

# Proyecto final: Predicción de crímenes en la Ciudad de Boston

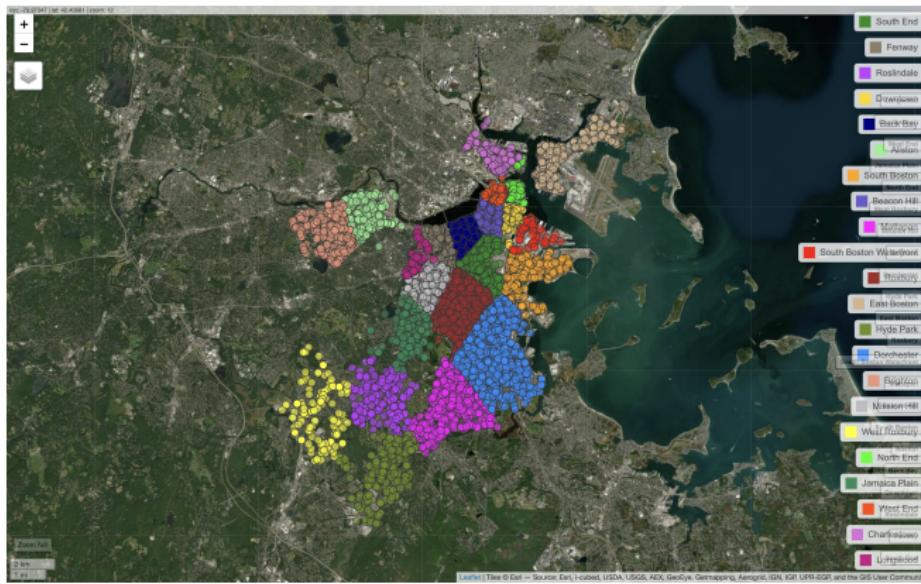
Joaquín Álvarez, Diego Velázquez y Marcelino Sánchez

Instituto Tecnológico Autónomo de México

23 de mayo de 2023

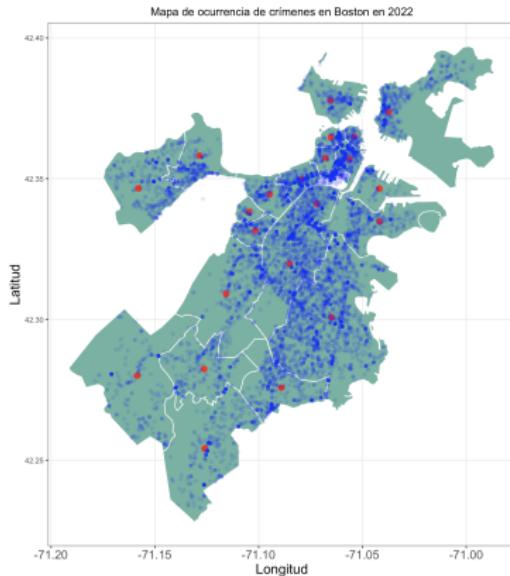
# Planteamiento

Se exploraron 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.



# Preprocesamiento de la variable respuesta

De la base de datos de crímenes en Boston consideramos los tipos de crímenes más relevantes. Posteriormente, a cada crimen le asociamos el distrito más cercano. Finalmente, se determinaron porcentajes.



**Figure:** Los puntos rojos son coordenadas de cada distrito. Los puntos azules son localizaciones donde se reportó alguno de los tipos de crimen.

