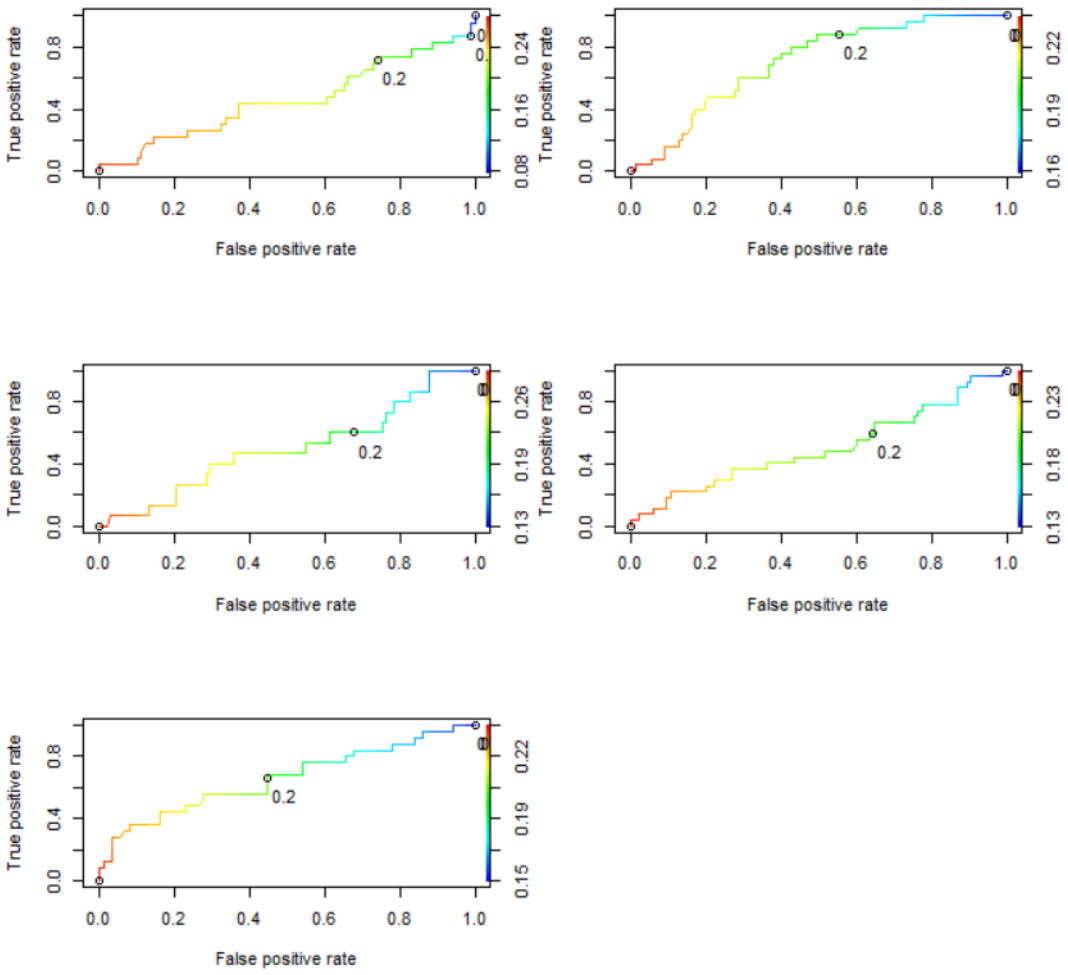


Figura 5.11.: Curvas ROC del procedimiento de validación cruzada de los datos de Lazega.

5.3. Datos de conflicto internacional

El conjunto de datos reporta los conflictos internacionales entre 130 países desde 1990-2000, recolectados por Mike Ard y Xun Cao del departamento de ciencias políticas de la universidad de Washington.
Se define $y_{i,j} = 1$ si el país i tuvo algún conflicto con el país j y $y_{i,j} = 0$ en otro caso. El

vector de covariables $\boldsymbol{x}_{i,j}$ corresponde a la medición de 7 atributos más el intercepto. Las 7 covariables se describen a continuación:

1. Logaritmo de la población de la nación agresora.

2. Puntaje político de la nación agresora.

3. Logaritmo de la población de la nación objetivo.

4. Puntaje político de la nación objetivo.

5. Distancia geográfica entre las dos naciones.

6. Término de interacción entre el score político de la nación agresora y la nación objetivo.

7. Número de organizaciones intergubernamentales en las que ambas naciones son miembros.

Dado que en este documento se trabaja sobre redes no dirigidas y que es de interés el estudio analítico y gráfico de la superficie subyacente, son conservadas únicamente dos de las covariables mencionadas (las dos con variabilidad más alta), el logaritmo de la población de la nación y el puntaje político de la nación. Para seguir de manera fidedigna el proceso de modelamiento presentado ambas covariables son re-escaladas al intervalo $[0, 1]$ (ver Figuras 5.12 y 5.13). Se puede consultar Ward and Hoff 2007 para más detalles sobre el conjunto de datos bajo estudio.

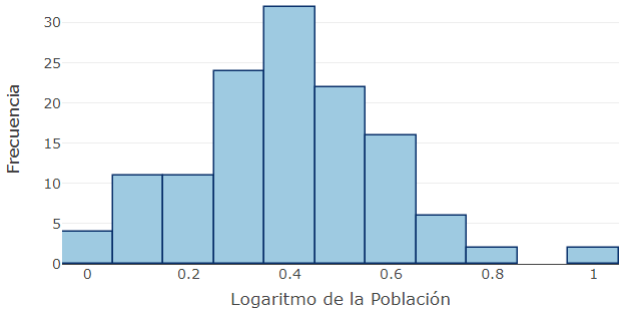


Figura 5.12.: Histograma del logaritmo de la población de los datos de conflicto.

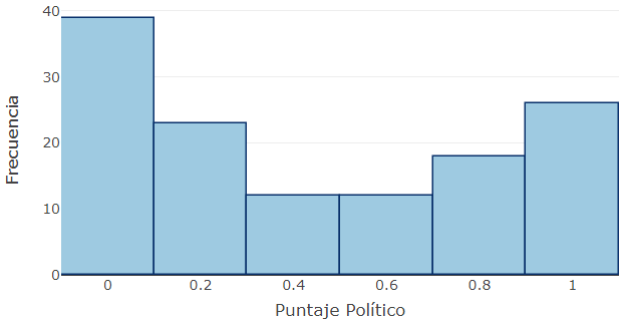


Figura 5.13.: Histograma del puntaje político por nación de los datos de conflicto.

5.3.1. Análisis Descriptivo de la Red

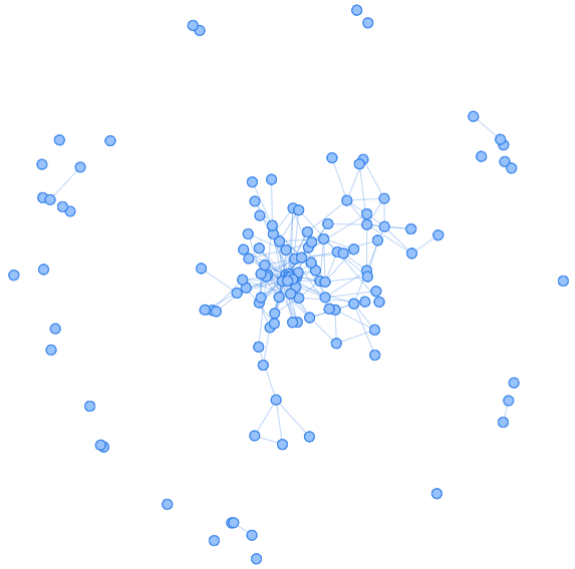


Figura 5.14.: Red de los datos de conflicto.

En la Figura 5.14 se presenta la red de contactos de conflictos. Se observa que existen varios vértices con grado cero o igual a uno. En la Figura 5.15 se muestra la distribución del grado de los vértices. Es una distribución sesgada a derecha, con más del 90 % de los vértices con grado menor o igual a 10. Iraq, Jordania y USA son los países más centrales para el grado, la centralidad por cercanía y la centralidad por intermediación. La Figura 5.16 presenta el grado de los vértices versus el promedio del grado de sus vecinos. Países con grados altos tienden a presentar conflicto con países de grados bajos. Países con grado bajo se enlazan con países de grado alto.

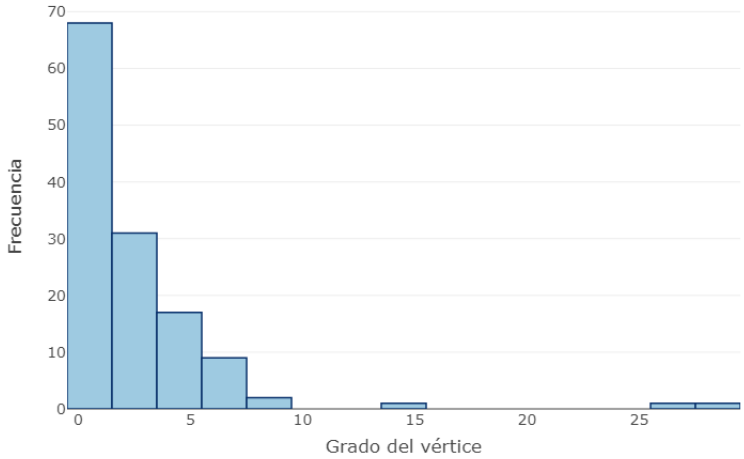


Figura 5.15.: Distribución del grado de los vértices de los datos de conflicto.

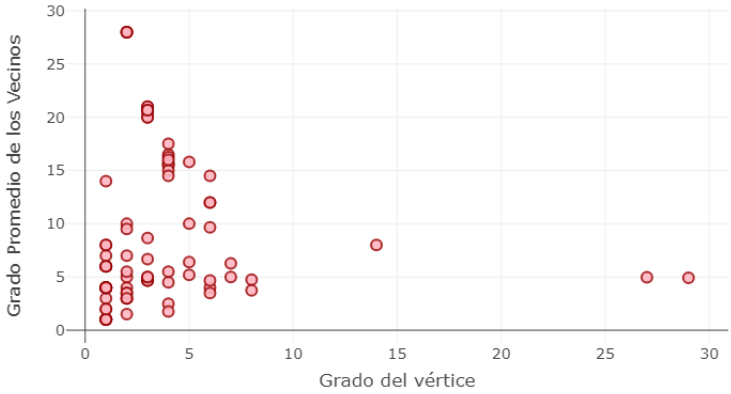


Figura 5.16.: Grado de los vértices de los datos de conflicto vs el promedio del grado de sus vecinos.

Dado que la red no está conectada se estudia su gran componente, compuesta por 83 países. Esta sub-red contiene más del 60 % de los vértices de la red total y es presentado en la Figura 5.17.

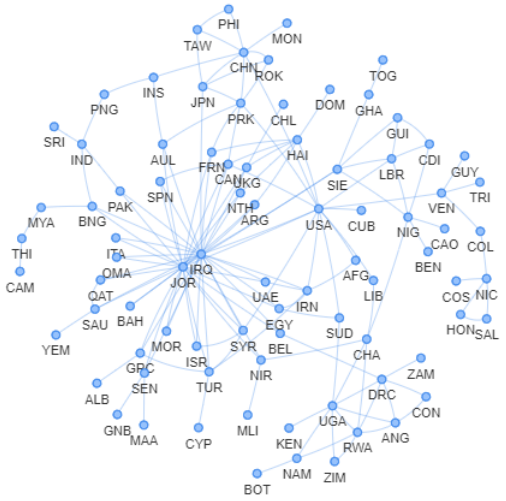


Figura 5.17.: Gran componente de la red de los datos de conflicto.

