

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO



REGRESIÓN AVANZADA

Primavera 2023

Predicción de Crímenes en la Ciudad de Boston

Diego Arturo Velázquez Trejo
Marcelino Sánchez Rodríguez
Joaquín Álvarez Galnares

Índice

1. Introducción	2
1.1. Contexto y Descripción del Problema	2
1.2. Variables	2
2. Análisis Exploratorio de Datos	5
3. Análisis de Componentes Principales	8
3.1. Motivación	8
3.2. Descripción del PCA	9
3.3. Formulación del PCA	9
3.4. Interpretación de los resultados	10
3.5. PCA para analizar información de Boston	10
4. Modelos propuestos	13
5. Estrategia de predicción y <i>kriging</i> bayesiano	14
6. Modelo 1	15
6.1. Monitoreo de convergencia de las cadenas (MCMC)	15
6.2. Resumen, inferencia bayesiana e interpretación	17
6.3. Evaluación de las Predicciones	18
7. Modelo 2	20
7.1. Monitoreo de convergencia de las cadenas (MCMC)	21
7.2. Resumen, inferencia bayesiana e interpretación	23
7.3. Evaluación de las Predicciones	24
8. Modelo 3	25
8.1. Monitoreo de convergencia de las cadenas (MCMC)	26
8.2. Resumen, inferencia bayesiana e interpretación	28
8.3. Evaluación de las Predicciones	28
9. Discusión	29
9.1. ¿Cómo interpretar las predicciones de las tasas de crimen si están en escala logarítmica?	29
10. Conclusiones	31
11. Bibliografía	31

1. Introducción

1.1. Contexto y Descripción del Problema

Utilizamos distintos modelos de regresión espaciales para analizar la situación de crímenes en la ciudad de Boston durante el 2022 usando como variables explicativas características demográficas de los distritos de Boston y evaluar la capacidad predictiva de dichos modelos. El gobierno de la ciudad de Boston mantiene un sitio en donde publica datos relativos a diversos aspectos de la ciudad. En particular, comparte los datos recopilados por el Departamento de Policía de Boston sobre la ocurrencia de crímenes. Entre otras cosas, los datos incluyen el tipo de crimen, el momento en que ocurrió y las coordenadas geográficas en las cuales ocurrió. Por otro lado, el gobierno de la ciudad de Boston también comparte información muy detallada recolectada en censos, sobre diversas características de los distintos distritos de la ciudad.

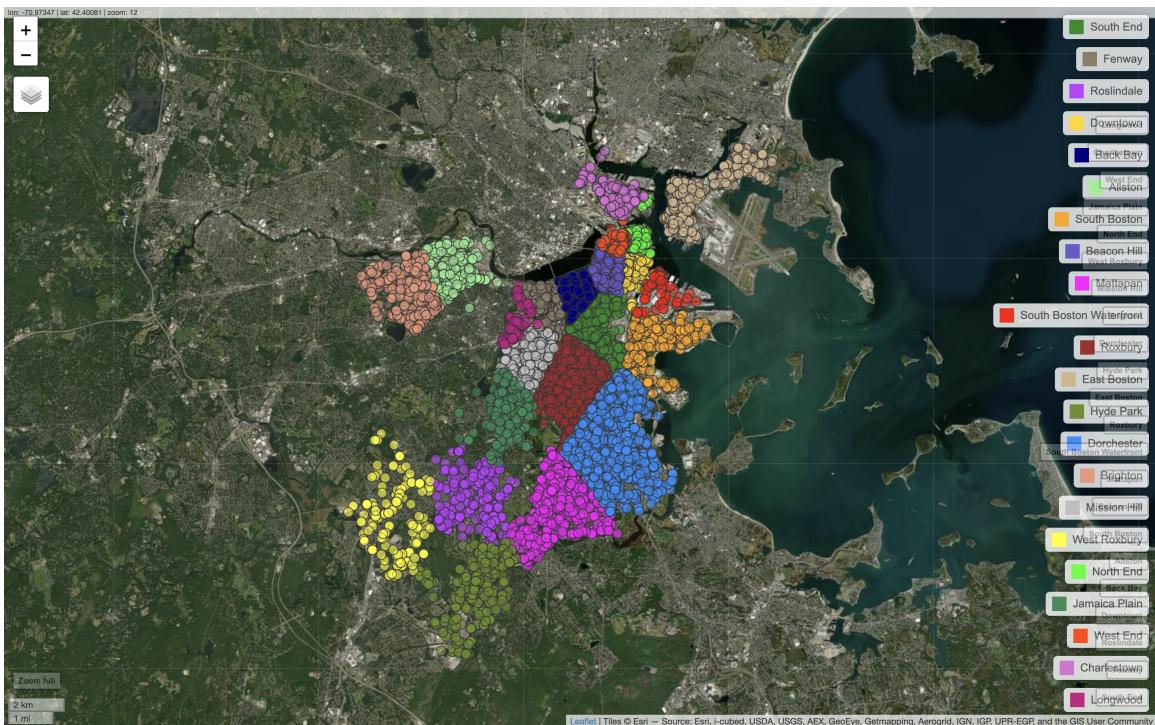


Figura 1: Una imagen aérea de Boston y las afueras de la ciudad. De colores marcamos puntos donde ocurrieron crímenes, estratificando por distrito.

1.2. Variables

La base de datos de crímenes en Boston que ocurrieron en 2022 cuenta con diversas categorías de tipos de crimen. De hecho, el Departamento de Policía de Boston no solo reporta crímenes sino también cualquier tipo de evento para el cual lo ciudadanos recurran al Departamento de Policía. Así por ejemplo, la base de datos también incluye suicidios, uso sospechoso de drogas, etcétera, que en estricto sentido no necesariamente son crímenes *per se*. Lo que nosotros hicimos fue restringirnos a ciertos tipos de categorías que consideramos interpretables y relevantes para comprender la situación de

crimen en Boston. En particular, decidimos considerar: robos, asaltos simples, asaltos agravados, robo de coches, robo de bicicletas, robo de auto-partes, hurto y vandalismo. Nosotros construimos la variable respuesta siguiendo el procedimiento que explicamos a continuación. Contamos con dos bases de datos: una de ocurrencia de crímenes en 2022 (restringida a los tipos de crimen antes mencionados) y otra de características demográficas para 22 distritos de Boston. La base de datos de los crímenes contiene el momento en que se reportó el crimen a la policía y la ubicación donde ocurrió en coordenadas geográficas (longitud, latitud). Entonces a cada distrito en la base de datos geográficos le asignamos una coordenada que consideramos central dentro de ese distrito. Luego a cada crimen le asignamos como etiqueta el distrito más cercano al cual ocurrió dicho crimen respecto a las coordenadas que asociamos a cada distrito. Finalmente, sumamos el número total de crímenes que ocurrieron en cada distrito en 2022 y a esos totales los dividimos entre la población del respectivo distrito. En notación, si $i \in \{1, \dots, 22\}$ es el i -ésimo distrito en Boston y ocurrieron $Z_i \in \mathbb{N}$ crímenes en 2022 en dicho distrito y de acuerdo al censo de 2019 habían $M_i \in \mathbb{N}$ habitantes en ese distrito entonces tomamos $Y_i := \frac{Z_i}{M_i} \in (0, \infty)$. Y_i nos dice el número de crímenes por habitante en el distrito i en el año 2022. De hecho, por la naturaleza de los datos, como es de esperarse que cada distrito tenga más habitantes que crímenes ocurridos en un año, lo que ocurre es que $Y_i \in (0, 1) \quad \forall i \in \{1, \dots, 22\}$. Para ajustar un modelo espacial con un proceso Gaussiano consideramos trabajar con una transformación invertible y que nos lleve a una variable respuesta que en principio pueda tomar valores en todo \mathbb{R} . Así pues, lo más natural es trabajar con $Y(s_i) := \log(Y_i)$, con $s_i \in \mathbb{R}^2$ las coordenadas geográficas asociadas al distrito i . Así pues, nuestra variable respuesta en los modelos que vamos a construir será la tasa de crimen en escala logarítmica.

En cuanto a las variables explicativas, los censos ofrecen información muy amplia y detallada de cada distrito.

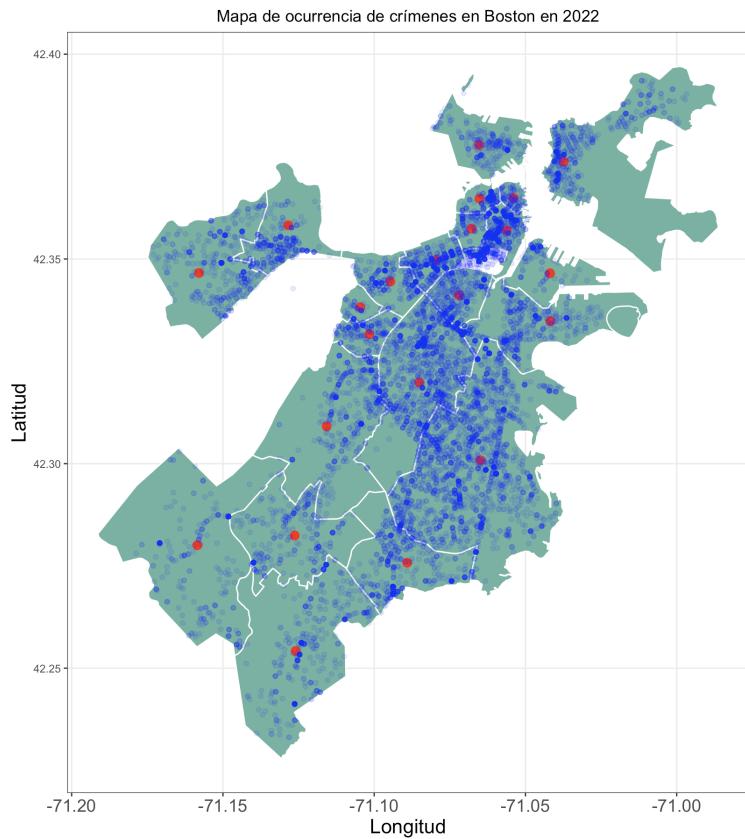


Figura 2: El mapa de Boston está dividido por distritos. Los puntos rojos son coordenadas con las cuales identificamos a cada distrito. Los puntos azules son localizaciones donde se reportó alguno de los tipos de crimen que estamos considerando.

Nótese que hay ocurrencia de crímenes al rededor de las coordenadas $(-71.06, 42.34)$, donde en principio no hay mapa. Esa región de blanco corresponde al barrio chino de Boston. Desafortunadamente, la base de datos de características demográficas no contiene información específica asociada a este distrito. Así que lo que hicimos fue distribuir los crímenes que ocurrieron en esa zona y asociarlos al centroide más cercano según la localización donde hayan ocurrido.

Por comodidad para el lector y para tener acceso a una representación visual que permita identificar cada distrito, a continuación presentamos un mapa que pone con etiqueta el nombre de cada distrito en su respectiva delimitación geográfica

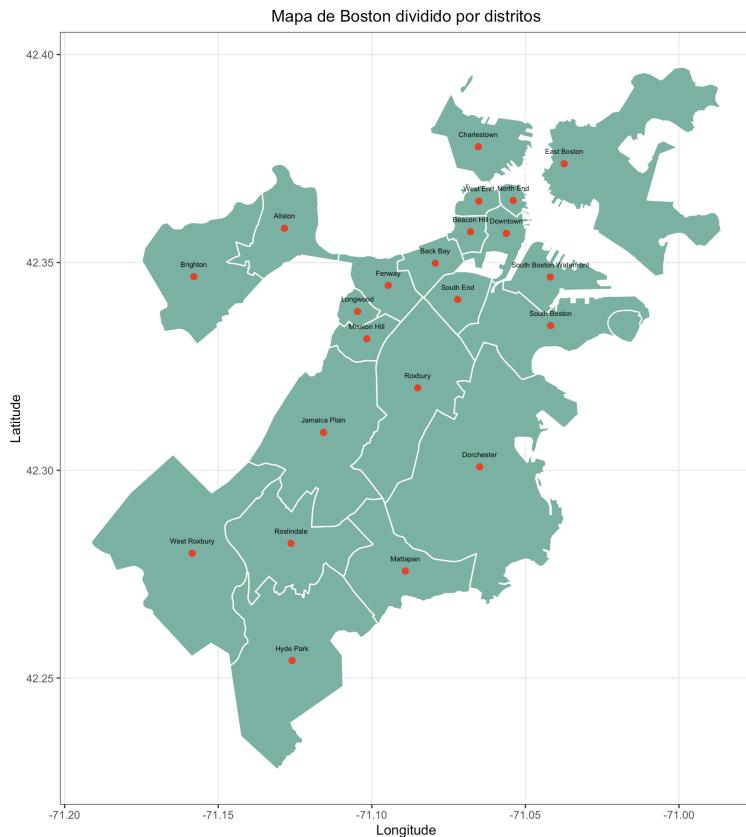


Figura 3: El mapa de Boston está dividido por distritos. Los puntos rojos son coordenadas con las cuales identificamos a cada distrito.

2. Análisis Exploratorio de Datos

La *enorme* cantidad de variables explicativas (más de 100) nos llevó a la necesidad de trabajar con menos por dos razones: interpretabilidad y cómputo. No todas las variables tienen una definición clara. Además no es viable correr modelos de regresión con una cantidad gigante de variables. Así que, a nuestro criterio consideramos algunas que nos parecieron interesantes para tratar de explicar los crímenes de Boston y que además fueran interpretables.