# Descripción de la variable respuesta

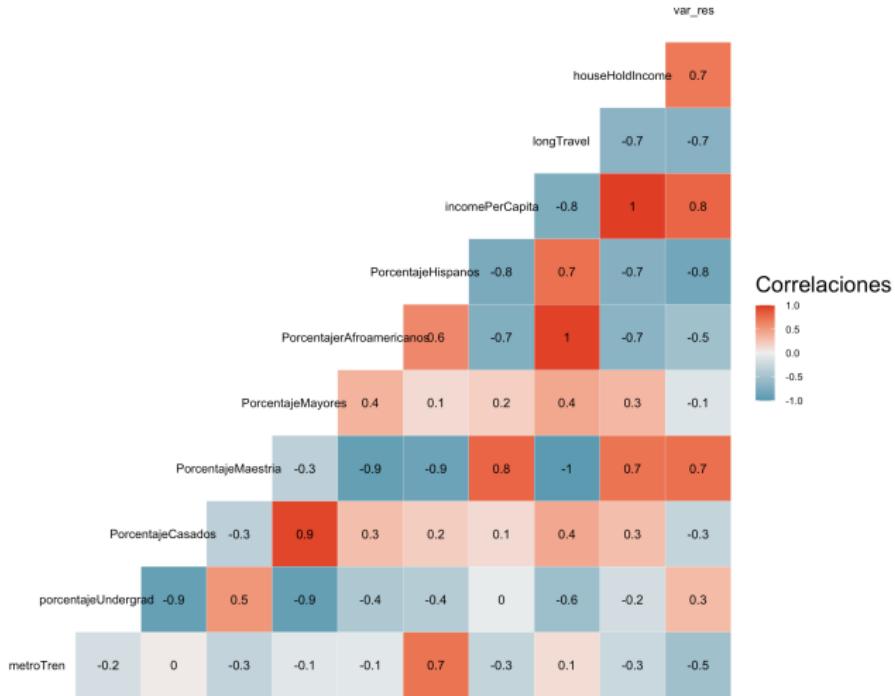
Variable	Descripción
Población ( $M_i$ )	Número de habitantes en el distrito según el censo de 2019
Crímenes ( $Z_i$ )	Número de crímenes reportados en 2022 en el distrito
Tasa de crimen ( $y_i$ )	Número de crímenes por habitante en el distrito en el año 2022. Se calcula como $y_i = \frac{Z_i}{M_i}$
$\ln(\text{Tasa de crimen}) (Y_i)$	Variable que se usó para ajustar. Se calcula como $Y_i = \ln\left(\frac{Z_i}{M_i}\right)$

# Descripción de las variables explicativas

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

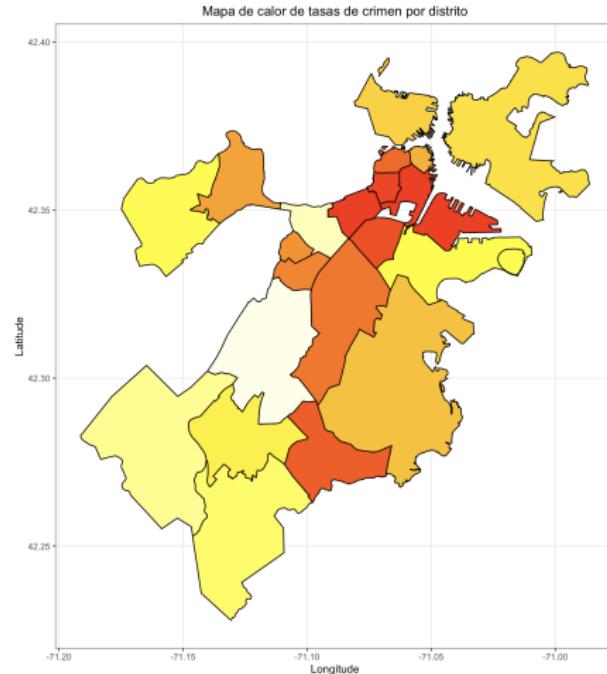
# Análisis exploratorio de los datos

## Matriz de correlaciones entre variables



# Análisis exploratorio de los datos

## Mapa de calor sobre la tasa de crímenes



**Figure:** 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.

# Análisis de componentes principales

## PCA

La enorme cantidad de variables explicativas proporcionadas por los datos y la necesidad de implementar modelos que corrieran en tiempos razonables nos hizo utilizar técnicas de reducción de dimensionalidad. 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.