La distancia media de la sub-red es 3.656. Su diámetro y transitividad son iguales que los de la red completa, 9 y 0.1634 respectivamente. Existen 22 puntos de articulación, i.e., basta con sacar del conflicto algunos países de manera estratégica para disminuir la ocurrencia de los conflictos. La densidad de la red completa corresponde a 0.0191.

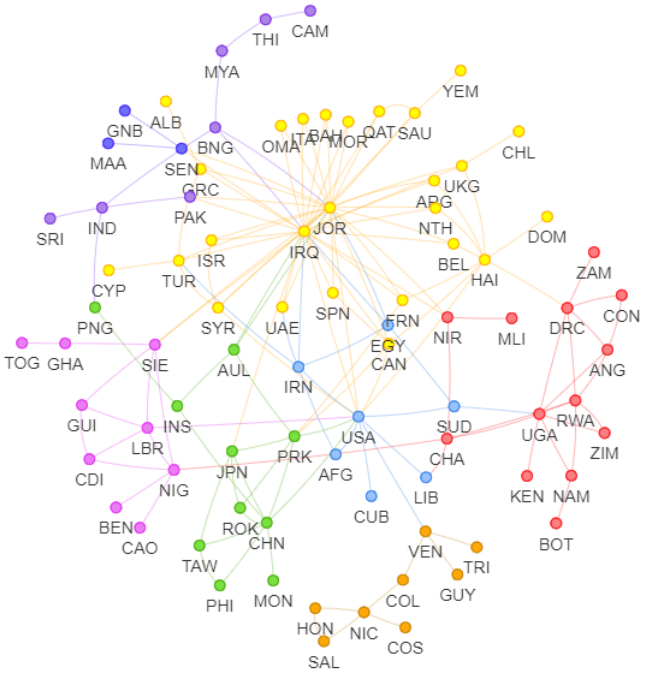


Figura 5.18.: Partición vía métodos jerárquicos de la gran componente de la red de los datos de conflicto.

5.3.2. Modelamiento

Los paneles a, b, c, d, e y f de la Figura A.5 permiten estudiar la convergencia de los diferentes parámetros de interés, las cadenas realizan una correcta exploración del espacio de estados. Con base es los paneles g y h se realiza el estudio de convergencia para z_i, para ambos casos se presenta un comportamiento simétrico al rededor del valor teórico. La Figura 5.19 presenta la superficie estimada, tiene un comportamiento suave, con máximos valores cuando el puntaje político toma valores altos y el logaritmo de la población valores bajos.

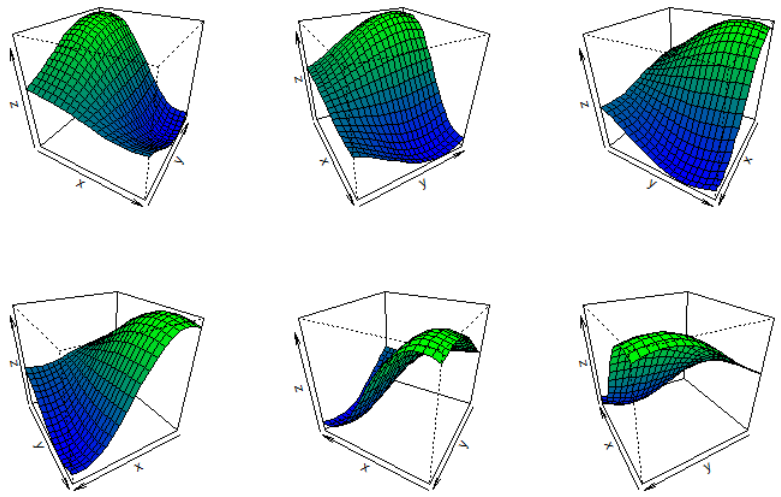
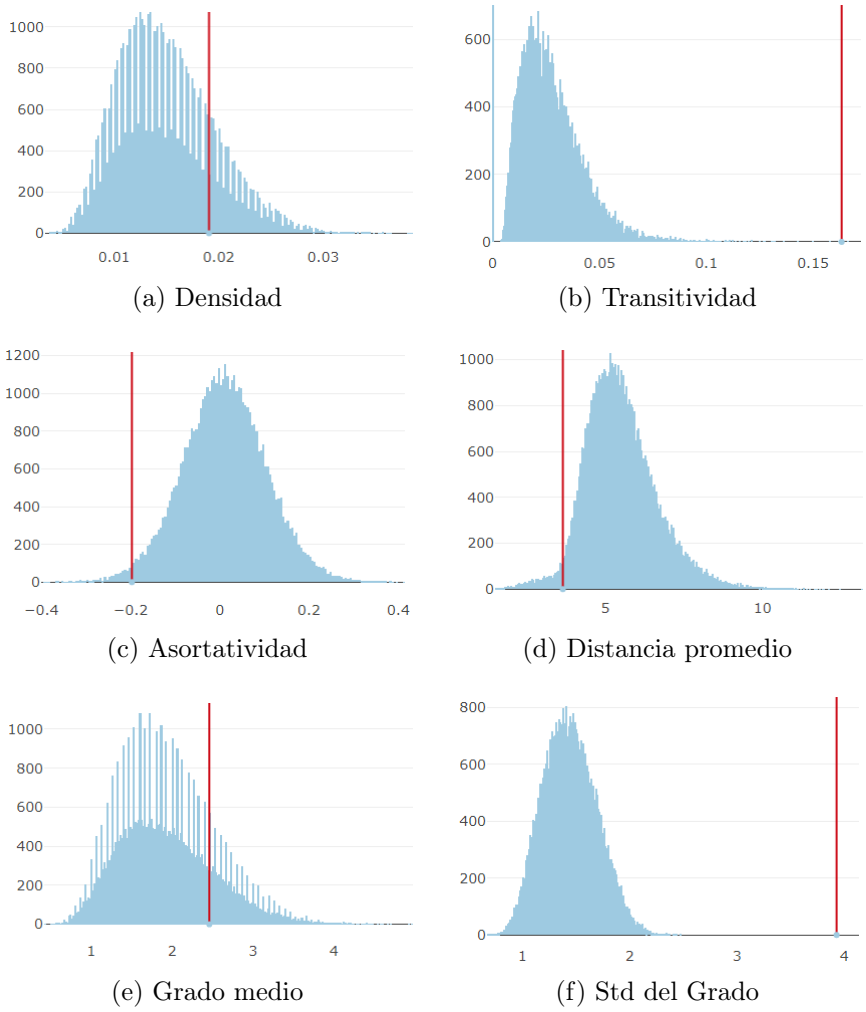


Figura 5.19.: Superficie estimada de los datos de conflicto.

5.3.3. Bondad de Ajuste

Las estadísticas para estudiar corresponden a la densidad, la transitividad, la asortatividad, la distancia promedio, el grado promedio de los vértices y su desviación estándar. La Figura 5.20 muestra que el modelo presenta problemas para replicar la transitividad, la asortatividad y la desviación estándar del grado de los vértices.



(a) Densidad (b) Transitividad

(c) Asortatividad (d) Distancia promedio

(e) Grado medio (f) Std del Grado

Figura 5.20.: Estadísticas para la evaluación de bondad de ajuste del modelo de los datos de conflicto.

5.3.4. Predicción

La Figura 5.21 presenta las curvas ROC para cada grupo del procedimiento de validación cruzada. Se presenta un desempeño bueno de la calidad predictiva del modelo dado que los valores se encuentran entre 0.65 y 0.75. El vector de AUC de los grupos es igual a $(0.69, 0.69, 0.65, 0.70, 0.73)$. El AUC estimado del modelo es igual a 0.70.

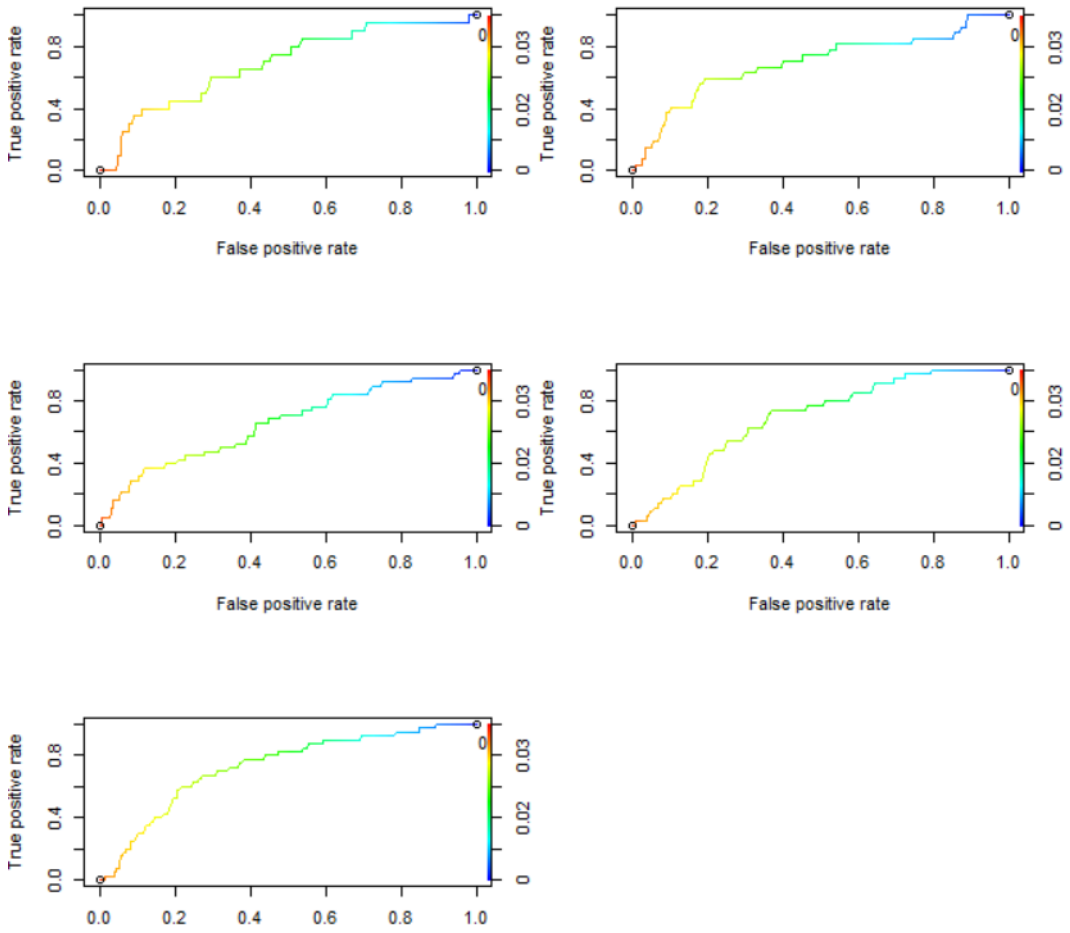


Figura 5.21.: Curvas ROC del procedimiento de validación cruzada de los datos de conflicto

5.4. Datos de Krackhardt

5.4.1. Análisis descriptivo de la red

La red está compuesta por las mediciones de la relación "¿quién es amigo de X?" para evaluar los efectos de un programa de intervención sobre la estructura social cognitiva del personal en una empresa de fabricación de maquinaria de alta tecnología estudiada por David Krackhardt. El orden de la red es de 21. Los atributos evaluados son la edad y la antigüedad.