Variable Explicativa	Descripción
$porcentajeUndergrads_i$	Porcentaje de personas que tienen un grado a nivel licenciatura en el distrito i
$PorcentajeMaestria_i$	Porcentaje de personas que tienen un título de maestría en el distrito i
$PorcentajeHispanos_i$	Porcentaje de personas de origen hispano/latino en el distrito i
$PorcentajeAfroamericanos_i$	Porcentaje de personas de origen étnico afroamericano en el distrito i
$PorcentajeMayores_i$	Porcentaje de personas en el distrito i que tienen más de 60 años de edad
$PorcentajeCasados_i$	Porcentaje de viviendas en el distrito i en las que residen personas casadas
$incomePerCapita_i$	Ingreso promedio anual de personas en el distrito i
$metroTren_i$	Porcentaje de personas en el distrito i cuyo principal medio de transporte es el metro o tren
$houseHoldIncome_i$	Porcentaje de vecindarios con ingreso superior a 150 mil dólares anuales en el distrito i
$longTravel_i$	Porcentaje de personas que hacen en promedio más de 60 minutos para llegar a su trabajo desde su hogar en el distrito i

Cuadro 1: La definición de cada variable explicativa

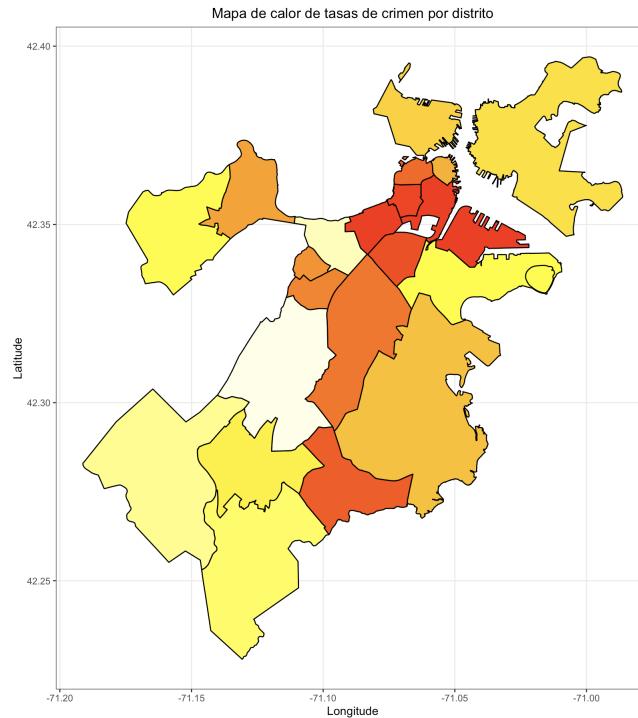


Figura 4: El color rojo está asociado a distritos con mayores tasas de crimen per cápita. El color blanco crema está asociado al distrito con menor tasa de crimen per cápita en 2022.

Resultó ser que *Jamaica Plain* es el distrito con menos crímenes per cápita, seguido de *Longwood*. Puede corroborarse con ayuda de las figuras 4 y 3 para identificar el nombre del distrito y la intensidad

asociada en el mapa de calor.

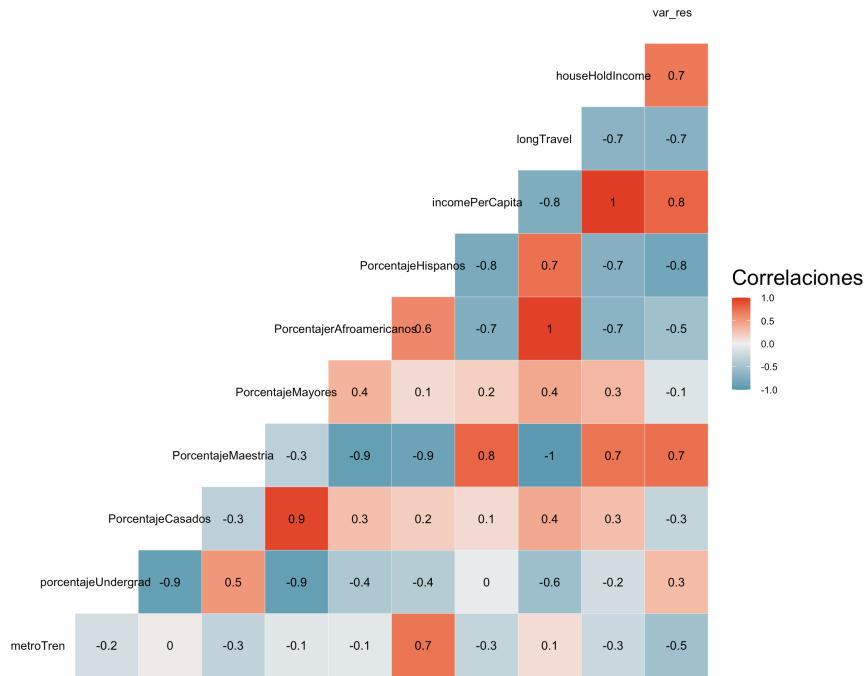


Figura 5: Matriz de correlaciones de variables explicativas y variable respuesta.

Vemos que algunas variables tienen correlación muy cercana a uno. Por ejemplo, *HouseHoldIncome* e *incomePerCapita*. Ambas se relacionan con el ingreso de las personas, por lo que es razonable que tengan correlación alta. Una correlación alta que es curiosa es *porcentajeAfroamericanos* con *longTravel*.

Si bien, no es la mejor práctica seleccionar variables en un análisis exploratorio, si dos variables tienen correlación cercana a 1, es razonable suponer que aportan casi la misma información al modelo de regresión lineal espacial y para para reducir el cómputo, nos quedaremos con *incomePerCapita* y *porcentajeAfroamericanos*.

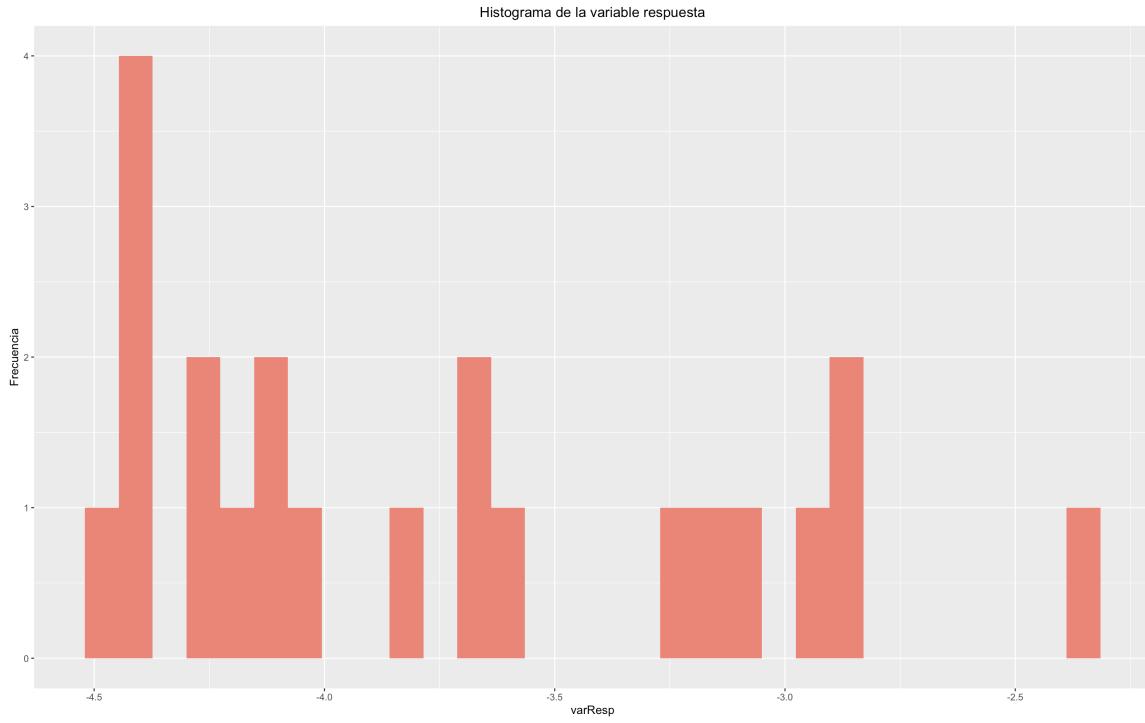


Figura 6: Histograma de la variable respuesta.

Notamos que hay un valor extremo que corresponde al máximo valor que toma la variable respuesta en la base de datos: -2.340. Podemos considerarlo extremo porque la media de la variable respuesta es de -3.743 y la desviación estándar es de 0.64, por lo que dicho valor dista de la media en más de dos veces la desviación estándar. De hecho, coherente con el mapa de calor, el distrito con tasas de crimen más grandes es *South Boston Waterfront*. Más adelante veremos que cuando particionamos el conjunto de datos en dos (uno para el aprendizaje y otro para validación de los modelos), este distrito fue uno de los que resultó seleccionados para la validación (predicción sobre una localización nueva). Es importante advertir esto desde ahorita, pues conviene tener presente que nuestra estrategia de predicción y de evaluación y análisis de los modelos nos permitirá tener una noción de que tan robustos son nuestros modelos en sus predicciones sobre observaciones atípicas. Por otro lado, el distrito con menos crímenes (pintado de blanco crema en el mapa de calor) es *Longwood*, que de hecho se conoce como *Longwood Medical and Academic Area*, lo cual nos sugiere que es un distrito con características muy particulares y diferentes al resto de los distritos en Boston: consiste prácticamente de hospitales e instituciones académicas, de hecho ahí se encuentra la Escuela de Medicina de Harvard.

3. Análisis de Componentes Principales

3.1. Motivación

La enorme cantidad de variables explicativas proporcionadas por los datos demográficos recabados en el censo así como la necesidad de poder implementar modelos que corrieran en tiempos razonables en OpenBUGS nos hizo recurrir a técnicas de reducción de dimensionalidad, para trabajar con menos

variables explicativas, tratando de conservar la información que aportan las variables que teníamos originalmente. En particular, decidimos aplicar Análisis de Componentes Principales a los datos demográficos de los distritos de Boston. Con esto, generamos los índices para educación, raza, grupos de edad y movilidad.

El análisis de componentes principales (PCA) es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos, mientras se conserva la mayor cantidad posible de su varianza. A continuación se presenta un resumen del proceso de PCA:

3.2. Descripción del PCA

1. **Normalización de los datos:** Los datos deben ser centrados en cero y escalados para tener una media de cero y una desviación estándar de uno. Es importante considerar que PCA tiende ser susceptible a cambios de escala. Es por ello que para nuestro caso de estudio, se trabajaron únicamente con bases de datos que registraban porcentajes.
2. **Cálculo de la matriz de covarianza:** Se calcula la matriz de covarianza para los datos normalizados. La matriz de covarianza se encarga de capturar las relaciones lineales entre las variables.
3. **Obtención de los eigenvalores y eigenvectores:** Se calculan los autovalores y autovectores de la matriz de covarianza. Los eigenvectores representan las direcciones principales en los datos, y los eigenvalores indican la cantidad de varianza explicada por cada autovector.
4. **Selección de componentes principales:** Se ordenan los autovalores de mayor a menor y se eligen los primeros k autovectores correspondientes a los k mayores autovalores. Estos autovectores forman las componentes principales. Para realizar el índice, se consideró una combinación lineal entre las dos primeras componentes.
5. **Proyección de los datos:** Los datos originales se proyectan en el nuevo espacio definido por las componentes principales. La proyección se realiza multiplicando los datos normalizados por la matriz de autovectores seleccionados.

3.3. Formulación del PCA

A continuación se presentarán las ecuaciones matemáticas que se ejecutan para realizar el PCA. Para efectos de este trabajo, se utilizó R, la función de **prcomp**¹.

- **Datos centrados en cero:** $\mathbf{Y} = \mathbf{X} - \bar{\mathbf{X}}$, donde \mathbf{X} es la matriz de datos y $\bar{\mathbf{X}}$ es el vector de medias de las columnas de \mathbf{X} .
- **Matriz de covarianza:** $\mathbf{C} = \frac{1}{n-1} \mathbf{Y}^T \mathbf{Y}$, donde n es el número de muestras en \mathbf{X} .
- **Autovalores y autovectores:** $\mathbf{Cv}_i = \lambda_i \mathbf{v}_i$, donde λ_i es el i -ésimo autovalor y \mathbf{v}_i es el i -ésimo autovector.

¹Esta función es nativa al lenguaje de R y no requiere de instalar ningún paquete en especial.

- **Proyección de los datos:** $\mathbf{T} = \mathbf{X}\mathbf{V}_k$, donde \mathbf{T} es la matriz de datos proyectados, \mathbf{X} es la matriz de datos normalizados y \mathbf{V}_k es una matriz formada por los primeros k autovectores seleccionados (nosotros tomamos las primeras $k = 2$ componentes principales para la construcción de cada índice).

3.4. Interpretación de los resultados

La interpretación para PCA resulta ser complicada. Se necesitan analizar de manera independiente cada componente y ver el peso que le asigna a cada variable. Se pueden utilizar los biplots para ver cómo se comportan los componentes en relación a las variables.

- **Varianza explicada por cada componente principal:** Los eigenvalores proporcionan una medida de la varianza explicada por cada componente principal. Cuanto mayor sea el autovalor, mayor será la varianza explicada por esa componente.
- **Contribución de las variables a las componentes principales:** Los coeficientes de los eigenvectores indican cómo cada variable contribuye a la formación de las componentes principales. Los coeficientes más grandes indican una mayor contribución de la variable correspondiente.
- **Visualización de los datos en un espacio de menor dimensión:** Al proyectar los datos originales en el espacio de las componentes principales seleccionadas, es posible visualizarlos en un espacio de menor dimensión. Esto puede facilitar la comprensión de los patrones y estructuras presentes en los datos.