# Análisis de componentes principales

## Biplots

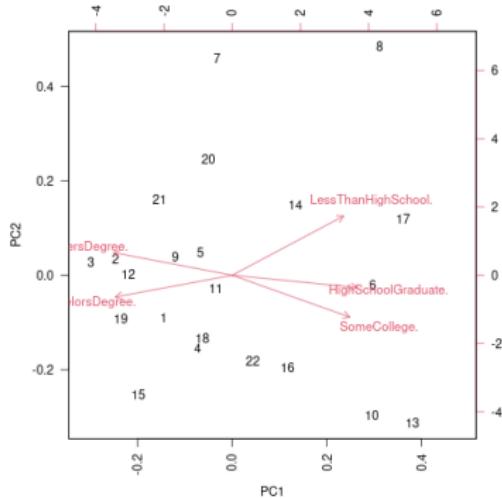


Figure: Biplot a educación

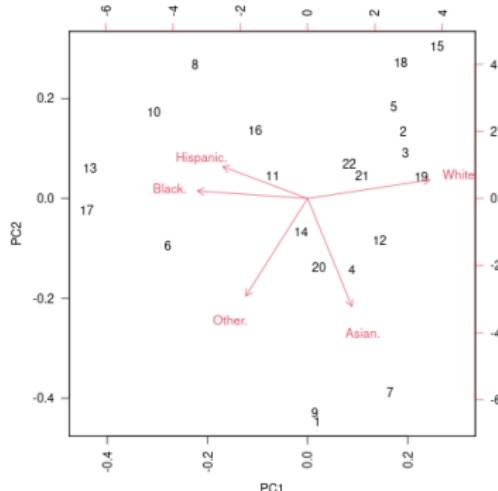
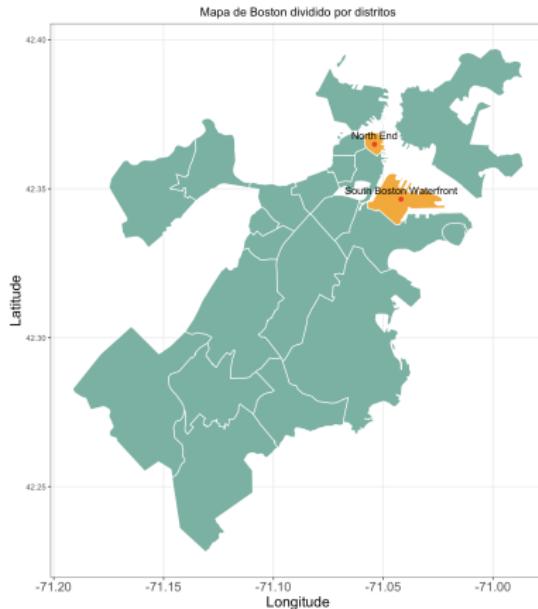


Figure: Biplot a raza

# Preparación para ajustar modelos

## Conjuntos de entrenamiento y aprendizaje



**Figure:** 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.

# Modelo espacial para datos referenciados puntualmente

## Parte 1

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

con  $Y(s) = \begin{pmatrix} Y(s_1) \\ Y(s_2) \\ \vdots \\ Y(s_n) \end{pmatrix}, \mu(s) = \begin{pmatrix} \mu(s_1) \\ \mu(s_2) \\ \vdots \\ \mu(s_n) \end{pmatrix}, \mu(s_i) = x(s_i)^T \underline{\beta},$

$$w(s) = \begin{pmatrix} w(s_1) \\ w(s_2) \\ \vdots \\ w(s_n) \end{pmatrix} \sim \mathcal{N}_n(\underline{0}, \Sigma(s, t))$$
 denota un proceso Gaussiano

donde  $Cov(w(s_i), w(s_j)) = \Sigma(s_i, s_j) = \sigma^2 \exp\{-\phi d_{i,j}\}, d_{i,j} = \|s_i - s_j\|_2$