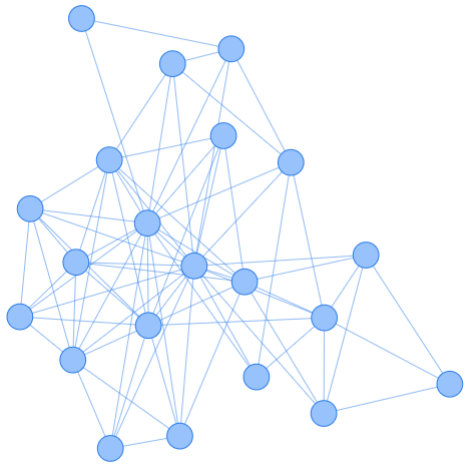


Figura 5.22.: Red de los datos de Krackhardt.

En la Figura 5.22 se presenta la red de Krackhardt. La Figura 5.23 presenta el grado de los vértices versus el promedio del grado de sus vecinos. Los vértices con mayor grado se agrupan con vértices de grado menor. Los vértices 2, 18 y 14 tienen los valores más altos del grado. De manera análoga sucede para la centralidad por intermediación. La distancia media de la red es 1.7095. Su diámetro y transitividad son 4 y 0.5694, respectivamente. En la Figura 5.24 se presenta la partición vía métodos jerárquicos. La densidad de la red es de 0.3667.

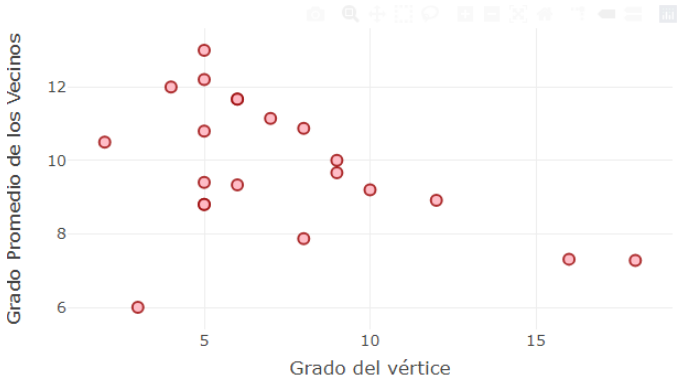


Figura 5.23.: Grado de los vértices de los datos de Krackhardt vs el promedio del grado de sus vecinos.

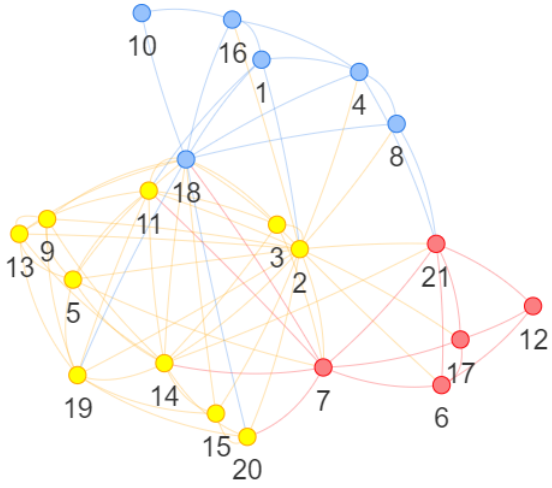


Figura 5.24.: Partición vía métodos jerárquicos de la red de los datos de Krackhardt.

5.4.2. Modelamiento

Los paneles a, b, c, d, e y f de la Figura A.6 permiten estudiar la convergencia de los diferentes parámetros de interés, las cadenas realizan una correcta exploración del espacio de estados. Con base es los paneles g y h se realiza el estudio de convergencia para z_i, para ambos casos se presenta un comportamiento simétrico al rededor del valor teórico. La Figura 5.25 presenta la superficie estimada.

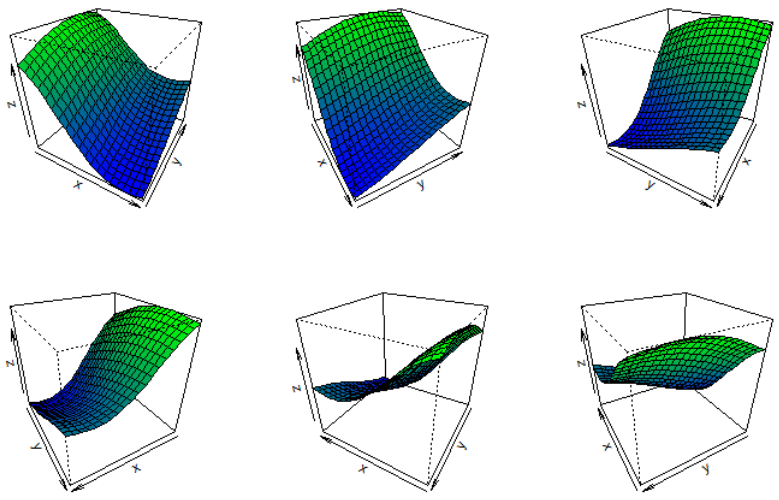
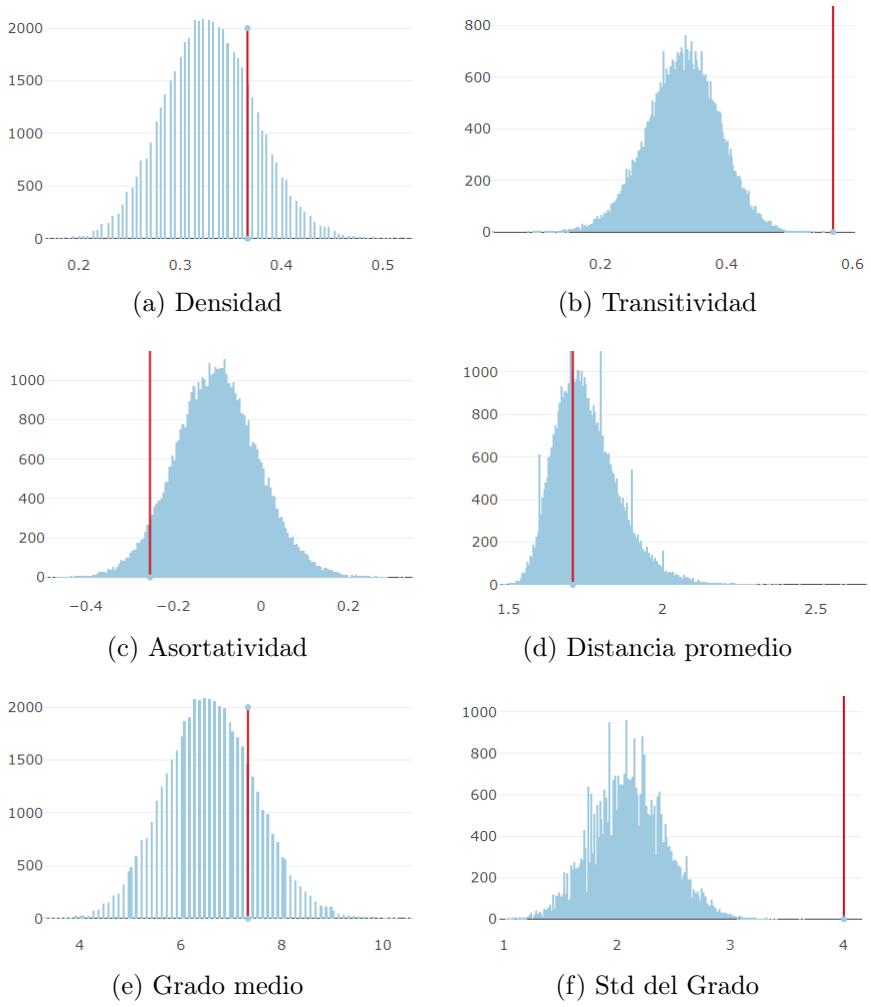


Figura 5.25.: Superficie estimada de los datos de Krackhardt.

5.4.3. Bondad de Ajuste

Las estadísticas para estudiar corresponden a la densidad, la transitividad, la asortatividad, la distancia promedio, el grado promedio de los vértices y su desviación estándar. La Figura 5.26 muestra que el modelo replica de manera correcta las estadísticas de interés con excepción de la transitividad y la desviación estándar del grado de los vértices.



(a) Densidad (b) Transitividad

(c) Asortatividad (d) Distancia promedio

(e) Grado medio (f) Std del Grado

Figura 5.26.: Estadísticas para la evaluación de bondad de ajuste del modelo de los datos Krackhardt.

5.4.4. Predicción

La Figura 5.27 presenta las curvas ROC para cada grupo del procedimiento de validación cruzada. Se presenta un desempeño regular de la calidad predictiva del modelo dado que los valores se encuentran entre 0.4 y 0.65. El vector de AUC de los grupos es igual a

$(0.42, 0.58, 0.53, 0.53, 0.63)$. El AUC estimado del modelo es igual a 0.54.

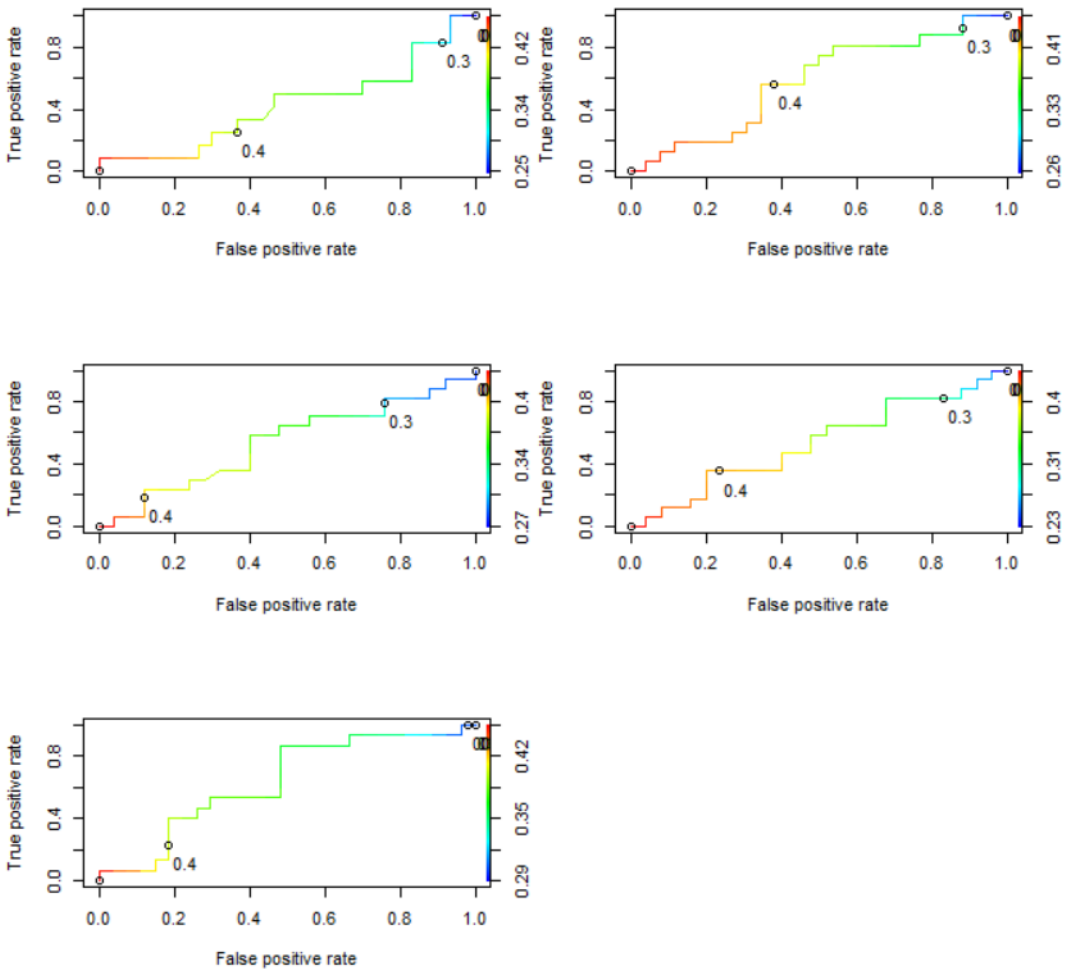


Figura 5.27.: Curvas ROC del procedimiento de validación cruzada de los datos de Krackhardt.