3.5. PCA para analizar información de Boston

Se aplicó PCA a la base de datos de edades (0-9 años, 10-17 años, 18-19 años, 20-34 años, 35-59 años, 60- +60 años), base de datos de educación (porcentaje que tienen educación menor a preparatoria, educación preparatoria, carrera técnica, licenciatura y maestría), base de datos de movilidad (movilidad de boston a otro estado, a diferente país, movilidad de cualquier estado a alguna parte de boston, movilidad de cualquier otro país a algún distrito de boston), base de datos de raza (porcentaje de blancos, raza afroamericana, hispanos, asiáticos y otros).

Para construir los índices, se consideró una combinación lineal con las primeras dos componentes para cada base de datos. Para la base de edades, se explicó el 87 % de la variabilidad, para la base de educación se explicó el 92.8 %, para la de movilidad se explicó el 86.07 % y para la información de raza se explicó el 78.6 % de la variabilidad.

Se observó que los componentes principales que mejor interpretabilidad tuvieron fueron los asociados a raza y a educación. En la figura 7, en el biplot de educación podemos ver que la primera componente disminuye a medida que aumenta el porcentaje de licenciados o de egresados de maestría; por otro lado, aumenta a medida que hay más ciudadanos que no terminan la licenciatura. Después de haber aplicado el índice, se observó que los distritos con mayor índice estaban asociados a aquellos que tenían mayores porcentajes de personas que no concluían la preparatoria o que no concluían la licenciatura; en contraposición con los distritos que tenían un menor índice (mayor porcentaje de licenciados y maestros).

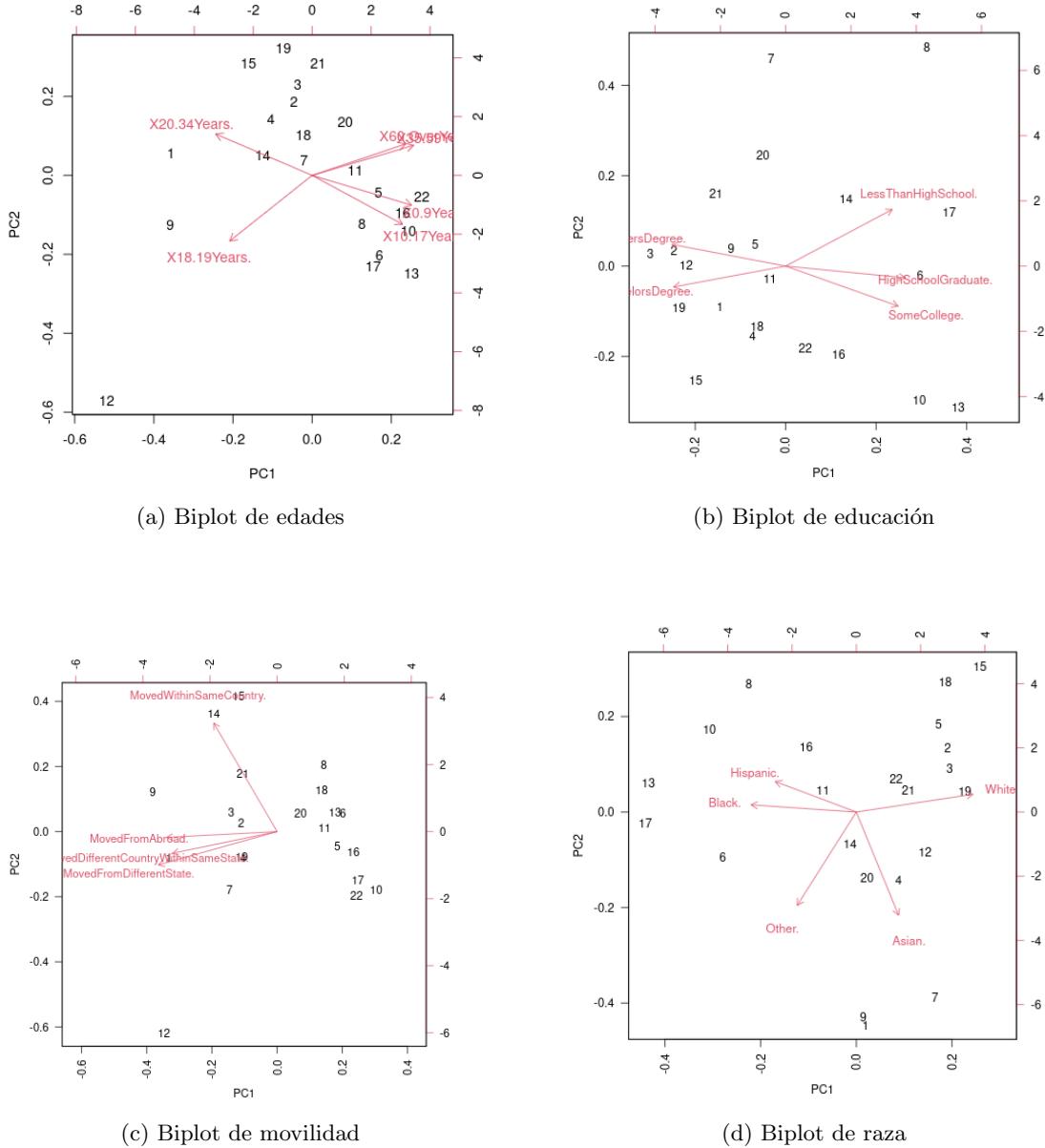


Figura 7: Biplots resultantes de aplicar PCA a cada conjunto de variables. El eje horizontal en cada gráfica corresponde a la primera componente principal y el eje vertical a la segunda componente.

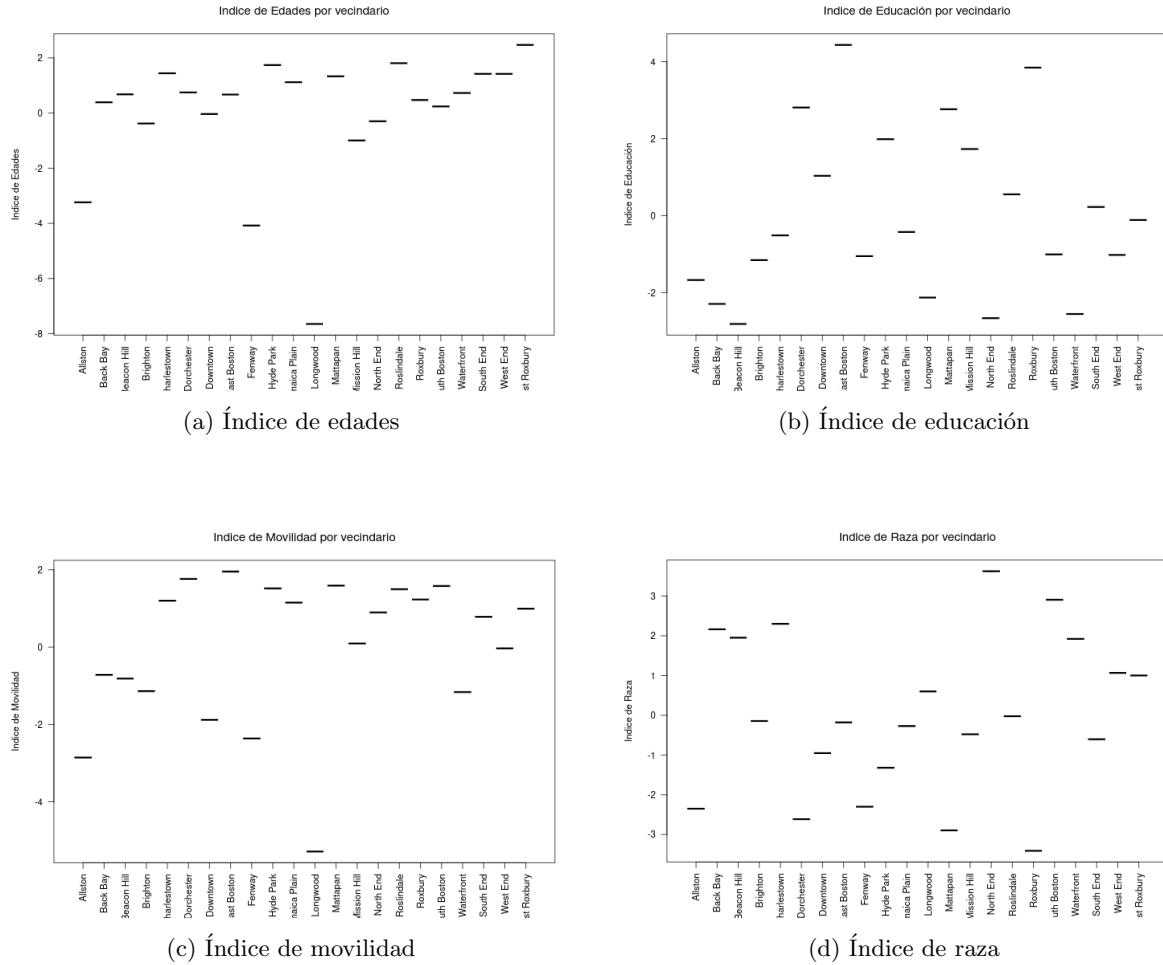


Figura 8: Resultados de los índices construidos. Cada gráfica muestra el valor que toma cada distrito para cada índice construido.

Por lo que se esperaría que en los modelos de regresión, si el coeficiente asociado a este índice es grande, esto significa que aumentará el porcentaje de crímenes a medida que hay más personas sin concluir sus estudios hasta licenciatura. Más adelante veremos en el Modelo 1 que esto es lo que efectivamente ocurre.

Para el índice de raza, en la figura 7, en el biplot de raza, se observa que la primera componente es menor a medida que hay mayor porcentaje de hispanos y de afroamericanos, y a medida que hay más blancos, el índice es mayor. No podemos hacer inferencias a priori sobre como esperaríamos que fuera el coeficiente de regresión asociado a este índice. En la gráfica asociada al índice de raza, los distritos con un mayor índice tienen mayor porcentaje de personas blancas, mientras que a menor índice, mayor porcentaje de hispanos y de raza afroamericana.

4. Modelos propuestos

Vamos a considerar un modelo espacial para datos referenciados puntualmente usando un proceso Gaussiano para incorporar una estructura de dependencia espacial en los datos. Si $s_i \in \mathbb{R}^2$ denota las coordenadas del i-ésimo distrito de Boston, el modelo de regresión con el que trabajaremos es

$$Y(s) = \mu(s) + \omega(s) + \epsilon(s),$$

$$\text{con } Y(s) := \begin{pmatrix} Y(s_1) \\ Y(s_2) \\ \vdots \\ Y(s_n) \end{pmatrix}, \quad \mu(s) := \begin{pmatrix} \mu(s_1) \\ \mu(s_2) \\ \vdots \\ \mu(s_n) \end{pmatrix},$$

$$\mu(s_i) = x(s_i)^T \underline{\beta}, \text{ con } x(s_i) \text{ vector de variables explicativas asociadas al i-ésimo distrito y } \underline{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$w(s) := \begin{pmatrix} w(s_1) \\ w(s_2) \\ \vdots \\ w(s_n) \end{pmatrix} \sim \mathcal{N}_n(\underline{0}, \Sigma(s, t)) \text{ denota un proceso Gaussiano que corresponde a la dependencia espacial en los datos, utilizando una función de covarianza exponencial, esto es, si } s_i, s_j \text{ son las coordenadas de cualesquiera dos distritos de Boston, entonces } Cov(w(s_i), w(s_j)) = \Sigma(s_i, s_j) = \sigma^2 \exp\{-\phi d_{i,j}\},$$

donde $d_{i,j} := \|s_i - s_j\|_2$.

$$\epsilon(s) := \begin{pmatrix} \epsilon(s_1) \\ \epsilon(s_2) \\ \vdots \\ \epsilon(s_n) \end{pmatrix} \sim \mathcal{N}_n(\underline{0}, \varphi^2 I_{n \times n}) \text{ son errores independientes de } w(s).$$

Como consecuencia de esta especificación se sigue que

$$Y(s) \sim \mathcal{N}_n(\mu(s), \Sigma(s, t) + \varphi^2 I_{n \times n})$$

Nótese que este es un modelo de regresión a la media, pues las variables explicativas afectan a la

media de la variable respuesta. Los modelos que vamos a proponer usan diferentes variables explicativas pero todos tienen esa estructura. Más aún, para la distribución inicial de las betas se utilizaron distribuciones normales no informativas con media en cero. Y para los parámetros como ϕ , φ^2 , σ^2 se usaron distribuciones Gamma no informativas. Es importante recordar que la función por defecto de BUGS para la normal usa la parametrización de precisión y no de varianza como aquí la presentamos. Para la implementación de cada modelo se corrieron dos cadenas de longitud 10,000 con adelgazamiento de 20 unidades y un periodo de calentamiento de 1000 iteraciones en OpenBUGS.

5. Estrategia de predicción y *kriging* bayesiano

Tenemos disponibles 22 distritos de Boston. Entonces para comparar los modelos en sus predicciones sobre “nuevas localizaciones”, lo que hicimos fue recurrir al clásico paradigma de aprendizaje de máquina de particionar el conjunto de datos en dos. Para esto fijamos una semilla (manteniendo reproducibilidad) y aleatoriamente seleccionamos $m = 2$ distritos (que corresponde a $\approx 10\%$ del total de datos). Esos dos distritos se reservaron exclusivamente para hacer predicciones con los modelos, para así poder evaluar las predicciones sobre datos que no se utilizaron en el proceso de aprendizaje al momento de hacer inferencia bayesiana para obtener la distribución posterior de los parámetros.