# Modelo espacial para datos referenciados puntualmente

## Parte 2

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

Como consecuencia de esta especificación se sigue que

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

# Modelo 1

## Estructura del predictor lineal

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,

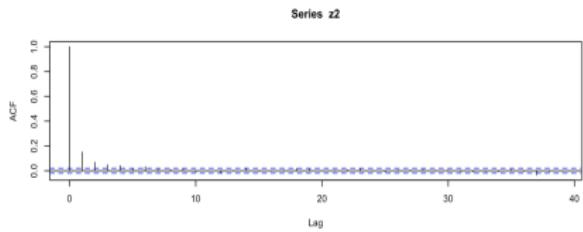
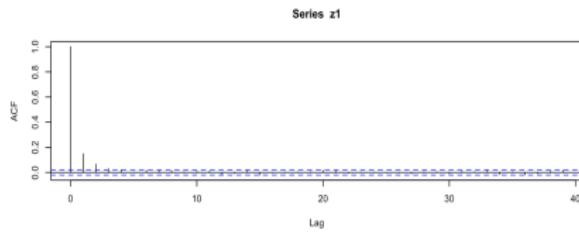
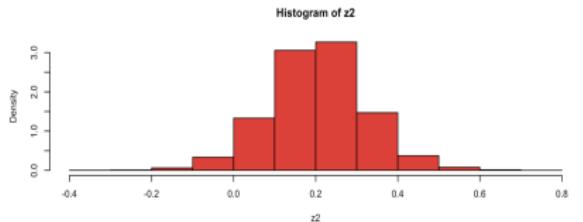
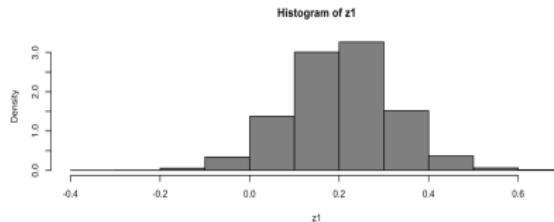
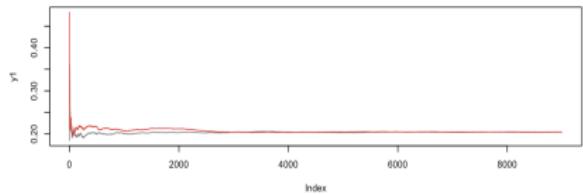
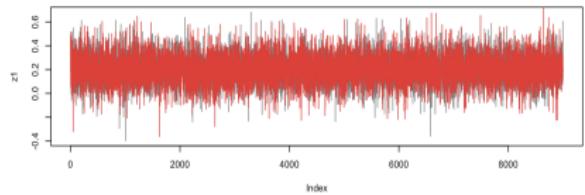
$X_4$  = Índice de origen étnico,

$X_5$  = Ingreso per cápita,

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

# Modelo 1

## Monitoreo de las cadenas de $\beta_3$



# Modelo 1

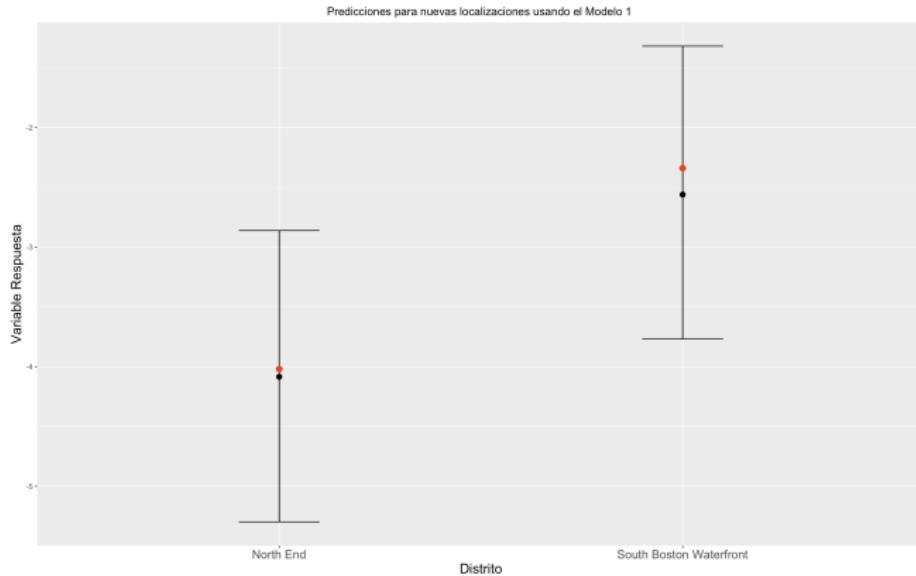
Tabla resumen

Parámetro	Media pos-terior	Cuantil 2.5%	Cuantil 97.5%	valor-p
$\beta_1$	-5.0756	-6.8410	-3.3190	0.0000
$\beta_2$	0.0558	-0.3262	0.4580	0.3883
$\beta_3$	0.2045	-0.0287	0.4367	0.0404
$\beta_4$	-0.2850	-0.6856	0.0993	0.0696
$\beta_5$	-0.0207	-0.2517	0.2129	0.4232
$\beta_6$	0.0000	-0.0000	0.0000	0.0319
$\beta_7$	0.4499	-4.3112	5.3460	0.4278
$\phi$	3.4287	0.0218	17.6502	0.0000