5.5. Consolidación de los Resultados

La calidad de la bondad de ajuste del modelo varía según el conjunto de datos estudiado. Se sugiere al analista realizar las inferencias cuidadosamente, sobre todo para las estadísticas que no fueron replicadas correctamente por el modelo. En términos de poder predictivo hay mucha fluctuación de los valores AUC entre modelos, también dentro de las particiones en los ejercicios de validación cruzada, lo que supone que las superficies a estimar posiblemente están dadas por representaciones funcionales complejas. La tabla 5.1 presenta el resumen de las medidas de bondad predictivo del modelo para los cuatro conjuntos de datos reales.

Modelo	Orden	Tamaño	AUC
Zachary	34	78	0.66
Lazega	34	115	0.57
Conflicto	130	160	0.70
Krackhardt	21	77	0.54

Tabla 5.1.: AUC estimados para los modelos de datos reales.

6. Aproximación Casos y Controles de la Verosimilitud

El costo computacional del modelo de espacio latente presentado en la Sección (3.1.1) para una red de n vértices es de $O(n^2)$, lo que hace su uso inviable en el caso de redes grandes.
Raftery et al. (2012) reemplaza la función de verosimilitud completa en el procedimiento MCMC por una estimación insesgada obtenida a través del enfoque epidemiológico de casos y controles. Esto reduce el costo computacional a $O(n)$.
En la epidemiología, el enfoque de casos y controles es ampliamente usado. Los casos corresponden a individuos portadores de una enfermedad o condición médica especial, en consecuencia, su medición es costosa o difícil. Por el contrario, los controles son fáciles de obtener.
En el contexto de redes, los casos corresponden a la realización de un enlace, la ausencia de un enlace corresponde a un control. Es de interés determinar qué factores medibles o latentes intervienen en la realización de un enlace. Para este fin se re-escribe la función de log-verosimilitud de la siguiente manera:

$$l = \mathsf{log}(\mathsf{p}(\boldsymbol{y} \mid \boldsymbol{\pi})) = \sum_{i=1}^{N} l_i \tag{6.1}$$

donde

$$\begin{aligned} l_i &= \sum_{i \neq j} \{\eta_{i,j} y_{i,j} - \mathsf{log}(1 + \mathsf{exp}\{\eta_{i,j}\})\} \\ &= \sum_{j \neq i, y_{i,j}=1} \{\eta_{i,j} - \mathsf{log}(1 + \mathsf{exp}\{\eta_{i,j}\})\} + \sum_{j \neq i, y_{i,j}=0} \{-\mathsf{log}(1 + \mathsf{exp}\{\eta_{i,j}\})\} \\ &= l_{i,1} + l_{i,0} \end{aligned} \tag{6.2}$$

Una estimación puntual de $l_{i,0}$ puede ser obtenida a través del estimador insesgado de un total para un muestreo aleatorio simple. Dados $N_{i,0}$ y $n_{i,0}$ el número total de vértices para los que $y_{i,j} = 0$ y su tamaño de muestra, respectivamente. Considerando $n_{i,0}$ suficientemente pequeño para reducir de manera significativa el tiempo computacional, el estimador puntual $\tilde{l}_{i,0}$ se se define como:

$$\tilde{l}_{i,0} = \frac{N_{i,0}}{n_{i,0}} \sum_{k=1}^{n_{i,0}} \{\mathsf{log}(1 + \mathsf{exp}\{\eta_{i,j}\})\} \tag{6.3}$$

El término de la derecha en la Ecuación (6.2) también puede ser aproximado a través de un muestreo estratificado de M estratos. Los estratos son definidos a partir del concepto de cercanía entre vértices, lo que lleva a la siguiente descomposición de la Ecuación (6.1),

$$\begin{aligned} l_i = &\sum_{j \neq i, y_{i,j}=1} \{\eta_{i,j} y_{i,j} - \mathsf{log}(1 + \mathsf{exp}\{\eta_{i,j}\})\} + \sum_{j:D_{i,j}=2} \{\mathsf{log}(1 + \mathsf{exp}\{\eta_{i,j}\})\} \\ &+ \cdots + \sum_{j:D_{i,j}=M} \{\mathsf{log}(1 + \mathsf{exp}\{\eta_{i,j}\})\} \end{aligned}$$

un estimador puntual insesgado para l_i basado en un muestreo estratificado se define como:

$$\hat{l}_i = \sum_{j \neq i, y_{i,j}=1} \{\eta_{i,j} y_{i,j} - \mathsf{log}(1 + \mathsf{exp}\{\eta_{i,j}\})\} + \sum_{h=2}^{M} \frac{N_{i,h}}{n_{i,h}} \sum_{j:Di,j=h} \{\mathsf{log}(1 + \mathsf{exp}\{\eta_{i,j}\})\} \tag{6.4}$$

donde $N_{i,h}$ es el número de vértices j con $D_{i,j} = h$, $n_{i,j}$ es el número de vértices j muestreados con $D_{i,j} = h$ y $D_{i,j}$ la distancia geodésica para un par de vértices i y j.

6.1. Cálculo de $n_{i,h}$

1. Determinar r con base en la siguiente igualdad $n_{i,0} = r\bar{d} = n_0$, donde $\bar{d}$ es el grado medio de la red.

2. Ejecutar un piloto del MCMC con muestreo aleatorio simple de tamaño $n_{i,0}$

   a. Dado el periodo de calentamiento y $z_i^{(t)*}$ un nuevo valor propuesto para z_i, en cada iteración $1 \leq t \leq T$ calcular:

$$\begin{aligned} \Delta\tilde{l}_i^{(t)} &\equiv \tilde{l}_i(z_i^{(t)*}) - \tilde{l}_i(z_i^{(t)}) \\ &= l_{i,1}(z_i^{(t)*}) - l_{i,1}(z_i^{(t)}) + \sum_h \{\tilde{l}_{i,h}(z_i^{(t)*}) - \tilde{l}_{i,h}(z_i^{(t)})\} \\ &\equiv \Delta l_{i,1}^{(t)} + \sum_h \{\Delta\tilde{l}_{i,h}^{(t)}\} \end{aligned}$$

   b. Definir

$$w_{i,h}^{(t)} = \mid \Delta\tilde{l}_{i,h}^{(t)} / \sum_{g=2}^{M} \Delta\tilde{l}_{i,g} \mid$$

c. Obtener

$$w_{i,h} = \frac{1}{T-1}\sum_{t=1}^{T-1} w_{i,h}^{(t)}$$

3. Fijar el tamaño del estrato h para el vértice i como

$$n_{i,h} = \frac{n_{i,0} w_{i,h}}{\sum_{g=2}^{M} w_{i,g}}$$

6.2. Algoritmo

Dada las estimaciones puntuales para $l_{i,0}$ y l_i, respectivamente presentadas en la Ecuación (6.3) y la Ecuación (6.4), el algoritmo de estimación vía MCMC se define como:

1. Ejecutar un piloto del MCMC.
   a. Definir $n_{i,0}$ para cada vértice $i = 1 \cdots N$.
   b. Dada la estimación $\tilde{l}_{i,0}$ para $l_{i,0}$ en la Ecuación (6.3) ejecutar el algoritmo MCMC reemplazando l por $\tilde{l} = \sum_i^N l_{i,1} + \tilde{l}_{i,0}$.
2. Para cada distancia geodésica $h = 2, \cdots, M$ y cada vértice i con $i = 1, \cdots, n$:
   a. Determinar $n_{i,h}$
   b. Muestrear $n_{i,h}$ vértices con $y_{i,j} = 0$ y distancia geodésica $D_{i,j} = h$.
3. Ejecutar el MCMC completo reemplazando la log-verosimilitud usual por su estimación presentada en la Ecuación (6.4), $\hat{l}$.

El estimador en la Ecuación (6.4) es un estimador puntual insesgado indistintamente del valor r, el cual es seleccionado arbitrariamente por el investigador. Diferentes valores de r llevarán a estimadores más o menos eficientes.

6.3. Estudio de Simulación y Aplicación en Datos de la Literatura

6.3.1. Primer Conjunto de Datos Simulados

Modelamiento y Bondad de Ajuste

La Figura A.7 permite estudiar la convergencia de los parámetros de interés se observa una buena exploración del espacio de estados de las cadenas. Para el estudio de los z_i se

presentan las gráficas g y h se observa un comportamiento simétrico al rededor del valor teórico. La Figura 6.1 muestra la superficie estimada, guarda similitud con la superficie teórica presentada en la Figura 4.2. La Figura 6.2 permite concluir que el modelo replica adecuadamente todos los estadísticos de interés de la red observada.

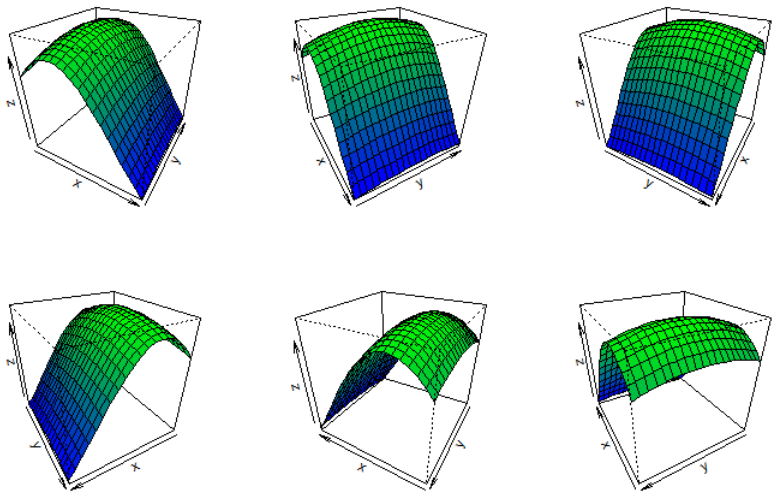
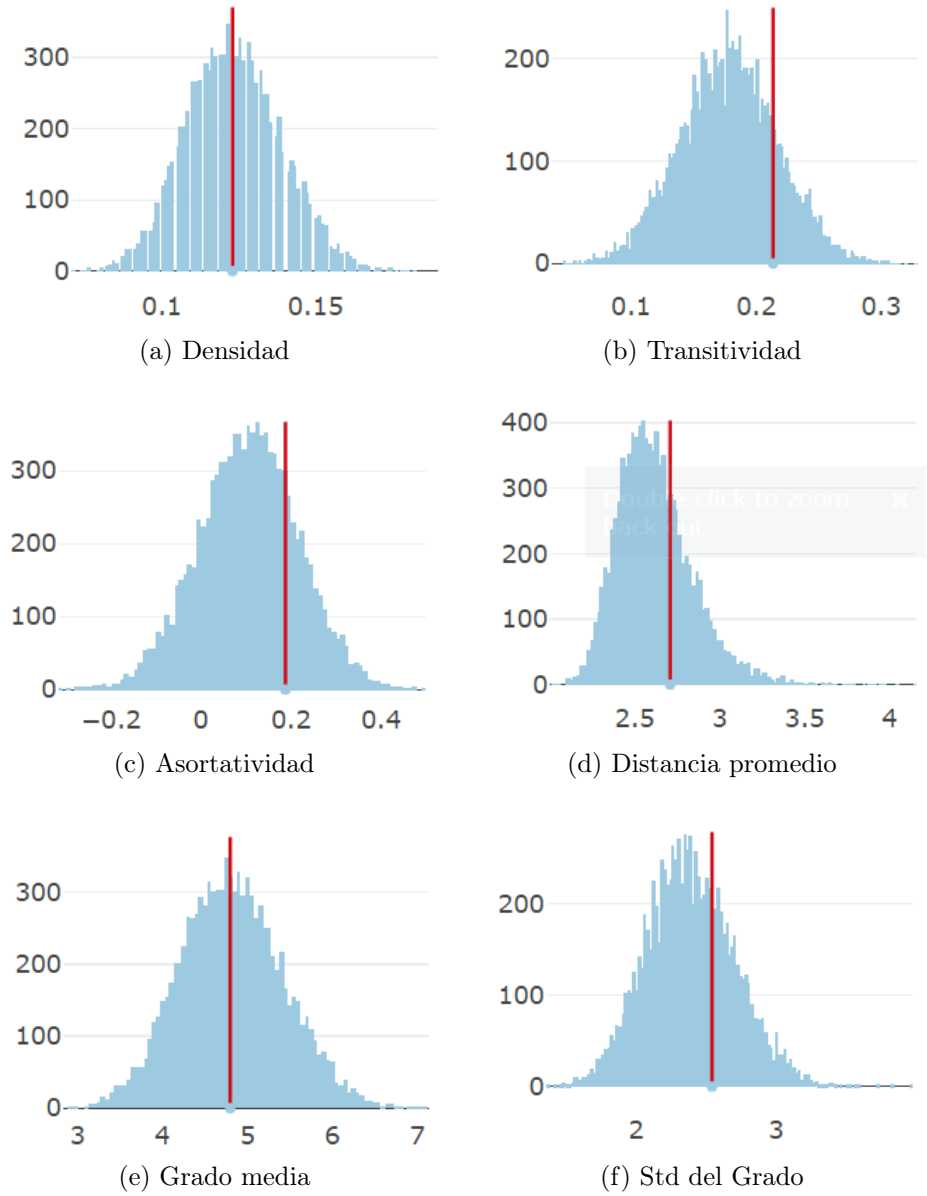