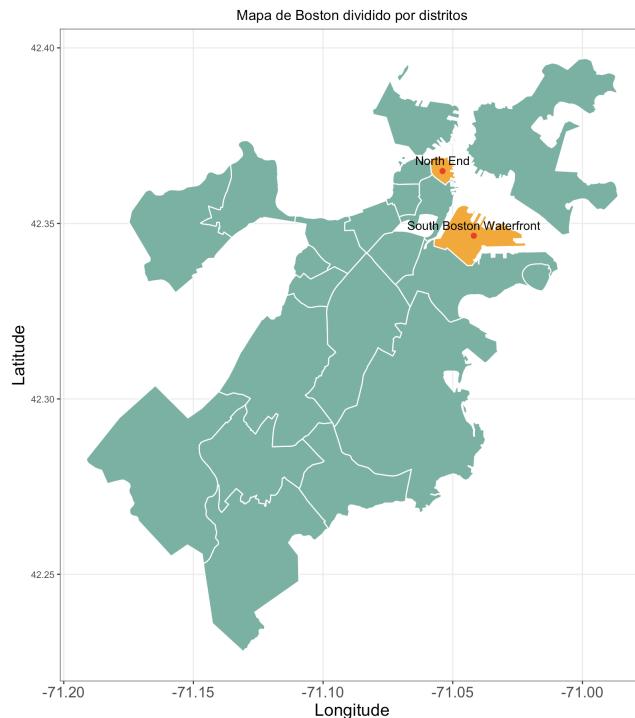


Figura 9: Mapa de los distritos de Boston. De naranja se marcan los distritos reservados exclusivamente para hacer predicciones. Los distritos de azul son los que se utilizaron en el proceso de aprendizaje así como para calcular la pseudo R^2 de los modelos.

6. Modelo 1

El predictor lineal que utilizaremos tiene la forma

$$\mu(s_i) = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \beta_4 X_{3,i} + \beta_5 X_{4,i} + \beta_6 X_{5,i} + \beta_7 X_{6,i},$$

donde las variables que utilizamos son:

X_1 = Índice de edades,

X_2 = Índice de educación,

X_3 = Índice de movilidad,

X_4 = Índice de origen étnico,

X_5 = Ingreso per cápita,

X_6 = Tasa de pobreza: porcentaje de la población en el distrito que vive en situación de pobreza.

6.1. Monitoreo de convergencia de las cadenas (MCMC)

Este modelo tiene varios parámetros. Para no hacer un proceso repetitivo y tedioso, solo mostraremos la convergencia de las cadenas para algunos de los parámetros. Para el resto se observaron propiedades similares a las que se presentan a continuación.

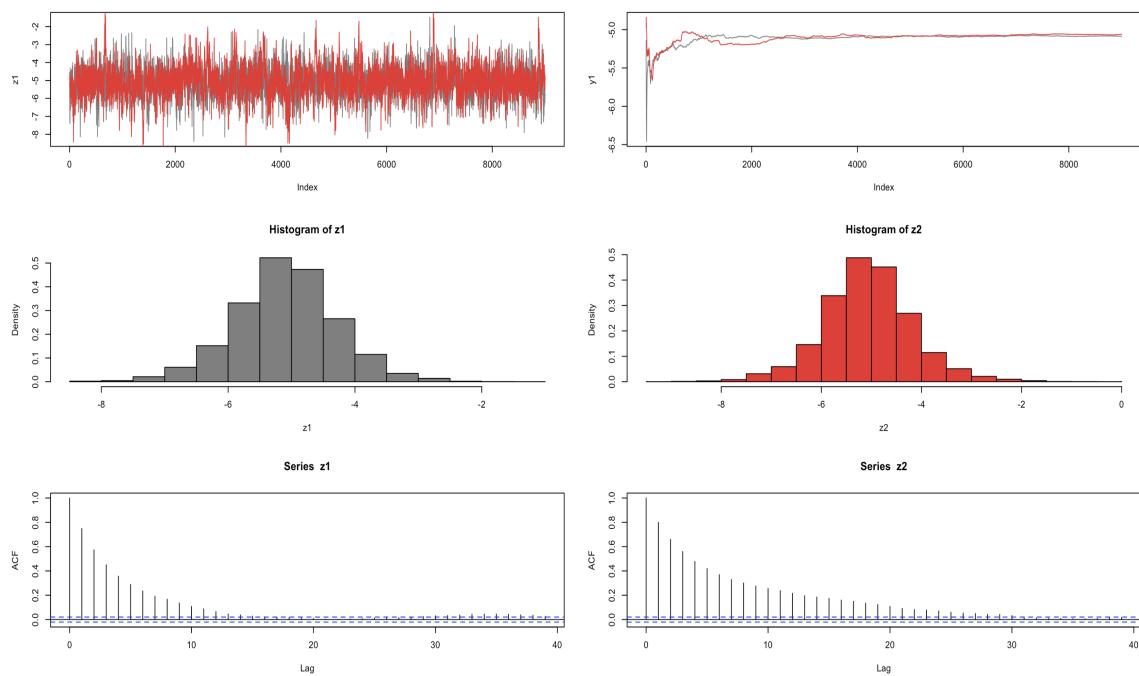


Figura 10: Monitoreo de las cadenas de β_1 .

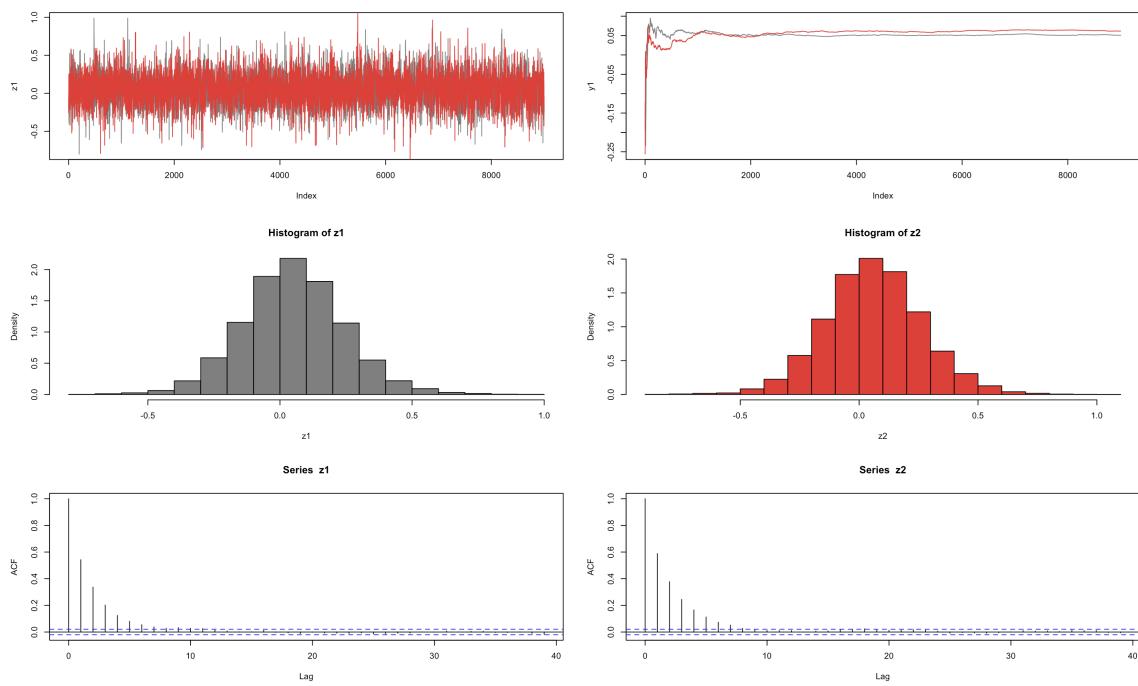


Figura 11: Monitoreo de las cadenas de β_2 .

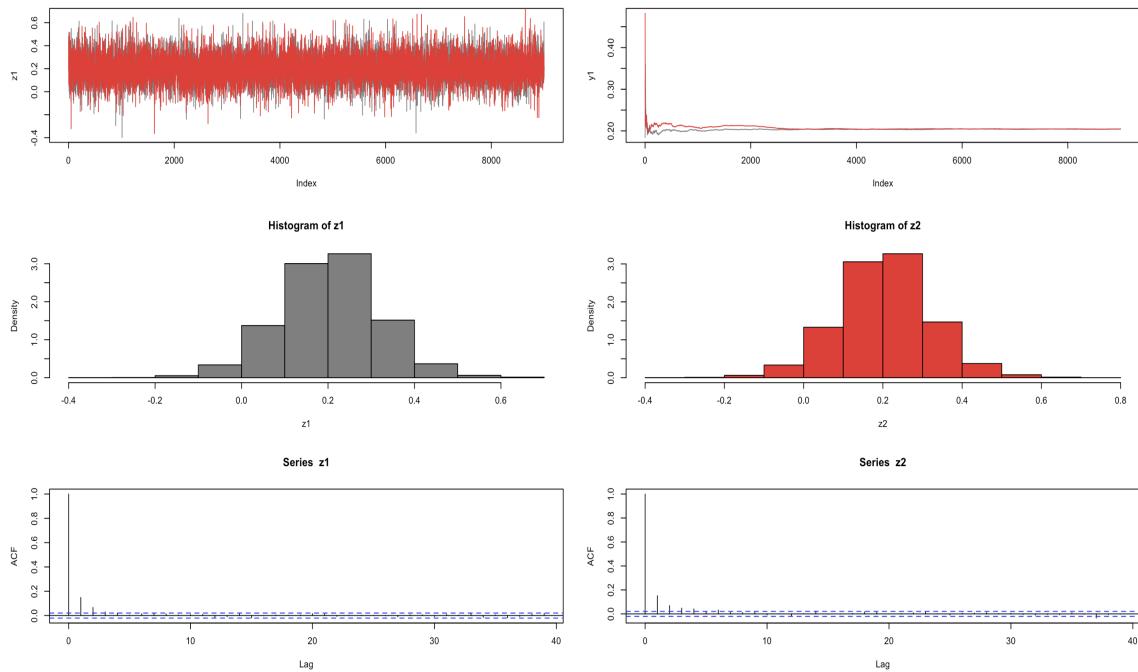


Figura 12: Monitoreo de las cadenas de β_3 .

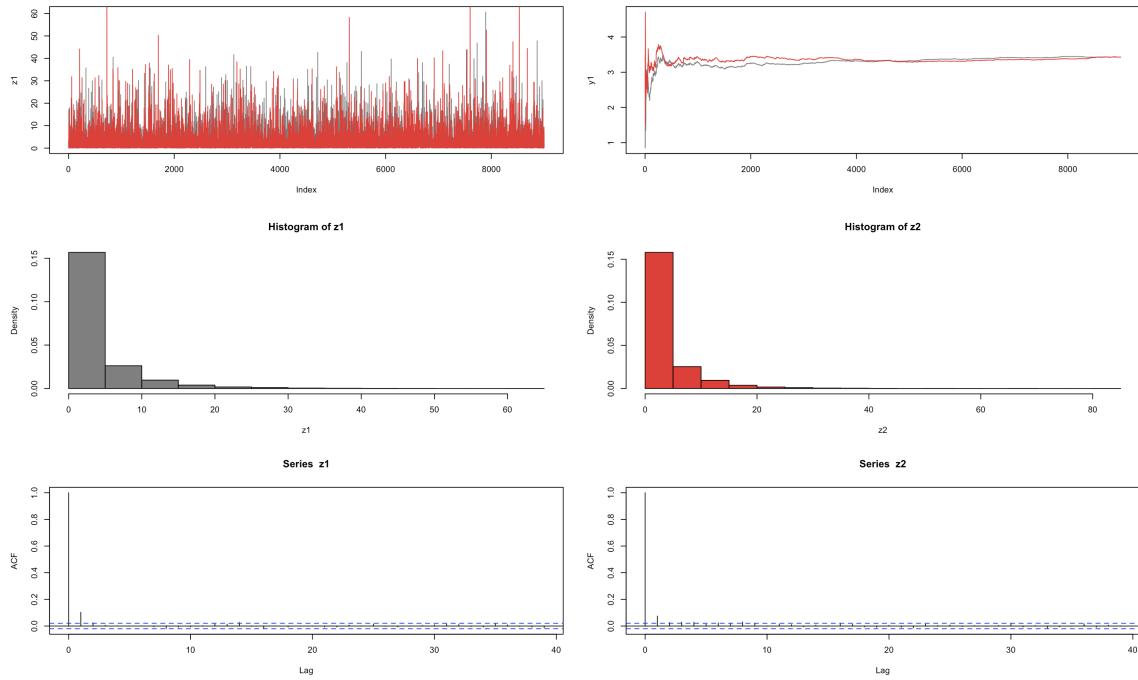


Figura 13: Monitoreo de las cadenas de ϕ .

Como comentario general de lo que se observa en el monitoreo de convergencia, las gráficas nos permiten concluir convergencia en promedios ergódicos así como convergencia a la distribución posterior. Por otro lado, para algunos parámetros (por ejemplo, β_1) tuvieron cierta autocorrelación pero no muy alta. Los demás parámetros tuvieron autocorrelaciones muy bajas en sus cadenas.

6.2. Resumen, inferencia bayesiana e interpretación

Parámetro	Media posterior	Cuantil 2.5 %	Cuantil 97.5 %	valor-p
β_1	-5.0756	-6.8410	-3.3190	0.0000
β_2	0.0558	-0.3262	0.4580	0.3883
β_3	0.2045	-0.0287	0.4367	0.0404
β_4	-0.2850	-0.6856	0.0993	0.0696
β_5	-0.0207	-0.2517	0.2129	0.4232
β_6	1.98e-05	-1.335e-06	3.945e-05	0.0319
β_7	0.4499	-4.3112	5.3460	0.4278
ϕ	3.4287	0.0218	17.6502	0.0000

Cuadro 2: Resumen de las cadenas con valor-p asociado a una prueba de significancia de los coeficientes.