DIC= -54.33

# Modelo 1

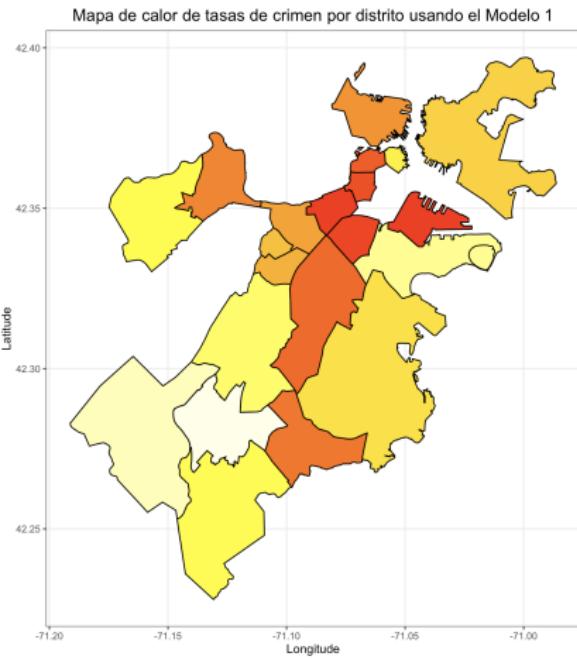
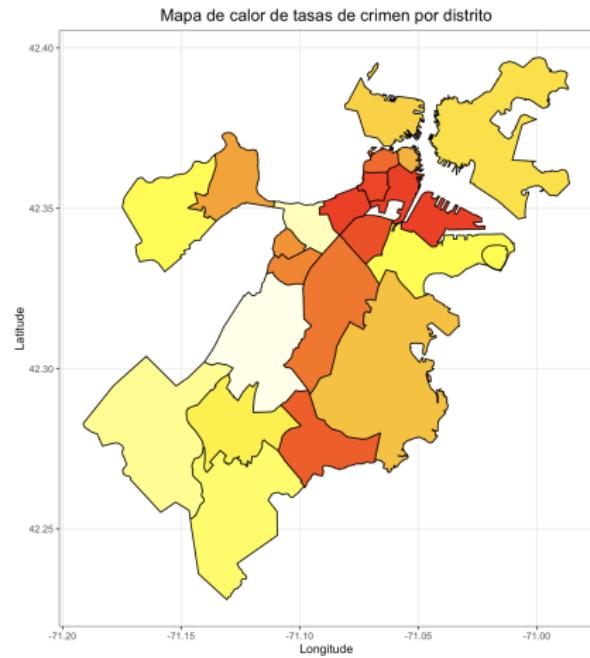
## Predicciones



**Figure:** 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.

# Modelo 1

## Mapa de calor comparativo



# Modelo 2

## Estructura del predictor lineal

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,

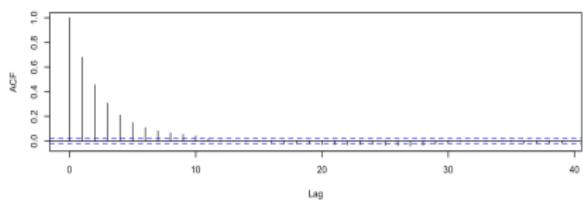
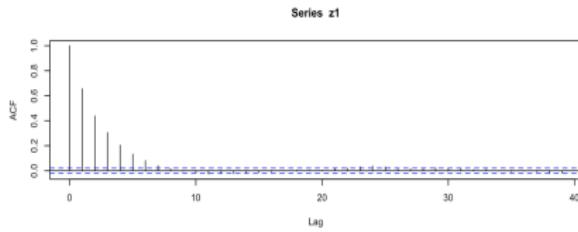
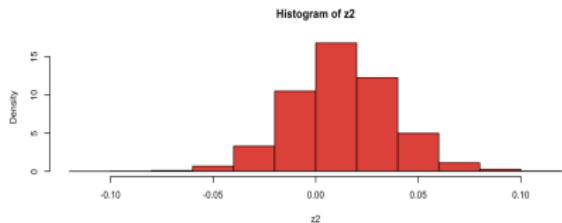
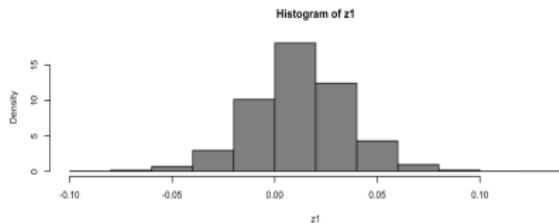
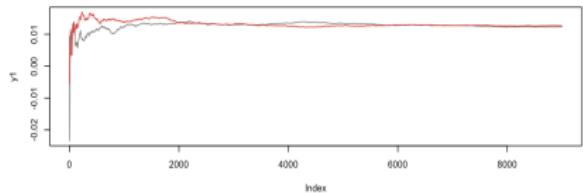
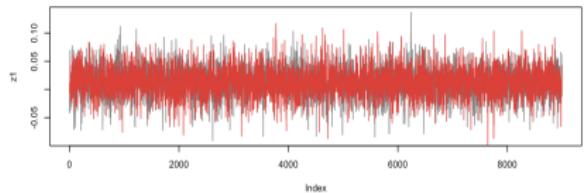
$X_6$  =PorcentajeCasados,