Figura 6.1.: Superficie estimada de los datos del primer caso de simulación para el modelo inspirado en la metodología de casos y controles.

6.3.2. Datos de Zachary

Modelamiento y Bondad de Ajuste

Las Figuras 6.4 y A.8 permiten estudiar la convergencia de las cadenas junto con la capacidad del modelo de replicar las estadísticas de interés de la red. Los resultados son similares a los presentados en las Figuras 5.5 y A.3.



(a) Densidad (b) Transitividad

(c) Asortatividad (d) Distancia promedio

(e) Grado media (f) Std del Grado

Figura 6.2.: Estadísticas para la evaluación de bondad de ajuste del modelo usando la metodología de casos y controles para los datos del primer caso de simulación.

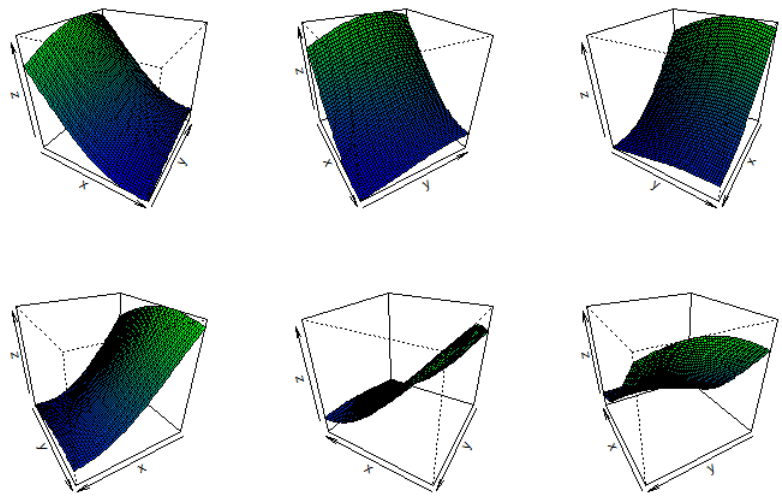
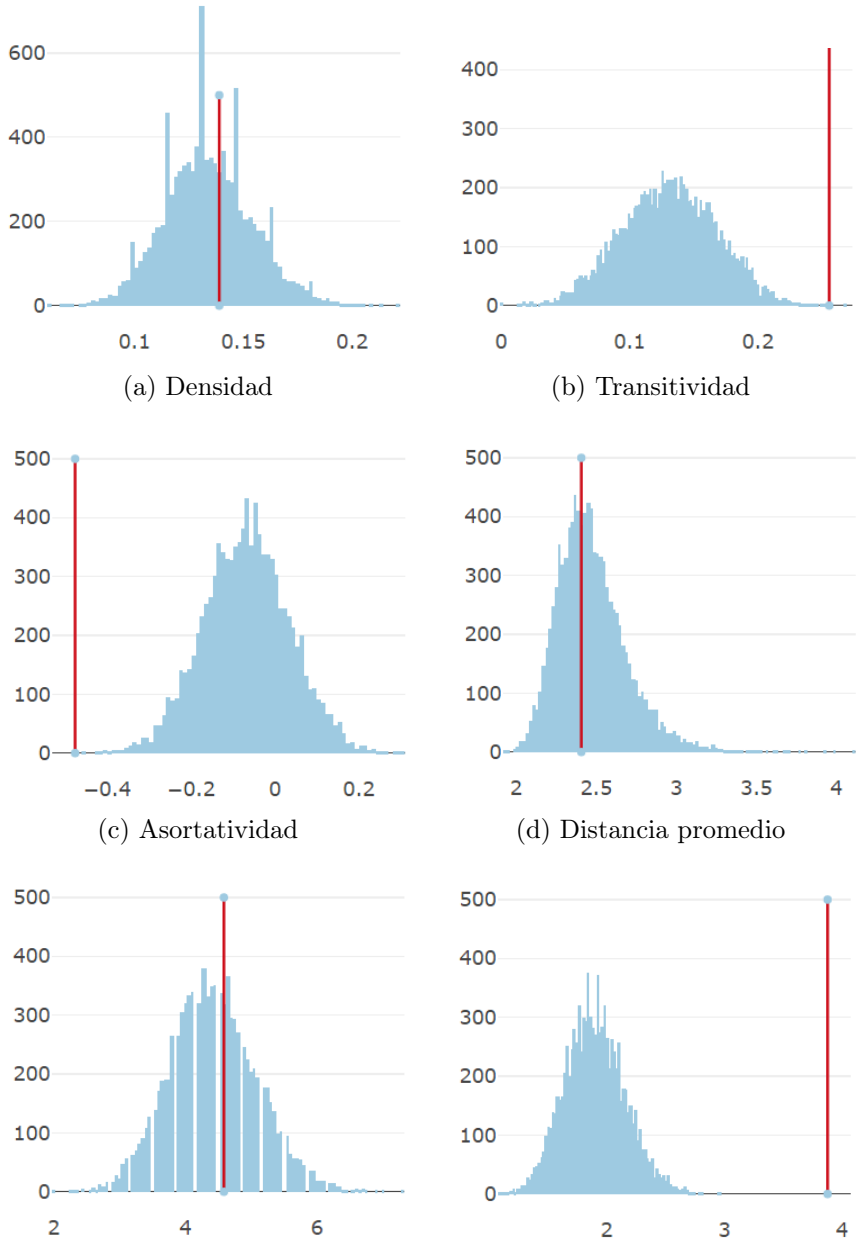


Figura 6.3.: Superficie estimada de los datos Zachary para el modelo inspirado en la metodología de casos y controles.



(a) Densidad (b) Transitividad

(c) Asortatividad (d) Distancia promedio

(e) Grado medio (f) Std del Grado

Figura 6.4.: Estadísticas para la evaluación de bondad de ajuste del modelo usando la metodología de casos y controles para los datos de Zachary.

7. Discusión

Los hallazgos muestran que la exploración del espacio de estados del modelo es adecuada para todos los casos de estudio desarrollados. El estudio de bondad de ajuste del modelo para los datos sintéticos presenta buenos resultados replicando todas las estadísticas de interés. La superficie asociada al segundo conjunto de datos sintéticos es más compleja comparada con el primer conjunto de datos sintético, se supone un impacto de la complejidad de la superficie sobre el poder predictivo del modelo. Se observa que la superficie estimada para el segundo conjunto de datos simulados presenta diferencias con respecto a la real y el AUC del modelo es igual a 0.59. En contraste, para el primer conjunto de datos sintéticos se tiene una estimación más precisa de la superficie y se obtiene que el AUC del modelo es igual a 0.76.

Hay deficiencias en la bondad de ajuste y en el poder predictivo del modelo para los ejercicios sobre datos reales, esto puede deberse a que la superficie a estimar es demasiado compleja o que las covariables de estudio no se relacionan lo suficiente con el fenómeno de interés. Se plantea el estudio de otras redes reales con diferentes tamaños y estructuras (e.g., Sosa and Buitrago 2021) y un estudio de simulación más exhaustivo para indagar con mayor profundidad las capacidades predictivas del modelo.

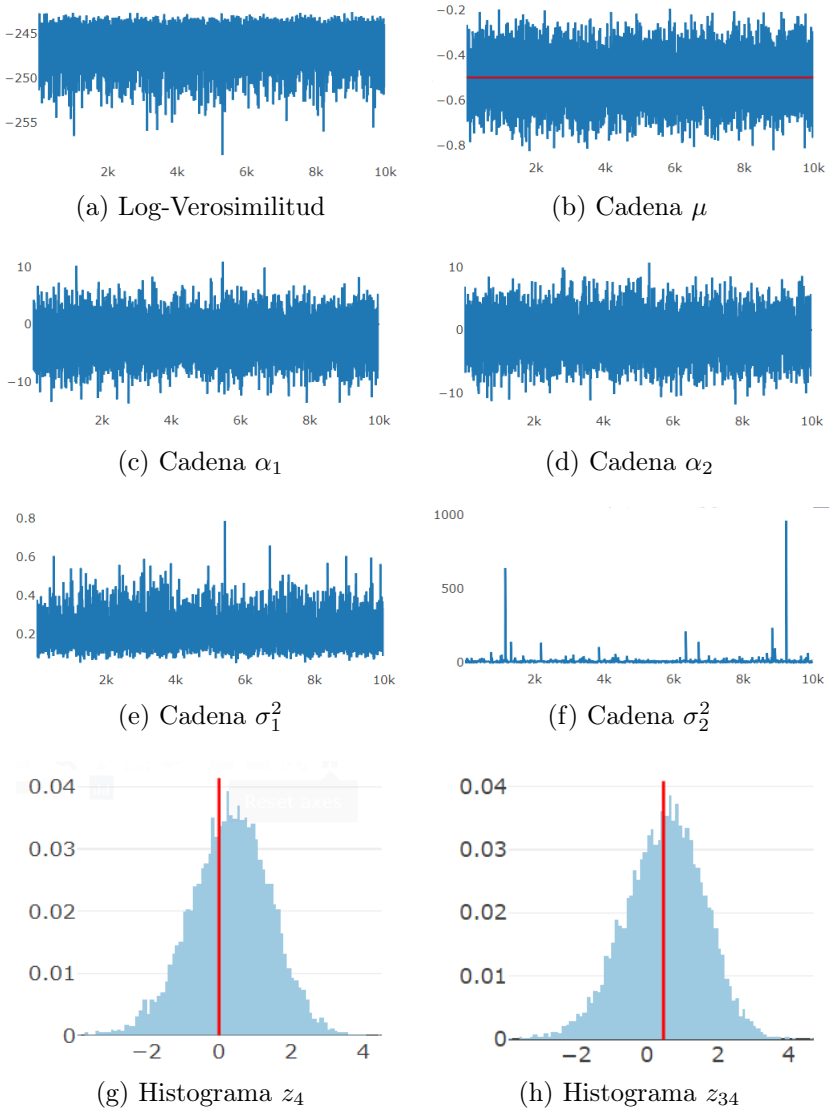
En este documento se realiza por primera vez la implementación de la propuesta de muestreo Raftery et al. 2012 aplicada al modelo de Proceso Socio-Espacial Latente de Linkletter 2007. Se realiza una exploración de la bondad de ajuste del modelo utilizando el primer conjunto de datos sintéticos y el conjunto de datos reales de Zachary. Los resultados obtenidos bajo este enfoque son equivalentes a los obtenidos utilizando la función de verosimilitud completa. Se recomienda esta metodología cuando se tengan conjuntos de datos grandes ya que reproduce correctamente la propuesta original y reduce los tiempos de estimación considerablemente. A futuro queda estudiar las capacidades predictivas del modelo bajo este enfoque.