Los valores p se calcularon usando la siguiente idea: si la estimación puntual fue negativa, entonces se calcula la probabilidad (posterior) de que el coeficiente sea positivo y eso es el valor-p asociado a la prueba de significancia de ese coeficiente (y viceversa si la estimación puntual fue positiva). El intercepto no es interpretable, pues no tiene sentido un distrito con ingreso per cápita igual a cero (*¿de qué comería la gente?*). Ahora bien, ¡los coeficientes que no sean significativos no se interpretan! Se

observó que los coeficientes β asociados a los regresores que fueron significativos son β_3 ($p_value = 0.0404$) y β_6 ($p_value = 0.0319$). Observamos que β_3 está asociado al índice de educación y β_6 está asociado al ingreso per cápita. Como se esperaba, β_3 es positivo, esto es que a medida que aumenta en una unidad el índice de educación, aumenta la tasa de crímenes en 0.2 unidades (en promedio y en escala logarítmica). En términos de tasa de crimen per cápita, incrementar una unidad el índice de educación provoca un incremento de $e^{0.02} = 1.02$, por lo tanto se incrementa en dos por ciento la tasa de crimen per cápita en promedio. Como se había discutido previamente, a medida que el índice de educación es mayor, esto se traduce en que hay más personas en ese distrito que no están llegando al nivel de licenciatura. Por lo que el hecho de que haya un alto porcentaje de licenciados y maestros, haría que el índice fuera menor y en consecuencia, (en promedio) tuviera una menor tasa de crímenes per cápita.

Incrementar en un dólar ingreso per cápita provoca un incremento en promedio de 1.98×10^{-5} unidades en el crimen per cápita en escala logarítmica.

6.3. Evaluación de las Predicciones

Este modelo tuvo un DIC de -54.33. Lo tendremos en cuenta para comparar con el resto de los modelos.

Si usamos como predictor puntual a la media de la distribución predictiva posterior, se obtuvo una correlación de 0.89 entre la predicción puntual y la variable respuesta observada usando los 20 distritos del proceso de aprendizaje. Elevando al cuadrado, esto da lugar a una pseudo- $R^2 = 0.793$.

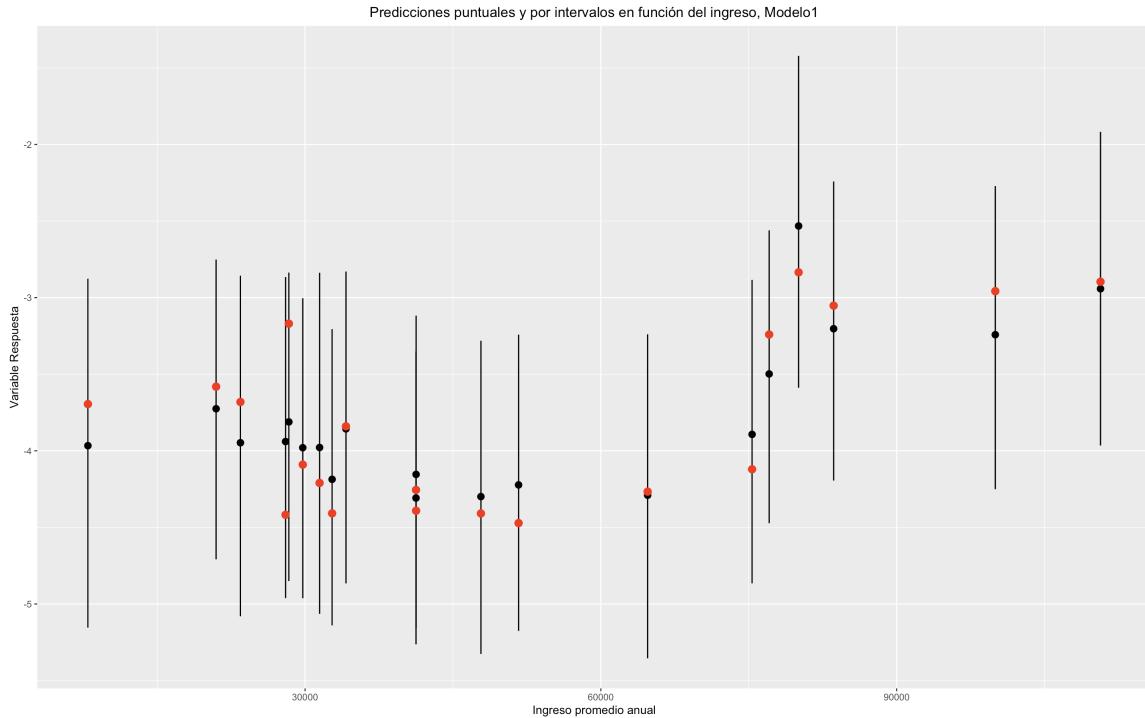


Figura 14: Predicción puntual y por intervalos al 95 % de probabilidad. De color negro tenemos predicciones y los puntos de color rojo corresponden a los valores observados de la variable respuesta (logaritmo de la tasa de crimen per cápita en 2022).

Vemos que las predicciones tienen muy poco sesgo. Los puntos negros quedan muy cerca de los rojos salvo en pocas excepciones.

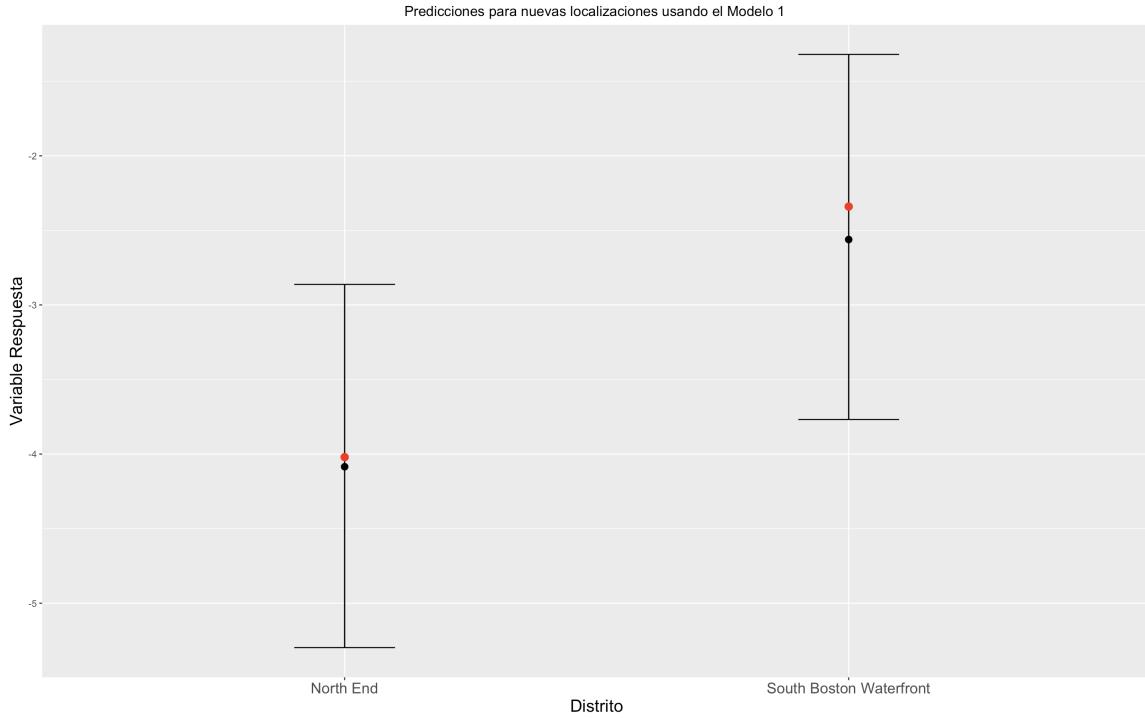


Figura 15: Predicción puntual y por intervalos al 95 % de probabilidad usando la distribución predictiva posterior. De color negro tenemos predicciones y los puntos de color rojo corresponden a los valores observados de la variable respuesta.

Para *North End* la longitud del intervalo de predicción fue de 2.436 y para *South Boston Waterfront* la longitud del intervalo fue de 2.448. Tal como se enfatizó en el análisis exploratorio de datos, la tasa de crimen per cápita de *South Boston Waterfront* es un dato atípico. Vemos que el modelo 1 tiene muy buena capacidad predictiva en esta observación atípica, con poco sesgo en la predicción puntual (la media de la predictiva final) y en su predicción por intervalo, la longitud no es muy grande, además de que la observación extrema está contenida en el intervalo de 95 % de probabilidad.

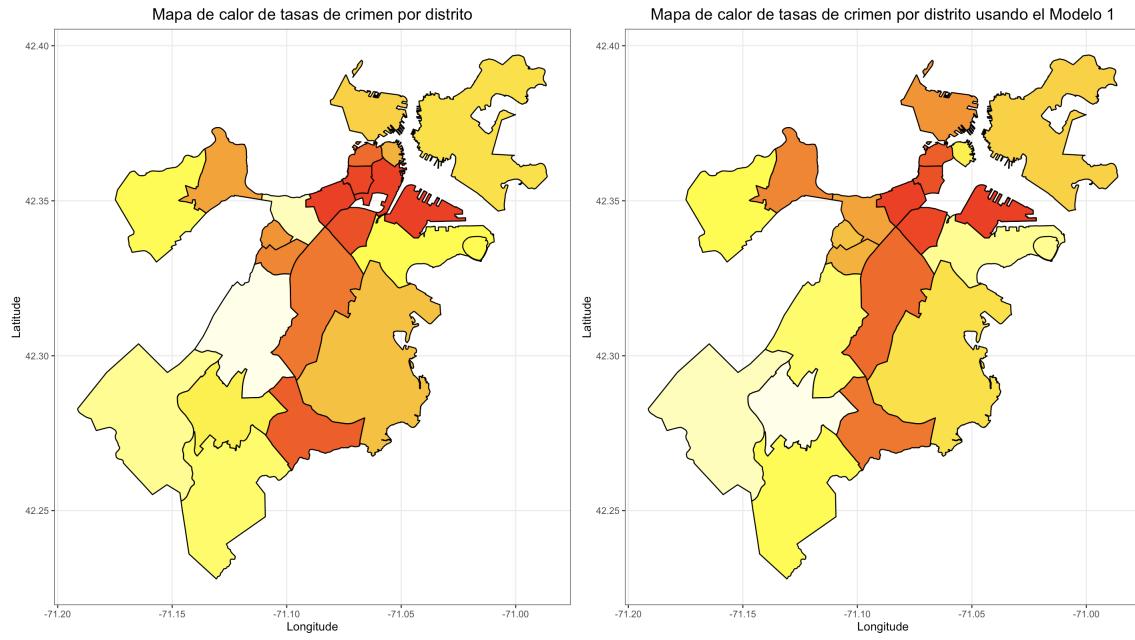


Figura 16: Comparación de mapas de calor. Del lado izquierdo está el mapa de calor de los datos sin ajustar ningún modelo. Del lado derecho presentamos el mapa de calor construido con las predicciones puntuales del modelo.

Podemos notar que el Modelo 1 hace un suavizamiento de los datos.² Por ejemplo, vemos que de los distritos del norte a los del sur va disminuyendo la intensidad de los colores más gradualmente en el mapa de la derecha. Mientras que el mapa de calor de las tasas de crimen de la izquierda tiene cambios de intensidad más drásticos (por ejemplo que distritos de color blanco y de color rojo sean vecinos). En el mapa de la derecha el patrón general es que distritos de color rojo serán vecinos de distritos de colores naranjas o rojos también, y no de distritos de color casi blanco.

7. Modelo 2

El predictor lineal que utilizaremos tiene la forma

$$\mu(s_i) = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \beta_4 X_{3,i} + \beta_5 X_{4,i} + \beta_6 X_{5,i} + \beta_7 X_{6,i} + \beta_8 X_{7,i} + \beta_9 X_{8,i} ,$$

donde las variables que utilizamos son:

X_1 = porcentajeUndergrads,

X_2 = PorcentajeMaestria,

X_3 = PorcentajeHispanos,

X_4 = PorcentajeAfroamericanos,

X_5 = PorcentajeMayores,

²Nota técnica: *Downtown* no aparece pintado en el mapa de calor del Modelo 1. No logramos identificar por qué `ggplot()` no lo colocó.