$X_7$  =incomePerCapita,

$X_8$  =metroTren

# Modelo 2

## Monitoreo de las cadenas de $\beta_3$



# Modelo 2

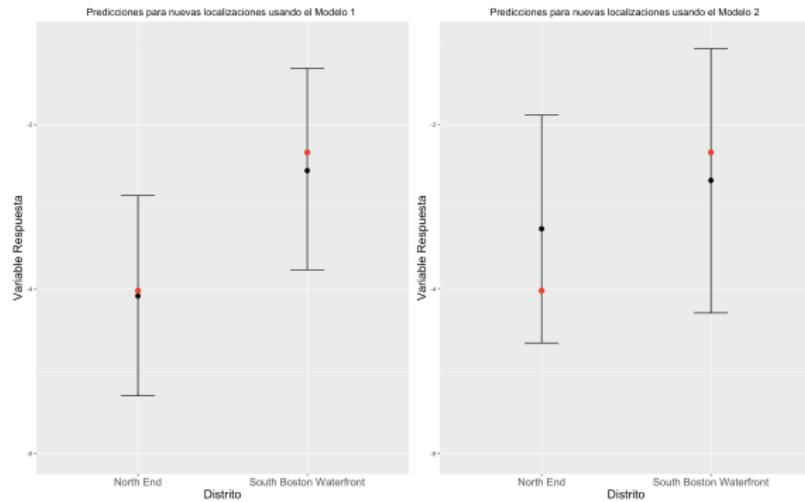
Tabla resumen

	mean	2.5%	97.5%	prob
$\beta_1$	-5.7514	-8.4741	-3.2090	0.0002
$\beta_2$	0.0104	-0.0080	0.0282	0.1156
$\beta_3$	0.0125	-0.0349	0.0605	0.2858
$\beta_4$	0.0041	-0.0471	0.0552	0.4272
$\beta_5$	0.0219	-0.0021	0.0464	0.0344
$\beta_6$	0.0151	-0.0645	0.0939	0.3381
$\beta_7$	-0.0192	-0.0707	0.0333	0.2153
$\beta_8$	1.65e-05	-2.82e-06	3.53 e-05	0.0413
$\beta_9$	0.0060	-0.0436	0.0554	0.3999
$\phi$	3.4379	0.0147	18.3510	0.0344

DIC=29.3

# Modelo 2

## Comparación de predicciones con el modelo 1



**Figure:** 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.

# Modelo 3

## Estructura del predictor lineal

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:

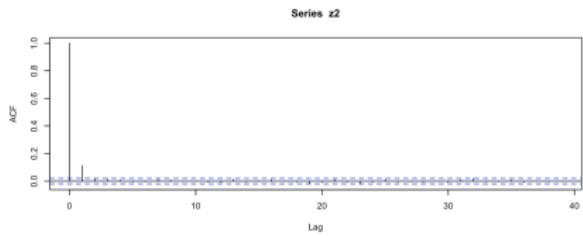
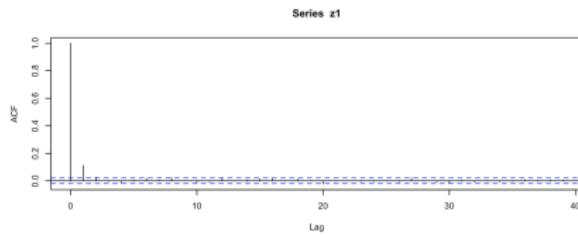
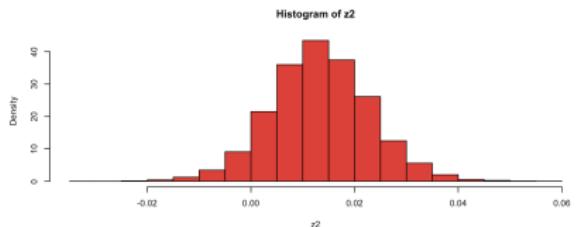
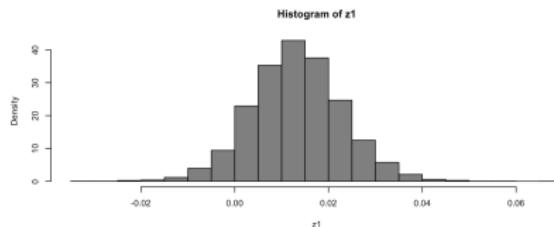
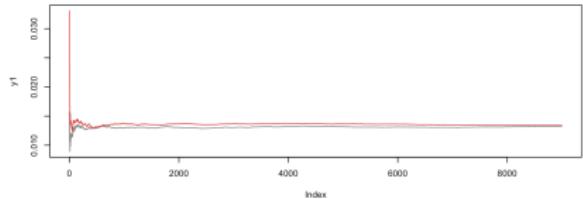
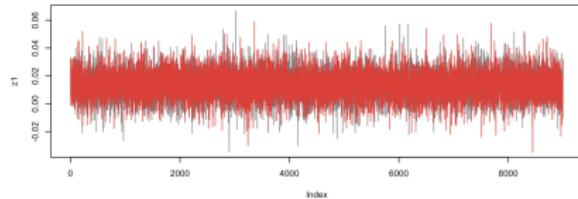
$X_1$  =PorcentajeAfroamericanos,

$X_2$  =incomePerCapita,

$X_3$  =Índice de educación (construido con PCA).

# Modelo 3

## Monitoreo de las cadenas de $\beta_2$



# Modelo 3

## Tabla resumen

Parámetro	Media	Cuantil 2.5%	Cuantil 97.5%	valor-p
$\beta_1$	-4.91754	-6.64205	-3.51600	0.00000
$\beta_2$	0.01330	-0.00554	0.03287	0.07417
$\beta_3$	0.00001	0.00000	0.00002	0.01828
$\beta_4$	-0.00575	-0.17190	0.16470	0.46767

DIC=25.8

# Modelo 3

Comparación de predicciones de los tres modelos para las dos localizaciones nuevas.

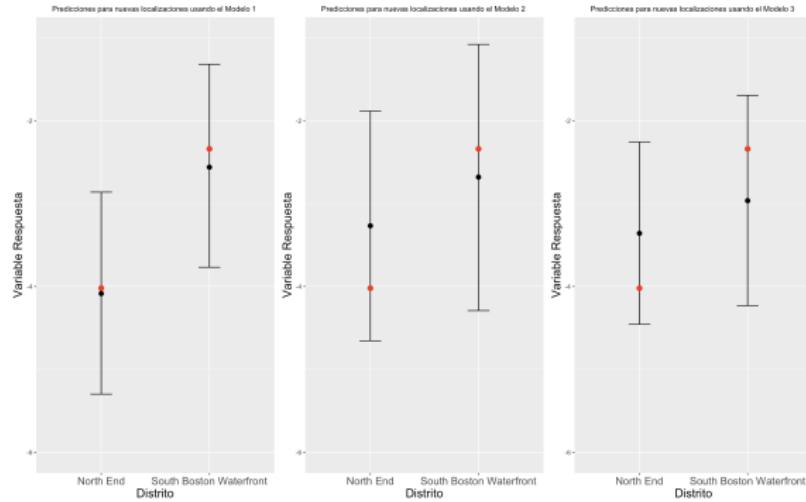


Figure: Modelo 1 del lado izquierdo. Modelo 2 en medio. Modelo 3 del lado derecho.

# Interpretación de resultados predictivos

Predicción en escala original

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	SBW	
$Y_{15}$	0.0381	0.0095	0.1518	0.0179	North End	Modelo 2
$Y_{19}$	0.0685	0.0137	0.3396	0.0963	SBW	
$Y_{15}$	0.0348	0.0116	0.1047	0.0179	North End	Modelo 3
$Y_{19}$	0.0515	0.0145	0.1838	0.0963	SBW	

**Table:** 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%