Finalmente, como trabajo futuro es de interés comparar la alternativa de Ciminelli et al. 2019 con la de Linkletter 2007 en términos tanto de bondad de ajuste como de predicción. Esta propuesta captura la correlación espacial de las redes sociales, modela conjuntamente los atributos de los vértices y, basado en los atributos y las conexiones de los vértices, modela sus ubicaciones latentes en el Espacio Social y el Proceso Espacial subyacente. También es de interés investigar el desempeño del modelo utilizando la estimación de la función de verosimilitud basada en la propuesta de casos y controles de Raftery et al. 2012

de forma análoga a como se implementó con el modelo de Linkletter 2007.

A. Anexo: Estudio de convergencia

Estudio de convergencia de los datos del primer caso de simulación



(a) Log-Verosimilitud (b) Cadena μ

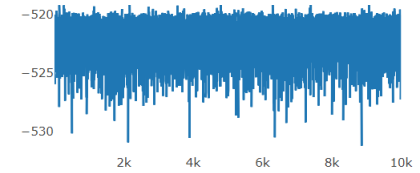
(c) Cadena α_1 (d) Cadena α_2

(e) Cadena σ_1^2 (f) Cadena σ_2^2

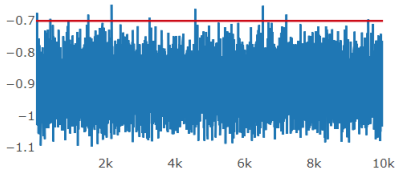
(g) Histograma z_4 (h) Histograma z_{34}

Figura A.1.: Convergencia del algoritmo MCMC de los datos del primer caso de simulación.

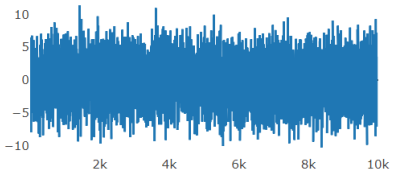
Estudio de convergencia de los datos del segundo caso de simulación



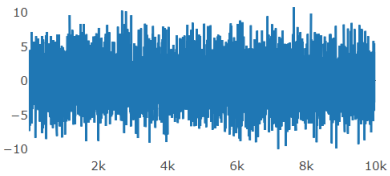
(a) Log-Verosimilitud



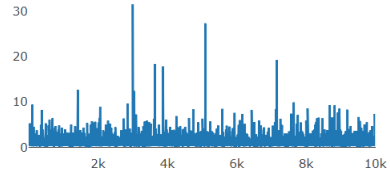
(b) Cadena μ



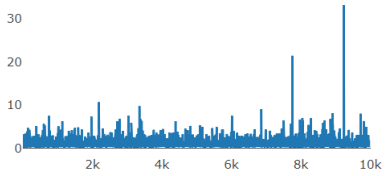
(c) Cadena α_1



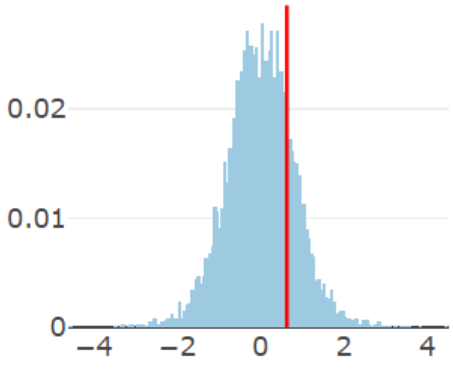
(d) Cadena α_2



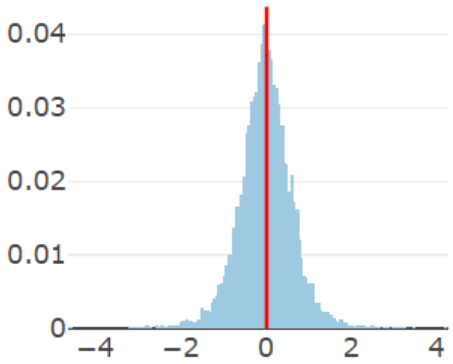
(e) Cadena σ_1^2



(f) Cadena σ_2^2



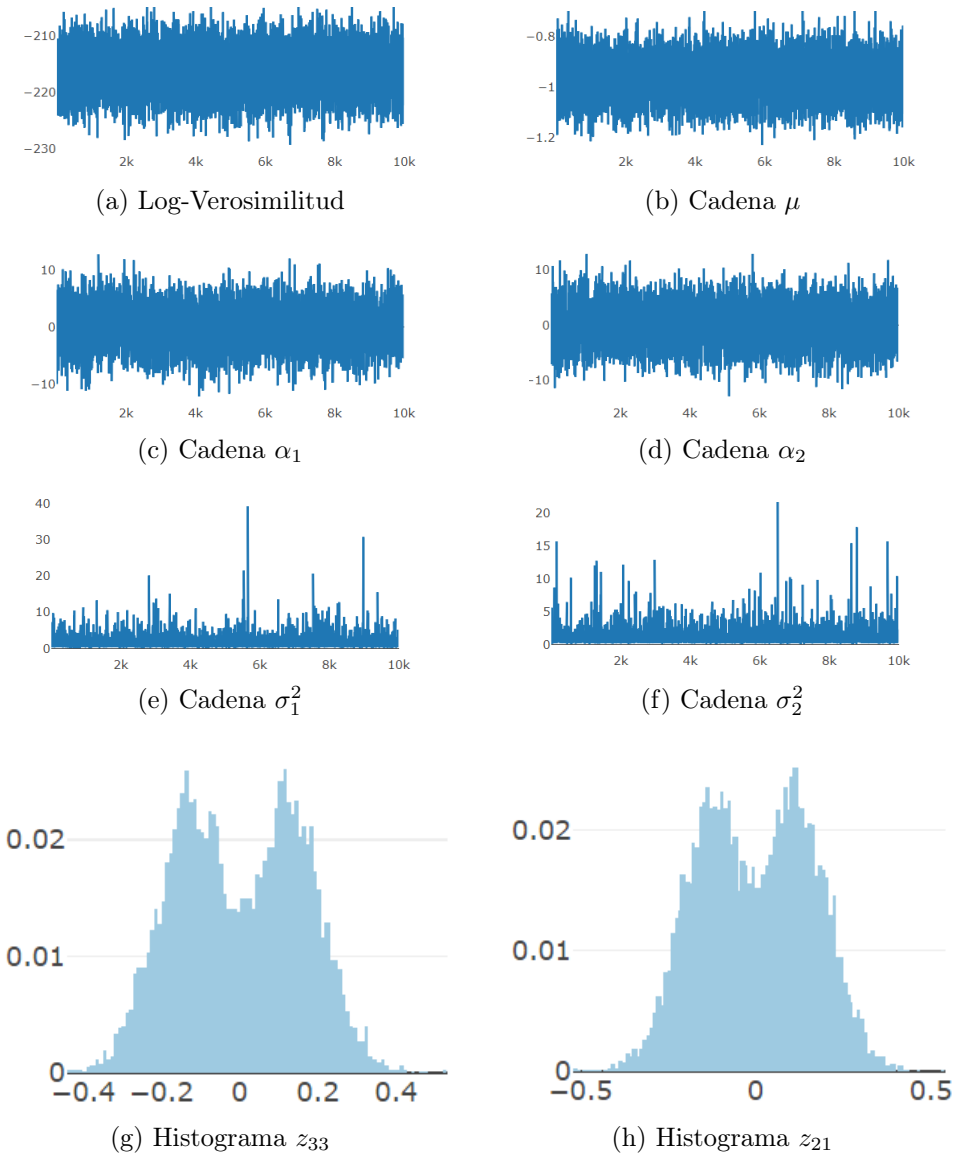
(g) Histograma z_{23}



(h) Histograma z_{43}

Figura A.2.: Convergencia del algoritmo MCMC de los datos del segundo caso de simulación.

Estudio de convergencia de los datos de Zachary



(a) Log-Verosimilitud

(b) Cadena μ

(c) Cadena α_1

(d) Cadena α_2

(e) Cadena σ_1^2

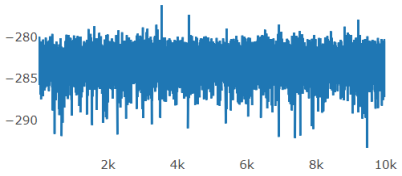
(f) Cadena σ_2^2

(g) Histograma z_{33}

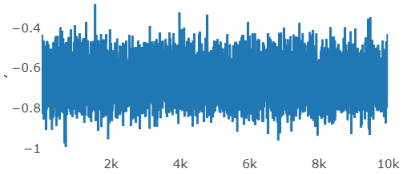
(h) Histograma z_{21}

Figura A.3.: Convergencia del algoritmo MCMC de los datos de Zachary.