X_6 =PorcentajeCasados,
 X_7 =incomePerCapita,
 X_8 =metroTren

7.1. Monitoreo de convergencia de las cadenas (MCMC)

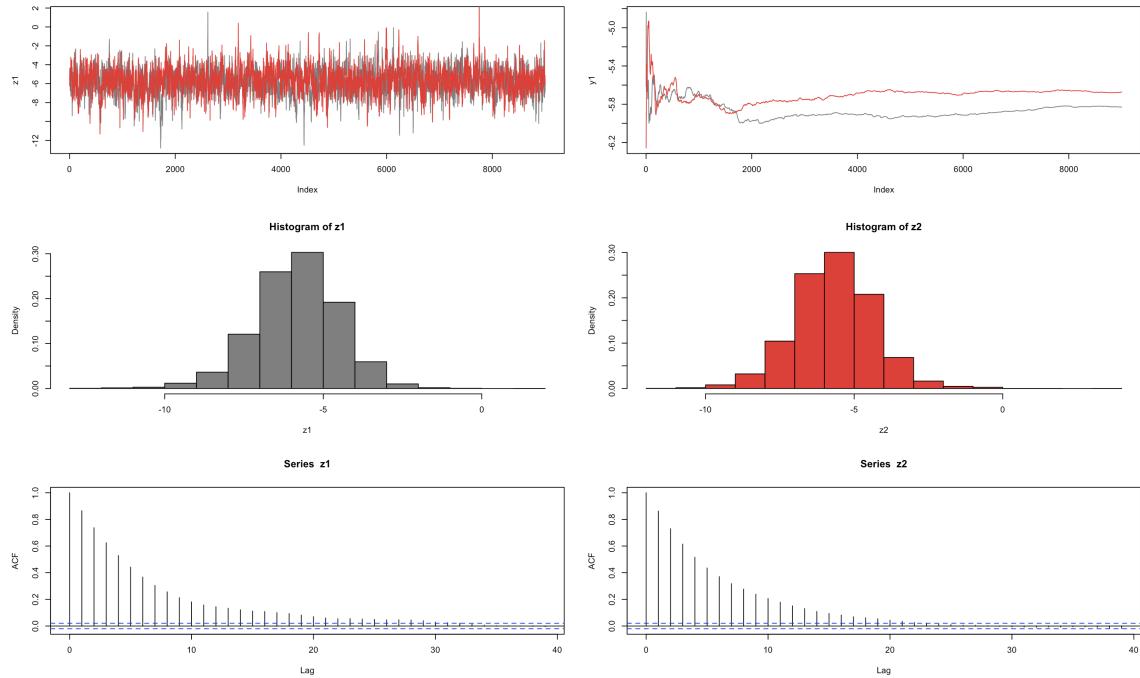


Figura 17: Monitoreo de convergencia de las cadenas de β_1 para el segundo modelo.

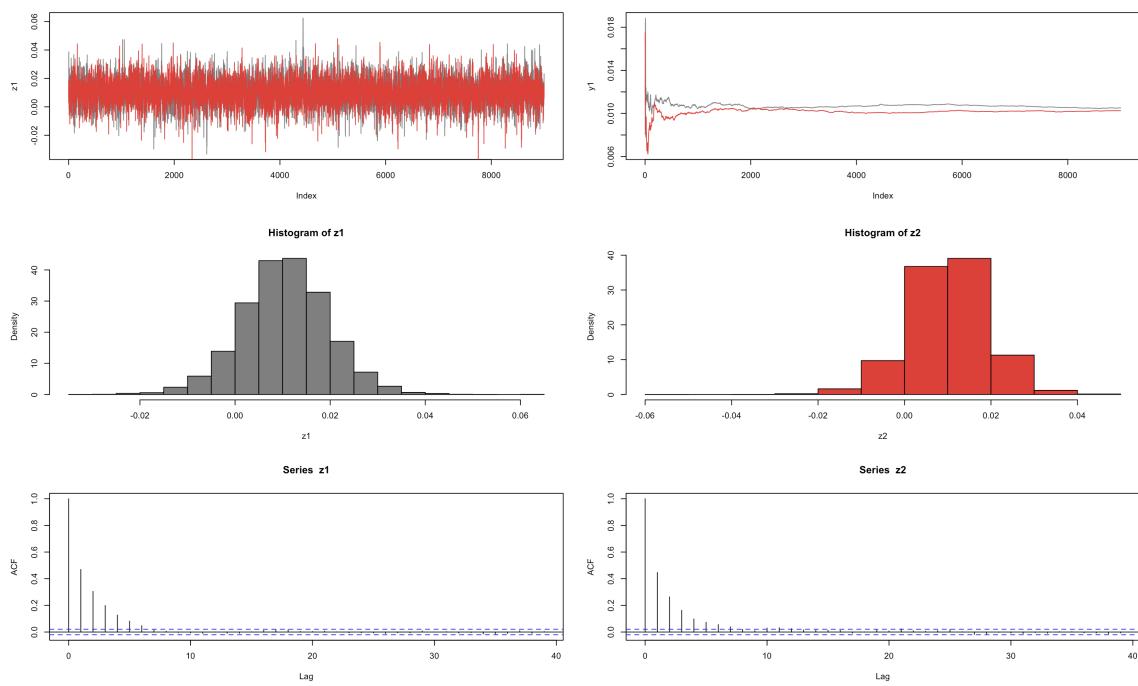


Figura 18: Monitoreo de convergencia de las cadenas de β_2 para el segundo modelo.

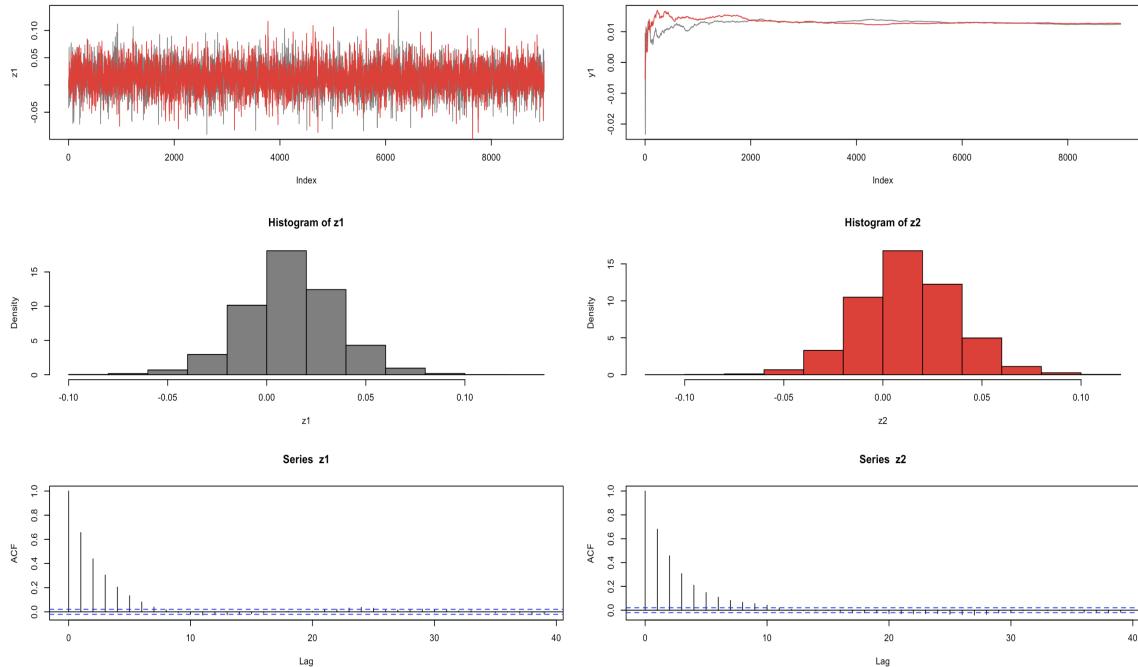


Figura 19: Monitoreo de convergencia de las cadenas de β_3 para el segundo modelo.

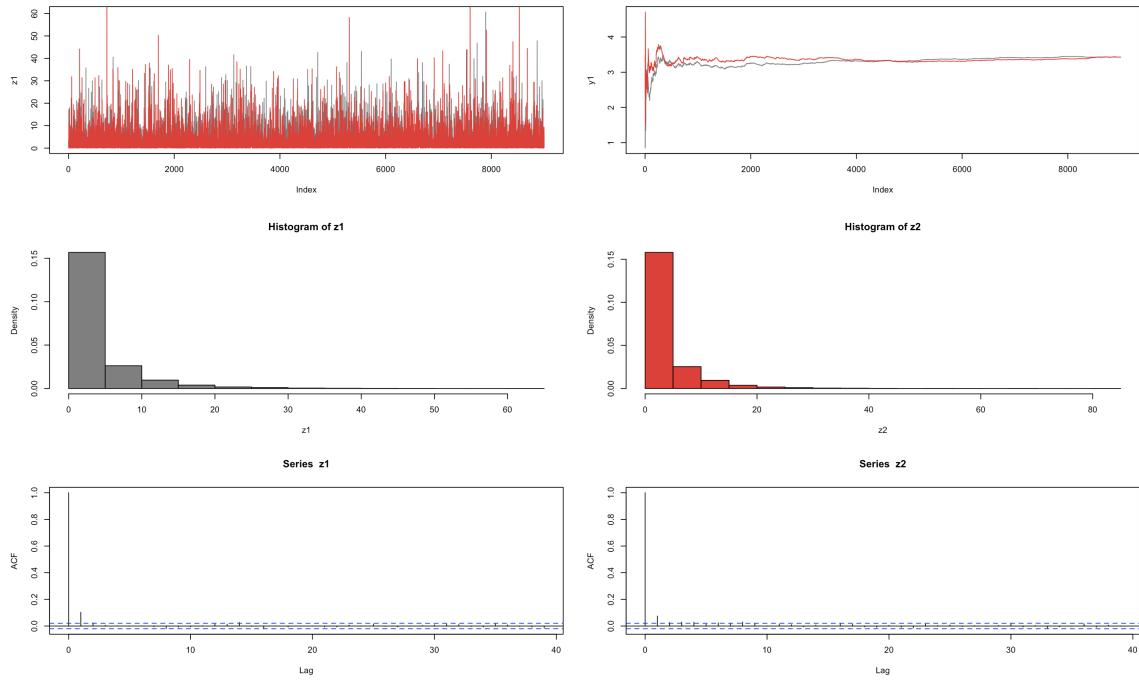


Figura 20: Monitoreo de convergencia de las cadenas de ϕ para el segundo modelo.

7.2. Resumen, inferencia bayesiana e interpretación

Parámetro	Media posterior	Cuantil 2.5 %	Cuantil 97.5 %	valor-p
β_1	-5.7514	-8.4741	-3.2090	0.0002
β_2	0.0104	-0.0080	0.0282	0.1156
β_3	0.0125	-0.0349	0.0605	0.2858
β_4	0.0041	-0.0471	0.0552	0.4272
β_5	0.0219	-0.0021	0.0464	0.0344
β_6	0.0151	-0.0645	0.0939	0.3381
β_7	-0.0192	-0.0707	0.0333	0.2153
β_8	1.65e-05	-2.82e-06	3.53 e-05	0.0413
β_9	0.0060	-0.0436	0.0554	0.3999
ϕ	3.4379	0.0147	18.3510	0.0344

Solamente fue significativo (con significancia del 5 %) β_5 , asociado a porcentaje de afroamericanos y β_8 . Tenemos que al incrementar en una unidad porcentual el porcentaje de personas afroamericanas (e.g. de 20 a 21 % o de 28 a 29 %) incrementa en promedio 0.021 la tasa de crímenes per cápita en escala logarítmica. Por lo que $e^{0.021} \approx 1.02$, así que incrementa en 2% la tasa de crimen per cápita en promedio (aproximadamente)³.

Incrementar en una unidad el ingreso per cápita genera un incremento promedio de 1.65×10^{-5} unidades en la tasa de crimen per cápita en escala logarítmica.

³Este tipo de interpretaciones y manipulaciones asumen que podemos sacar el logaritmo de la esperanza, ¡lo cual no es del todo cierto! Pero funge como una vaga aproximación. Más adelante entraremos en una discusión más detallada de esto respectivo a la **Desigualdad de Jensen**!

$$\begin{aligned}\mathbb{E}(\log(Y_i)|\mathbf{X}_i) - \mathbb{E}(\log(Y_j)|\mathbf{X}_j) &= \mathbb{E}(Y(s_i)|\mathbf{X}_i) - \mathbb{E}(Y(s_j)|\mathbf{X}_j) \\ &= \beta_5\end{aligned}$$

Si \mathbf{X}_i es igual a \mathbf{X}_j excepto en la entrada correspondiente a la variable de porcentaje de afroamericanos, donde se tiene $\mathbf{X}_{4,i} = \mathbf{X}_{4,j} + 1$

7.3. Evaluación de las Predicciones

Este modelo tiene un DIC de 29.3, el cual es mayor al del primer modelo. La correlación de las predicciones puntuales con la variable respuesta fue de 0.881. Y elevando al cuadrado, se obtuvo una pseudo $R^2 = 0.776$, la cual es menor a la del modelo 1.

$$\begin{aligned}\mathbb{E}(\log(Y_i)|\mathbf{X}_i) - \mathbb{E}(\log(Y_j)|\mathbf{X}_j) &= \mathbb{E}(Y(s_i)|\mathbf{X}_i) - \mathbb{E}(Y(s_j)|\mathbf{X}_j) \\ &= \beta_5\end{aligned}$$

Si \mathbf{X}_i es igual a \mathbf{X}_j excepto en la entrada correspondiente a la variable de porcentaje de afroamericanos, donde se tiene $\mathbf{X}_{4,i} = \mathbf{X}_{4,j} + 1$

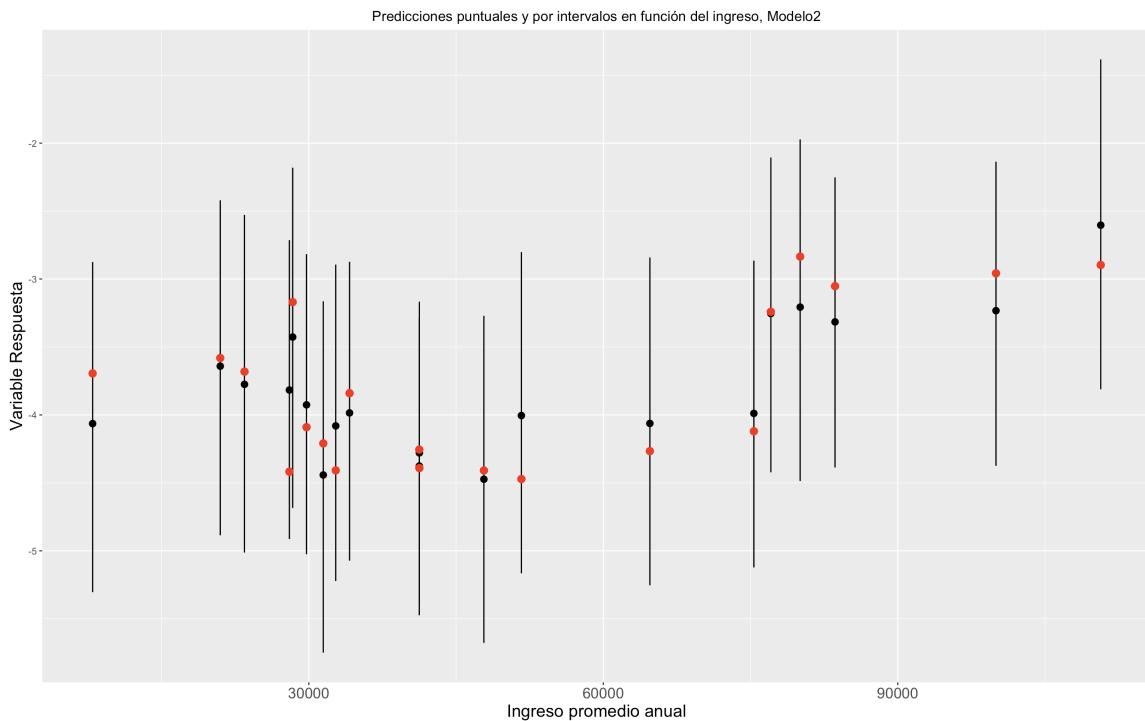


Figura 21: Predicción puntual y por intervalos al 95 % de probabilidad. De color negro tenemos predicciones y los puntos de color rojo corresponden a los valores observados de la variable respuesta (logaritmo de la tasa de crimen per cápita en 2022).

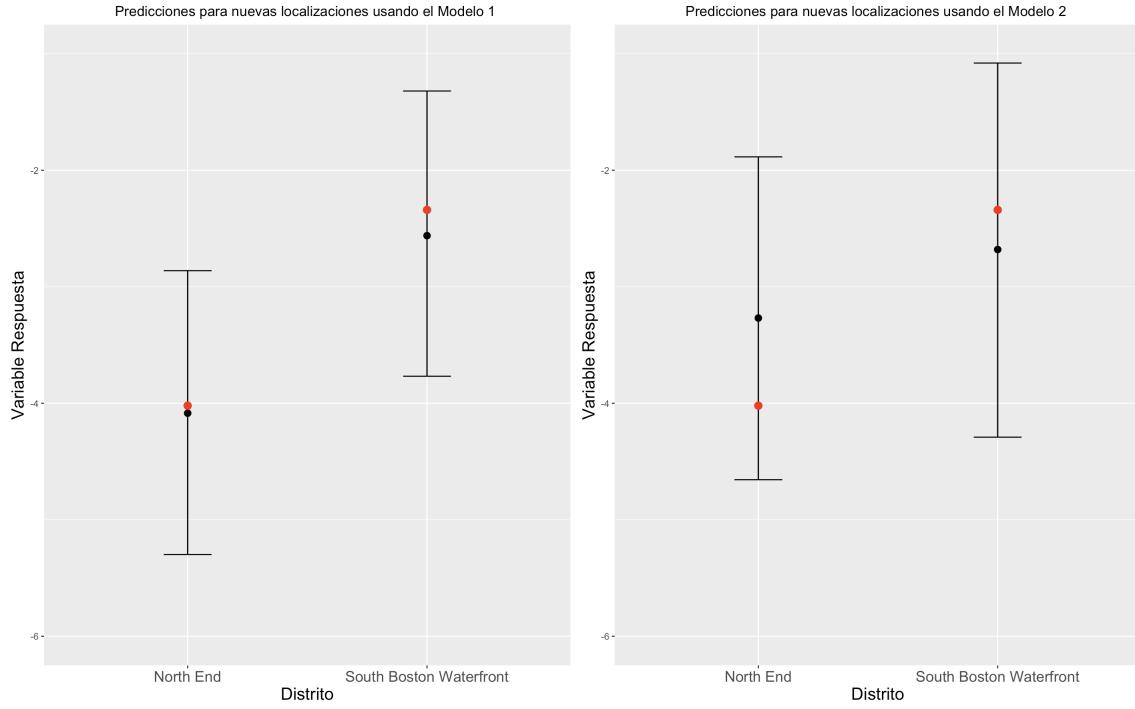


Figura 22: Comparación en las predicciones para nuevos distritos. Las asociadas al Modelo 1 están del lado izquierdo y las asociadas al modelo 2 del lado derecho. Bandas de predicción al 95 %. Puntos negros son estimaciones puntuales usando la distribución predictiva final. Los puntos rojos corresponden a las observaciones del logaritmo de tasa de crimen en cada distrito.

Esta figura nos permite apreciar que las predicciones puntuales y por intervalos del modelo 1 son mejores que las del modelo 2. Para *North End* la longitud del intervalo de predicción con el modelo 2 fue de 2.771 y para *South Boston Waterfront* la longitud del intervalo fue de 3.211. El modelo 1 nos dio intervalos con menores longitudes para ambas predicciones por intervalos, al mismo nivel de probabilidad. Algo que vale la pena notar es que (y de hecho es algo que se satisface en el modelo 1 y en el 2) la longitud asociada al intervalo de predicción para el valor extremo es mayor que la longitud del intervalo de predicción para *North End*. Desde una perspectiva intuitiva, esto hace sentido, pues predecir un valor extremo es una tarea más compleja, y por lo tanto sería de esperarse que tenga más incertidumbre asociada. Así que, bajo cualquier métrica o criterio que se utilice para evaluar la capacidad predictiva de los modelos, el modelo 1 es mejor que el modelo 2.

8. Modelo 3

La propuesta de este modelo es usar variables explicativas cuyos coeficientes hayan resultado significativos en los modelos anteriores. El predictor lineal que utilizaremos tiene la forma

$$\mu(s_i) = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \beta_4 X_{3,i} ,$$

donde las variables que utilizamos son:

X_1 =PorcentajeAfroamericanos,

X_2 = incomePerCapita,
 X_3 = Índice de educación (construido con PCA).

8.1. Monitoreo de convergencia de las cadenas (MCMC)

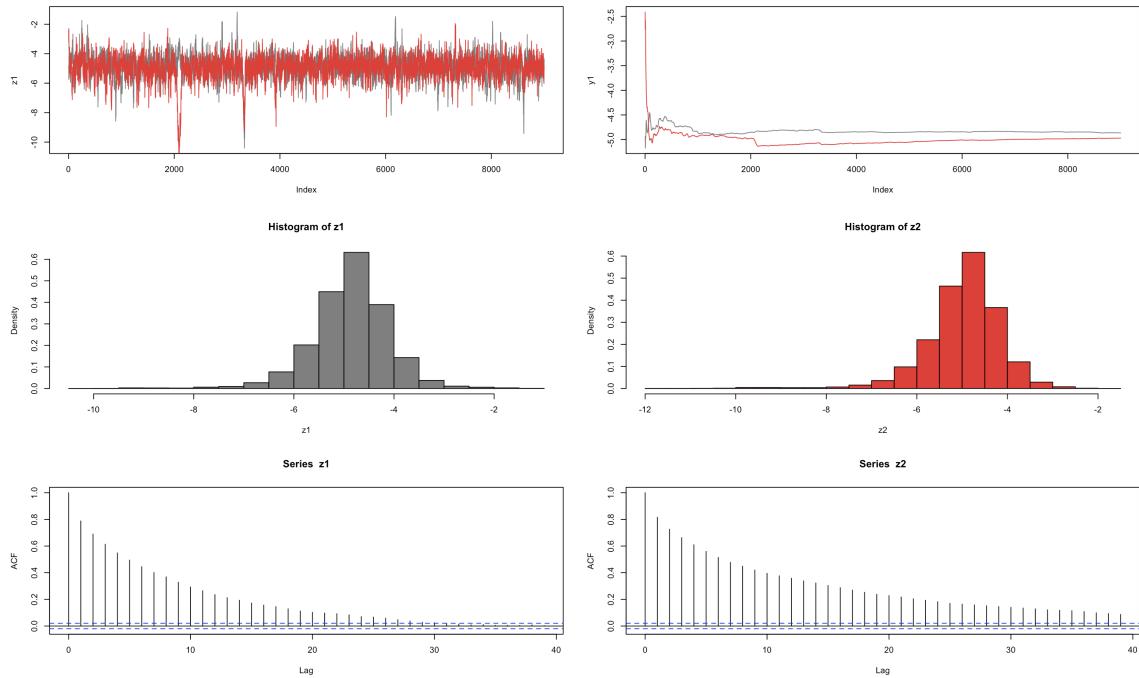


Figura 23: Monitoreo de las cadenas de β_1 para el tercer modelo.

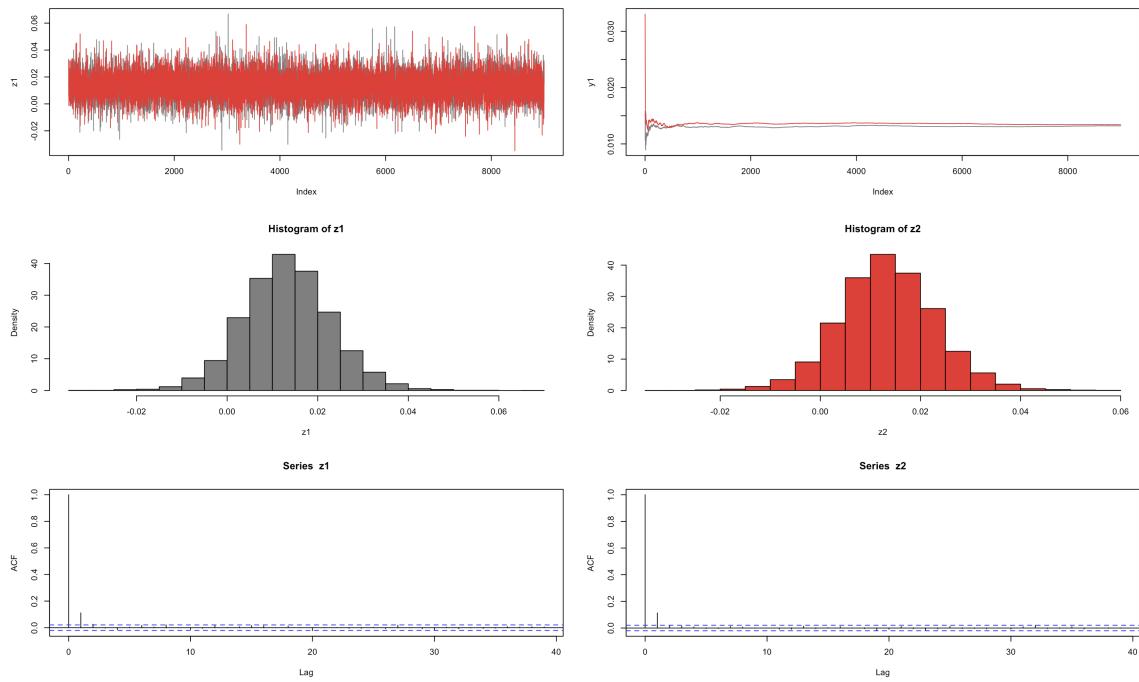


Figura 24: Monitoreo de las cadenas de β_2 para el tercer modelo.

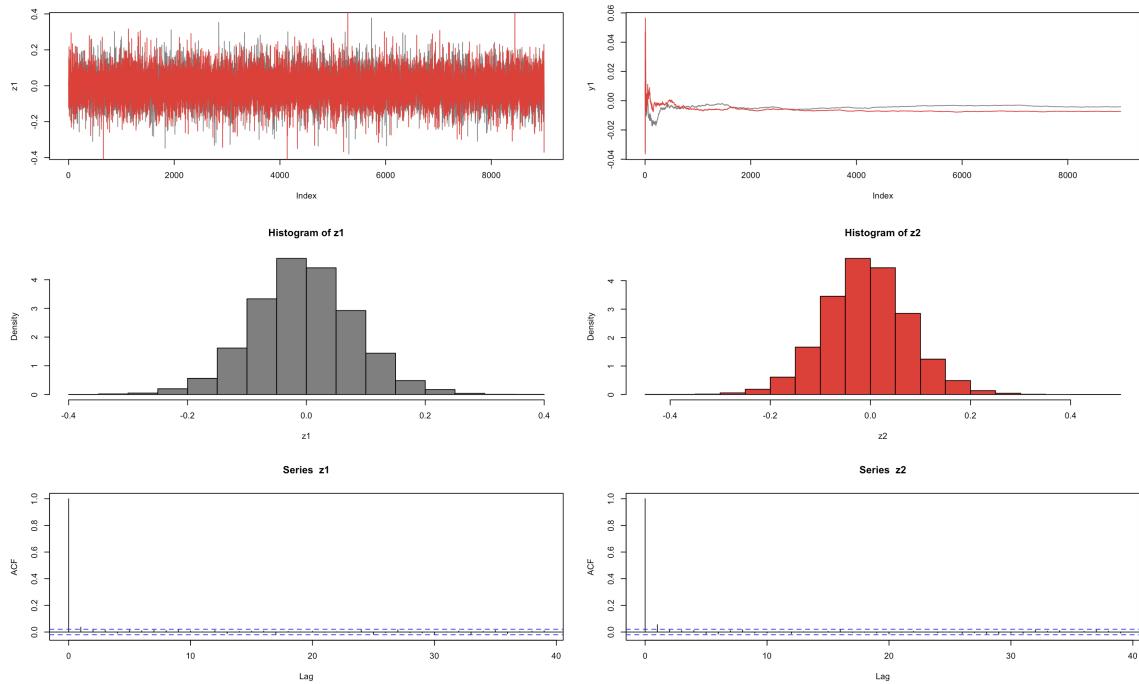


Figura 25: Monitoreo de las cadenas de β_4 para el tercer modelo.