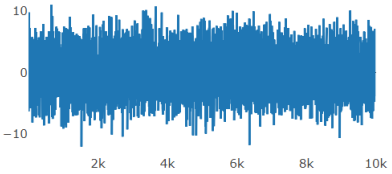
Estudio de convergencia de los datos de Lazega



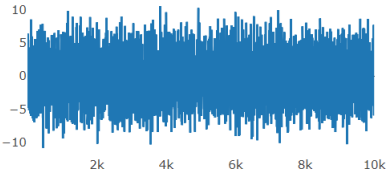
(a) Log-Verosimilitud



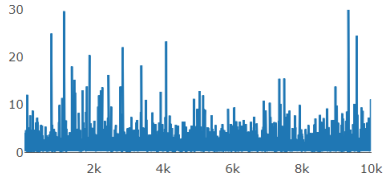
(b) Cadena μ



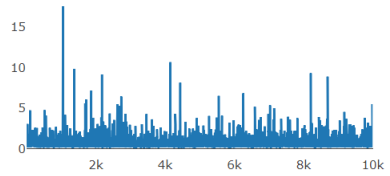
(c) Cadena α_1



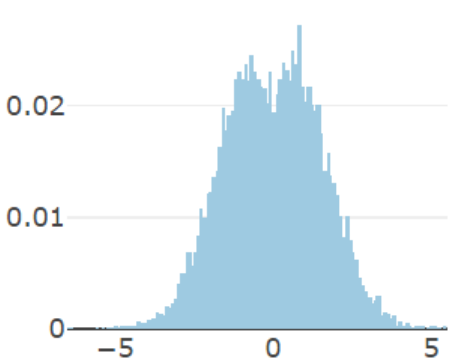
(d) Cadena α_2



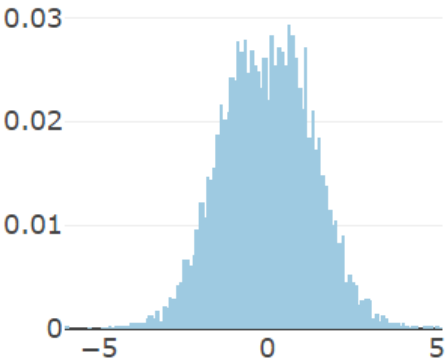
(e) Cadena σ_1^2



(f) Cadena σ_2^2



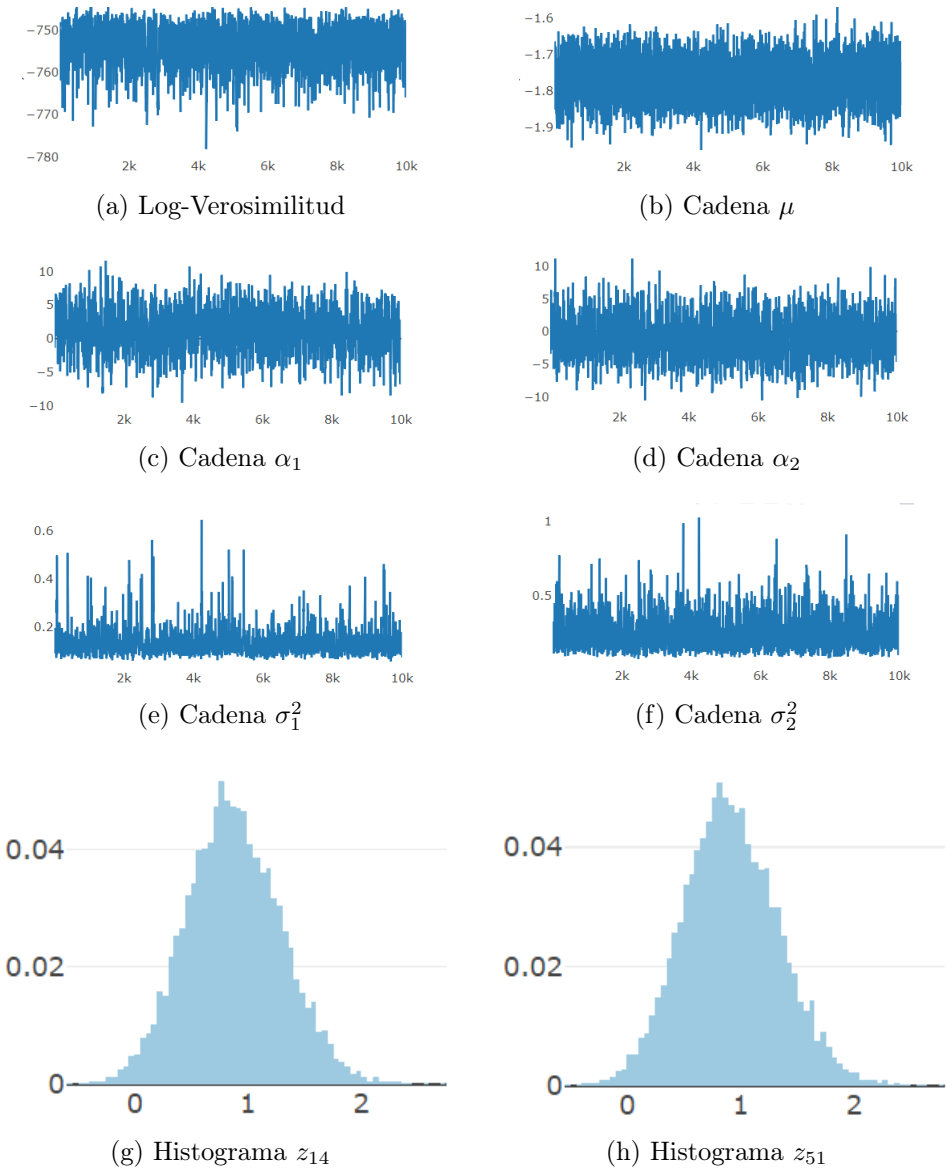
(g) Histograma z_{21}



(h) Histograma z_{10}

Figura A.4.: Convergencia del algoritmo MCMC de los datos de Lazega.

Estudio de convergencia de los datos de conflicto



(a) Log-Verosimilitud

(b) Cadena μ

(c) Cadena α_1

(d) Cadena α_2

(e) Cadena σ_1^2

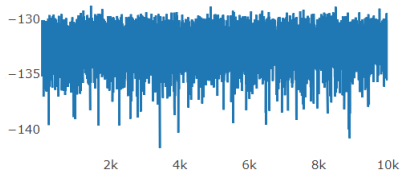
(f) Cadena σ_2^2

(g) Histograma z_{14}

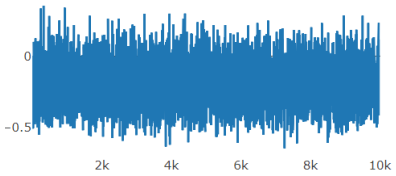
(h) Histograma z_{51}

Figura A.5.: Convergencia del algoritmo MCMC de los datos de conflicto.

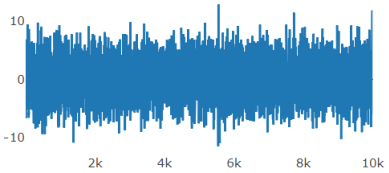
Estudio de convergencia de los datos de Krackhardt



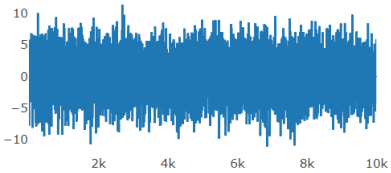
(a) Log-Verosimilitud



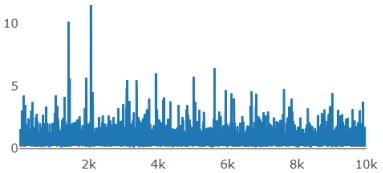
(b) Cadena μ



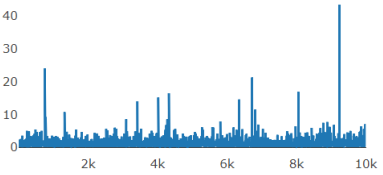
(c) Cadena α_1



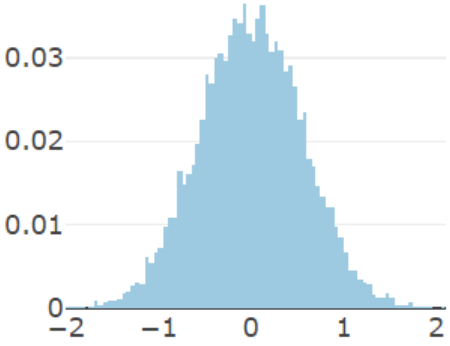
(d) Cadena α_2



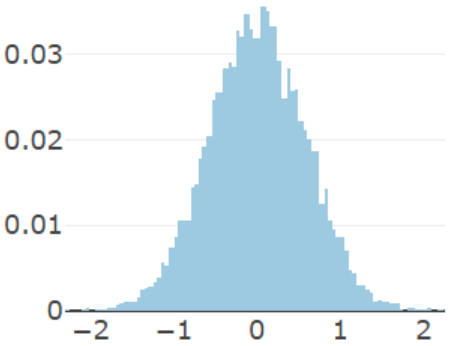
(e) Cadena σ_1^2



(f) Cadena α_2^2



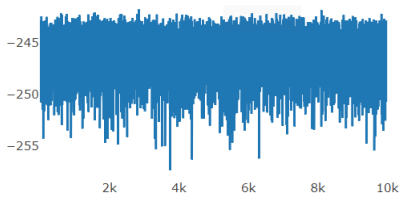
(g) Histograma z_{11}



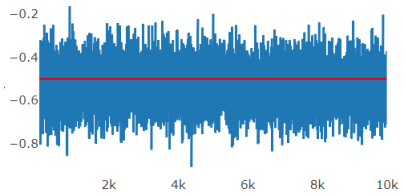
(h) Histograma z_{14}

Figura A.6.: Convergencia del algoritmo MCMC de los datos de Krackhardt.

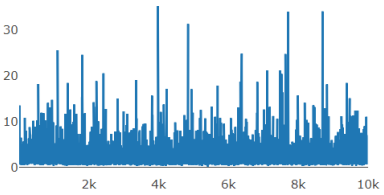
Estudio de convergencia de los datos del primer caso de simulación (Raftery)



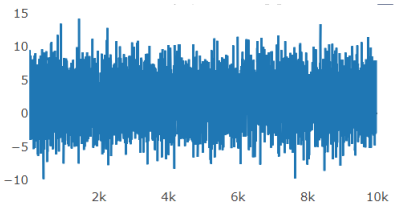
(a) Log-Verosimilitud



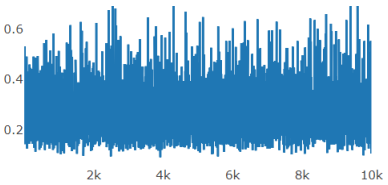
(b) Cadena μ



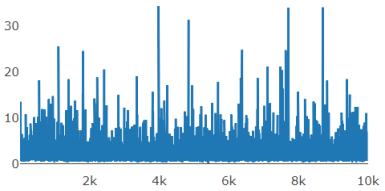
(c) Cadena α_1



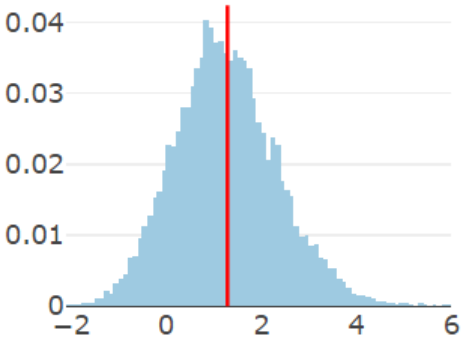
(d) Cadena α_2



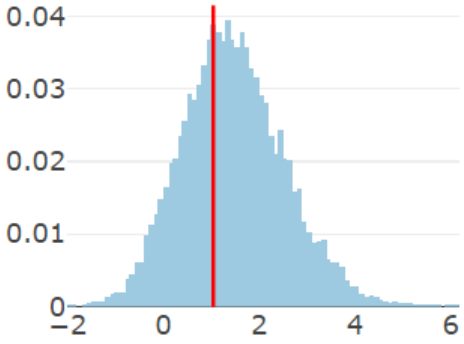
(e) Cadena σ_1^2



(f) Cadena α_2^2



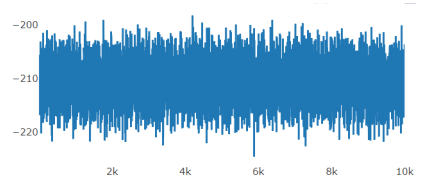
(g) Histograma z_{11}



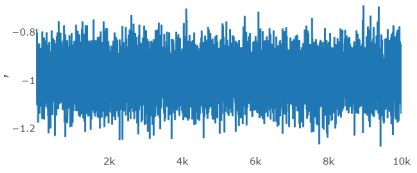
(h) Histograma z_{14}

Figura A.7.: Convergencia del algoritmo MCMC modificado de los datos del primer conjunto de simulación.