8.2. Resumen, inferencia bayesiana e interpretación

Parámetro	Media	Cuantil 2.5 %	Cuantil 97.5 %	valor-p
β_1	-4.91754	-6.64205	-3.51600	0.00000
β_2	0.01330	-0.00554	0.03287	0.07417
β_3	0.00001	0.00000	0.00002	0.01828
β_4	-0.00575	-0.17190	0.16470	0.46767

Solo resultó ser significativo al 5% β_3 , asociada al ingreso per cápita. La interpretación es analoga a la que ya presentamos en los otros modelos. El intercepto no es interpretable, pues no tiene sentido un distrito con ingreso per cápita igual a cero.

8.3. Evaluación de las Predicciones

El modelo tuvo un DIC de 25.8. Correlación entre predicciones puntuales y variable respuesta fue de 0.85. Elevando al cuadrado, se obtuvo una pseudo $R^2 = 0.73$.

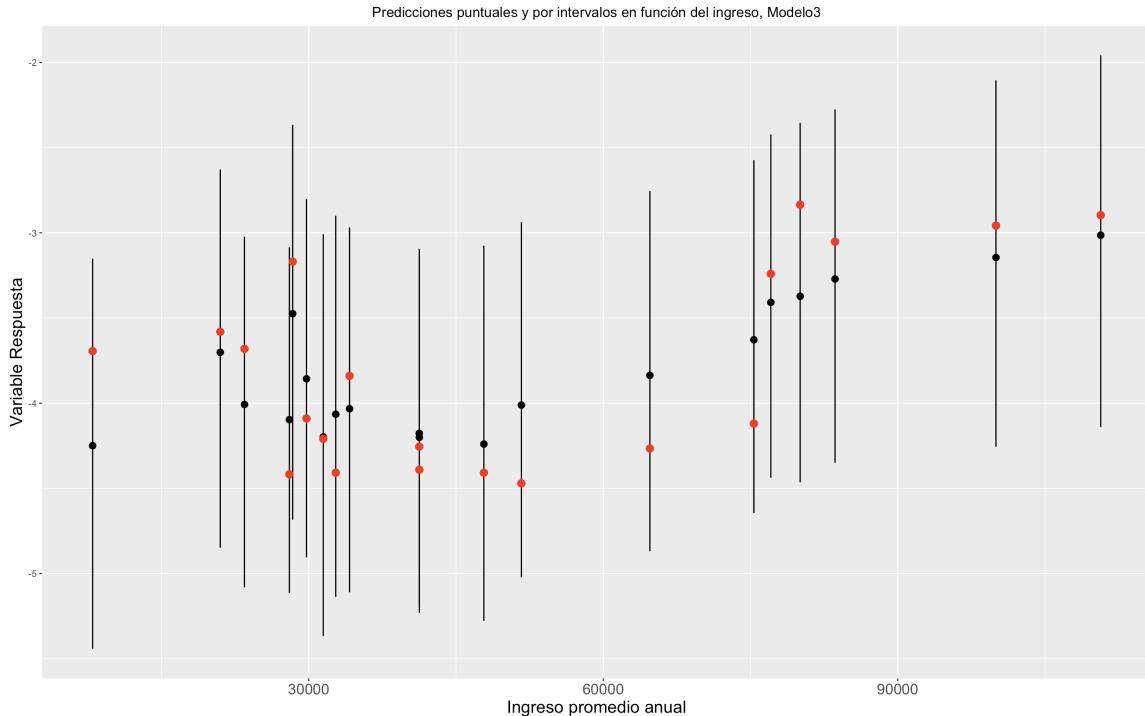


Figura 26: Predicciones del tercer modelo en función del ingreso promedio de los residentes en cada barrio. Negro son predicciones y rojo son los datos observados. Los intervalos son al 95 % de probabilidad.

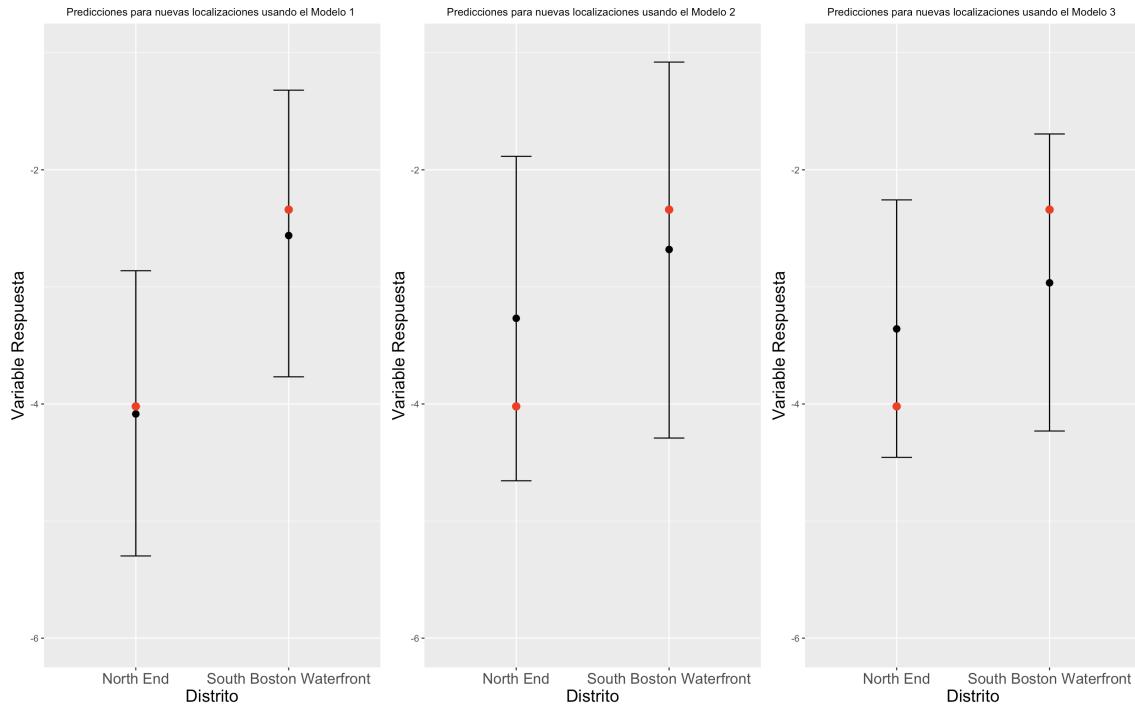


Figura 27: Comparación de predicciones de los tres modelos para las dos localizaciones nuevas. Modelo 1 del lado izquierdo. Modelo 2 en medio. Modelo 3 del lado derecho.

Longitudes de intervalos de predicción con el modelo 3 fueron: 2.199 para *North End* y 2.537 para *South Boston Waterfront*. Nótese que son menores a las obtenidas en con el modelo 2. Preferimos longitudes más cortas porque nos ofrecen más precisión al mismo nivel de probabilidad (95 % en este caso).

9. Discusión

Una buena práctica sería hacer predicciones sin basura, es decir, hacer las predicciones sobre las nuevas localizaciones habiendo quitado las variables que no resultaron significativas. Tratamos de remediar esa situación en el Modelo 3, para el cual una de las variables de hecho resultó no ser significativa, aunque dicha variable si lo fue para el primer modelo (índice de educación). Uno de los motivos por los cuales nos desviamos de esa práctica fue pragmatismo: el tiempo de computo fue un factor importante a considerar para lograr la realización de este trabajo en tiempo y forma. Correr estos modelos en OpenBUGS tomó mucho tiempo de máquina. Sin embargo, consideramos que es importante estar conscientes de esa situación.

9.1. ¿Cómo interpretar las predicciones de las tasas de crimen si están en escala logarítmica?

Una primera estrategia intuitiva es aplicar la función exponencial a todos nuestros resultados de predicciones puntuales e incluso a las cotas en los intervalos de predicción para así convertir todo a

términos originales. Por desigualdad de Jensen, si h es una función convexa y X es una variable aleatoria, entonces $h(\mathbb{E}(X)) \leq \mathbb{E}(h(X))$. En particular, pensemos en h la función dada por $x \rightarrow h(x) := e^x$, la cual es convexa. Esto quiere decir que cuando apliquemos la función exponencial a nuestras predicciones puntuales en escala logarítmica para obtener estimaciones de la media de la predictiva final (las cuales utilizamos como predictores puntuales) para las tasas de crimen, estas predicciones puntuales van a subestimar el valor real de la esperanza de la predictiva posterior. Teniendo ese tipo de cuestiones en mente, podemos transformar todo y llevarlo a una escala más interpretable. Respecto a los cuantiles tenemos que si X es una variable aleatoria y q_p denota un cuantil de orden p en la cola izquierda para la distribución de X , entonces

$$0.95 = \mathbb{P}(q_{0.025} \leq X \leq q_{0.975}) = \mathbb{P}(e^{q_{0.025}} \leq e^X \leq e^{q_{0.975}}),$$

pues la función exponencial es creciente por lo que no altera desigualdades.

Pero conviene tener presente ese tipo de sutilezas para no malinterpretar resultados y al mismo tiempo expresarlos en términos de nuestra variable de interés original.

Variable	Predicción puntual	$q_{0.025}$	$q_{0.975}$	Observación	Distrito	Modelo
$Y(s_{15})$	-4.09	-5.30	-2.86	-4.02	North End	Modelo 1
$Y(s_{19})$	-2.56	-3.77	-1.32	-2.34	South Boston Waterfront	Modelo 1
$Y(s_{15})$	-3.27	-4.66	-1.88	-4.02	North End	Modelo 2
$Y(s_{19})$	-2.68	-4.29	-1.08	-2.34	South Boston Waterfront	Modelo 2
$Y(s_{15})$	-3.36	-4.46	-2.26	-4.02	North End	Modelo 3
$Y(s_{19})$	-2.97	-4.23	-1.69	-2.34	South Boston Waterfront	Modelo 3

Cuadro 3: Predicciones en escala logarítmica

Recuérdese que en nuestra notación, $Y(s_i) = \log(Y_i)$, donde Y_i es la tasa de crimen per cápita en el distrito i

Entonces ahora aplicamos la función exponencial a los resultados numéricos. Así por ejemplo, nuestra predicción puntual para Y_{15} usando el primer modelo es $e^{-4.09} \approx 0.0167$ y en nuestra base de datos se tiene que la tasa de crimen per cápita en el distrito *North End* es de $0.0179 \approx e^{-4.02}$. Siguiendo ese razonamiento se puede interpretar la siguiente tabla.

Variable	Predicción puntual	$e^{q_{0.025}}$	$e^{q_{0.975}}$	Observación	Distrito	Modelo
Y_{15}	0.0168	0.0050	0.0572	0.0179	North End	Modelo 1
Y_{19}	0.0772	0.0231	0.2671	0.0963	South Boston Waterfront	Modelo 1
Y_{15}	0.0381	0.0095	0.1518	0.0179	North End	Modelo 2
Y_{19}	0.0685	0.0137	0.3396	0.0963	South Boston Waterfront	Modelo 2
Y_{15}	0.0348	0.0116	0.1047	0.0179	North End	Modelo 3
Y_{19}	0.0515	0.0145	0.1838	0.0963	South Boston Waterfront	Modelo 3

Cuadro 4: Predicción puntual y por intervalos para la tasa de crimen per cápita. Los intervalos de la forma $(e^{q_{0.025}}, e^{q_{0.975}})$ pueden interpretarse como intervalos de probabilidad al 95 % por la discusión anterior.

Nótese que las observaciones de las tasas de crimen para los dos distritos están contenidas en los intervalos que se reportan. Esto era algo que ya se podía notar en la figura 27, pero decidimos hacerlo explícito y analítico en tablas. De igual manera, se aprecia que el Modelo 1 tiene predicciones puntuales

que quedan más cerca de las tasas observadas.

10. Conclusiones

En el presente trabajo se exploró un modelo espacial bayesiano para explicar el porcentaje de crímenes en la ciudad de Boston. Las variables explicativas fueron variables demográficas, económicas, de educación y de diversidad étnica. Debido a que la complejidad del modelo también está en función a sus parámetros, a medida que aumenta la cantidad de parámetros, los modelos bayesianos aumentan de complejidad, y se vuelve difícil tanto la simulación como la interpretación. Es por ello que exploramos técnicas de reducción de dimensionalidad que resultaron ser efectivas al momento de explicar y predecir el porcentaje de crímenes en Boston. El modelo compuesto por los índices de edades, educación, movilidad, origen étnico, ingreso per cápita y tasa de pobreza resultó ser el mejor en DIC y para realizar los pronósticos, teniendo predicciones muy acertadas. Por otra parte, identificamos que es muy importante realizar PCA y verificar que se puede desarrollar una interpretación intuitiva para poderle dar significado a los coeficientes de la regresión en los modelos posteriores.

La construcción de índices vía PCA permitió incorporar mucha información de los distritos en pocas variables explicativas. El tercer modelo tuvo un mejor desempeño en cuanto a DIC y longitud de los intervalos de predicción en comparación al segundo. Como bien se señaló, el primer modelo fue mejor que los otros en prácticamente todos los criterios. Sin embargo, tal como ocurre comúnmente en los modelos de aprendizaje de máquina, esto viene a un costo: interpretabilidad. Los índices aportan mucha información y resultaron tener una capacidad predictiva muy buena, pero su interpretación no es tan sencilla como la de otras variables, pues por su construcción, los índices representan variables agregadas.

11. Bibliografía

Notas del curso de Regresión Avanzada del profesor Luis Enrique Nieto Barajas.