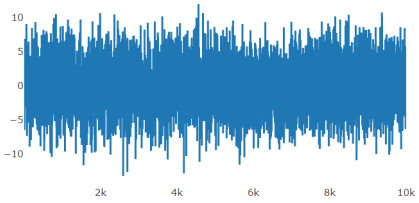
Estudio de convergencia de los datos de Zachary (Raftery)



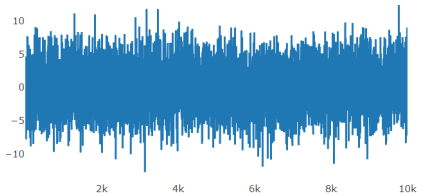
(a) Log-Verosimilitud



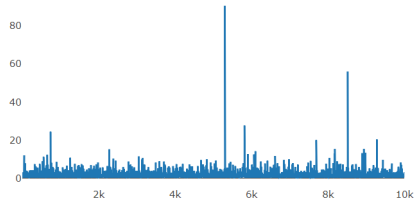
(b) Cadena μ



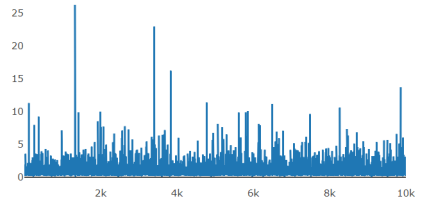
(c) Cadena α_1



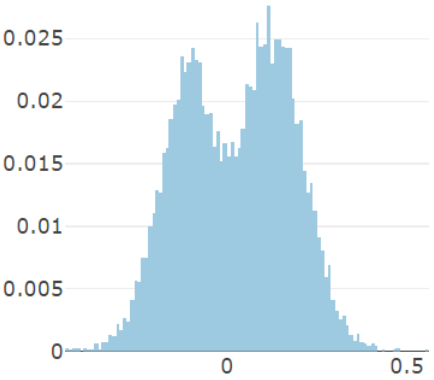
(d) Cadena α_2



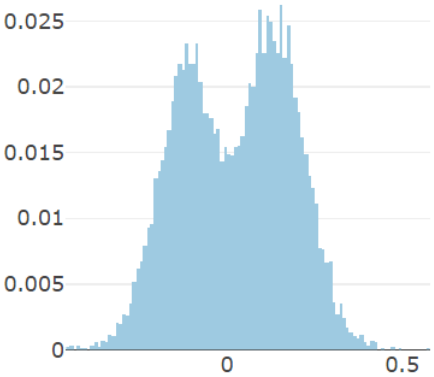
(e) Cadena σ_1^2



(f) Cadena α_2^2



(g) Histograma z_{11}



(h) Histograma z_{14}

Figura A.8.: Convergencia del algoritmo MCMC modificado de los datos de Zachary.

B. Anexo: Algoritmo

Reemplazando las Ecuaciones (3.7) y (3.8) en (3.9), la distribución posterior conjunta está dada por:

$$\begin{aligned}\mathsf{p}(\mu, \boldsymbol{\alpha}, \boldsymbol{\rho} \mid \boldsymbol{y}) &= \mathsf{p}(\boldsymbol{y} \mid \mu, \boldsymbol{\alpha}, \boldsymbol{\rho})\mathsf{p}(\mu)\mathsf{p}(\boldsymbol{\alpha})\mathsf{p}(\boldsymbol{\rho}) \\ &= \prod_{i<j}^{n} \pi_{i,j}^{y_{i,j}}(1-\pi_{i,j})^{1-y_{i,j}} \times \frac{e^{-\frac{\mu^2}{2\psi_\mu}}}{\sqrt{2\pi\psi_\mu}} \times \prod_{k=1}^{m} \frac{e^{-\frac{\alpha_k^2}{2}}}{\sqrt{2\pi}} \times \prod_{h=1}^{p} 1\end{aligned}$$

El algoritmo procede a generar un nuevo estado $\boldsymbol{\theta}^{(b+1)}$ a partir de $\boldsymbol{\theta}^{(b)}$ para $b = 1, \ldots, B$, como sigue:

1. Muestrear $\mu^{(b+1)}$, de acuerdo al algoritmo Metropolis-Hastings, considerando la distribución condicional completa:

$$\mathsf{p}(\mu \mid \boldsymbol{resto}) \propto \prod_{i<j}^{n} \pi_{i,j}^{y_{i,j}}(1-\pi_{i,j})^{1-y_{i,j}} \times \mathsf{exp}\{-\frac{\mu^2}{2\psi_\mu}\}$$

2. Muestrear $\alpha_k^{(b+1)}$, de acuerdo al algoritmo Metropolis-Hastings, considerando la distribución condicional completa:

$$\mathsf{p}(\alpha_k \mid \boldsymbol{resto}) \propto \prod_{i<j}^{n} \pi_{i,j}^{y_{i,j}}(1-\pi_{i,j})^{1-y_{i,j}} \times \mathsf{exp}\{-\frac{\alpha_k^2}{2}\} \quad \text{para} \quad k = 1, \ldots, m.$$

3. Muestrear $\rho_d^{(b+1)}$, de acuerdo al algoritmo Metropolis-Hastings, considerando la distribución condicional completa:

$$\mathsf{p}(\rho_h \mid \boldsymbol{resto}) \propto \prod_{i<j}^{n} \pi_{i,j}^{y_{i,j}}(1-\pi_{i,j})^{1-y_{i,j}} \quad \text{para} \quad h = 1, \ldots, p.$$

Para todo $\ell = 1 \ldots p$ se asume $\mathsf{q}(\theta_\ell^* \mid \theta_\ell^{(b-1)}) \sim N(\theta_\ell^{(b-1)}, \delta_\ell)$, δ_ℓ se calibra con base al siguiente algoritmo.

Sea λ_ℓ la tasa de aceptación del parámetro θ_ℓ. Para la iteración b del algoritmo MCMC, sí $b \quad \text{módulo} \quad 100 = 0$ y si $\mid \lambda_\ell - 0.375 \mid > 0.075$:

1. Iniciar $\kappa = 1$.
2. Calcular $\delta_\ell^* = \delta_\ell + (0.1/\kappa)(\lambda_\ell - 0.375)$.
3. Actualizar $\kappa = \kappa + 1$.
4. Repetir los pasos 2 y 3 hasta qué $\delta_\ell^* \geq 0$ o $\kappa \geq 0$.
5. Actualizar $\delta_\ell = \mid \delta_\ell^* \mid$.

C. Anexo: Notación

Los vértices cuyas entradas son variables son notados por una versión en negrilla de la letra que representa la variable. Por ejemplo, $\boldsymbol{x} = (x_1, \ldots, x_n)$ es un vector de dimensión $1 \times n$ cuyas entradas son $x_1, \ldots, x_n$. Las matrices cuyas entradas son variables son notadas por una versión en negrilla y mayúscula de la letra que representa la variable. Por ejemplo, $\mathbf{X}_n = [x_{i,j}]$ es una matriz cuyas entradas son representadas por los $x_{i,j}$. La dimensión de la matriz se indica en el subíndice y se asume como una matriz cuadrada a menos que se indique lo contrario, i.e, $\mathbf{X}_n$ es una matriz cuadrada de dimensión $n \times n$.

La norma de un vector representada por $|| \cdot ||$ corresponde a la norma euclidiana. El transpuesto de un vector es notado por $\boldsymbol{x}^T$ y de forma análoga aplica para matrices (e.g., $\mathbf{X}^T$). Finalmente, se presentan las distribuciones de probabilidad empleadas en el documento.

- Bernoulli:

  Una variable aleatoria X tiene una distribución Bernoulli de parámetro θ, notada por $X \mid \theta \sim \mathsf{Ber}(\theta)$ si su función de probabilidad es

  $$\mathsf{p}(x \mid \theta) = x^{\theta}(1-x)^{1-\theta}, \quad \text{para} \quad x \in \{0, 1\}$$

- Uniforme Continua:

  Una variable aleatoria X tiene una distribución Uniforme Continua de parámetros a y b, notada por $X \mid a, b \sim U(a, b)$; si su función de probabilidad es

  $$\mathsf{p}(x \mid a, b) = \frac{1}{b-a}, \quad \text{para} \quad x \in [a, b]$$

- Normal:

  Una variable aleatoria X tiene una distribución Normal de parámetros μ y σ, notada por $X \mid \mu, \sigma \sim N(\mu, \sigma^2)$; si su función de probabilidad es

  $$\mathsf{p}(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\mathsf{exp}\{-\frac{(x-\mu)^2}{2\sigma^2}\}, \quad \text{para} \quad x \in \mathbf{R}$$

- Normal multivariada:

  Un vector aleatorio $\boldsymbol{X} = (X_1, \ldots, X_n)$ tiene una distribución Normal Multivariada n-dimensional de parámetros $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, notada por $\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; si su función de probabilidad es

$$\mathsf{p}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} \mid \boldsymbol{\Sigma} \mid^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\} \quad \text{para} \quad \boldsymbol{x} \in \mathbf{R}^n$$

Bibliografía

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Bender, E. A. and Canfield, E. R. (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335.

Ciminelli, J. T., Love, T., and Wu, T. T. (2019). Social network spatial model. *Spatial statistics*, 29:129–144.

Durante, D., Dunson, D. B., et al. (2018). Bayesian inference and testing of group differences in brain networks. *Bayesian Analysis*, 13(1):29–58.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the american Statistical association*, 81(395):832–842.

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.

Gonçalves, L., Subtil, A., Oliveira, M. R., and de Zea Bermudez, P. (2014). Roc curve estimation: An overview. *REVSTAT-Statistical journal*, 12(1):1–20.

Handcock, M. S. and Krivitsky, P. N. (2008). Fitting latent cluster models for networks with latentnet. *Journal of Statistical Software*, 24(05).

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the american Statistical association*, 100(469):286–295.

Hoff, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Computational and mathematical organization theory*, 15(4):261.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098.

Kim, B. (2018). *Latent Modeling of Dynamic Social Networks*. The Pennsylvania State University.

Kolaczyk, E. D. and Csárdi, G. (2020). *Statistical analysis of network data with R,2nd edn,*. Springer.

Linkletter, C. D. (2007). *Spatial process models for social network analysis*. PhD thesis, Citeseer.

Padgett, J. F. and Ansell, C. K. (1993). Robust action and the rise of the medici, 1400-1434. *American journal of sociology*, 98(6):1259–1319.

Raftery, A. E., Niu, X., Hoff, P. D., and Yeung, K. Y. (2012). Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics*, 21(4):901–919.

Salter-Townshend, M. and McCormick, T. H. (2017). Latent space models for multiview network data. *The annals of applied statistics*, 11(3):1217.

Schweinberger, M. and Snijders, T. A. (2003). Settings in social networks: A measurement model. *Sociological Methodology*, 33(1):307–341.

Sewell, D. K. (2019). Latent space models for network perception data. *Network Science*, 7(2):160–179.

Sewell, D. K. and Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657.

Sosa, J. and Betancourt, B. (2022). A latent space model for multilayer network data. *Computational Statistics & Data Analysis*, page 107432.

Sosa, J. and Buitrago, L. (2021). A review of latent space models for social networks. *Revista Colombiana de Estadística*, 44(1):171–200.

Sosa, J. and Rodríguez, A. (2021). A latent space model for cognitive social structures data. *Social Networks*, 65:85–97.

Wang, L., Zhang, Z., Dunson, D., et al. (2019). Common and individual structure of brain networks. *The Annals of Applied Statistics*, 13(1):85–112.

Ward, M. D. and Hoff, P. D. (2007). Persistent patterns of international commerce. *Journal of Peace Research*, 44(2):157–175.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473.