ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

PROCESOS PUNTUALES ESPACIO TEMPORALES CON APLICACIÓN A LA MODELIZACIÓN DE CRIMEN

ANÁLISIS TEMPORAL Y ESPACIAL DE ROBOS EN LA CIUDAD DE VALENCIA - ESPAÑA

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERA MATEMÁTICA

KATHERYN ALEXANDRA YANES CHANALATA

katheryn.yanes@epn.edu.ec

DIRECTOR: YANDIRA DENISSE CUVERO CALERO

yandira.cuvero@epn.edu.ec

DMQ, AGOSTO 2023

CERTIFICACIONES

Yo, KATHERYN ALEXANDRA YANES CHANALATA, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

Katheryn Alexandra Yanes Chanalata

Certifico que el presente trabajo de integración curricular fue desarrollado por Katheryn Alexandra Yanes Chanalata, bajo mi supervisión.

Yandira Denisse Cuvero Calero
DIRECTOR

.

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como los productos resultantes del mismo, son públicos y estarán a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

Katheryn Alexandra Yanes Chanalata

Yandira Denisse Cuvero Calero

RESUMEN

Valencia, la tercera ciudad más poblada de España, centra principalmente su economía en actividades industriales y agrícolas. Registra, según el Ministerio del Interior, los robos como el delito de mayor incidencia entre el 2010 y 2019. Se ajusta una serie temporal con frecuencia mensual y un modelo econométrico espacial de la forma $\hat{y} = X\hat{\beta}+E\hat{\gamma}$, donde X será la matriz de covariables y E la matriz de adyacencia. Con esta metodología se analiza la situación de seguridad de Valencia.

ESF o filtrado de vectores espaciales, en español, es una metodología que permite analizar la dependencia espacial de las variables en el modelado, corrigiendo errores de exogeneidad residual debido a autocorrelación espacial. Consiste en tomar un subconjunto de vectores propios de la matriz de adyacencia y añadirlos al modelo de regresión lineal como variables predictoras adicionales. La selección de los vectores propios se la realiza con ayuda del método Lasso y la estimación del parámetro de ajuste del método con el estadístico de la prueba I de Moran que corresponde a la contribución de Cherodian del presente año. El modelo econométrico espacial se ajusta a través del método Lasso, el algoritmo de Cherodian y la metodología ESF.

Palabras clave: Robos en Valencia, Econometría Espacial, Filtrado de vectores espaciales, Método Lasso, Método Cherodian.

ABSTRACT

Valencia, the third most populous city on Spain, base its economy on industrial and agricultural activities. According to the Ministerio del Interior, the robberies is the crime with the highest incidence between 2010 and 2019. A time series is fitted with a monthly frequency and a spatial econometric model of the form $\hat{y} = X\hat{\beta} + E\hat{\gamma}$, where X will be the covariate matrix and E the adjacency matrix. This metodology is used to study Valencia's security events.

Eigenvector Spatial Filtering (ESF), is a methodology that allows analyzing the spatial dependence of covariables in model. It consists of taking a subset of eigenvectors from E and adding them to the linear regression model as additional predictor variables. The selection of the eigenvectors is carried out by the help of the Lasso method and the estimation of the adjustment parameter of the method with the Moran's I test statistic that corresponds to the Cherodian contribution of this year.The spatial econometric model is adjusted through the Lasso method, the Cherodian algorithm and the ESF methodology.

Keywords: Thefts in Valencia, Spatial Econometrics, Eigenvector Spatial Filtering, Lasso Method, Cherodian Method.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Descripción del componente desarrollado

1.1. Objetivo general

Analizar temporal y espacialmente los robos en la ciudad de Valencia entre los años 2010-2019 utilizando un modelo temporal y la metodología Mi Lasso de Rowan Cherodian en la estimación de un modelo econométrico espacial.

1.2. Objetivos específicos

1. Analizar espacial y temporalmente el registro de robos en la ciudad de Valencia, buscando identificar zonas que registran mayor número de robos en la ciudad de Valencia.

2. Estimar un modelo temporal a través de la metodología de Box-Jenkins para pronosticar el conteo anual de robos en la ciudad de Valencia.

3. Desarrollar un modelo econométrico espacial a través del método Mi Lasso, Lasso y el filtrado espacial de vectores propios.

1.3. Alcance

El presente trabajo pretende analizar los robos registrados en la ciudad de Valencia, España. El análisis temporal es de frecuencia mensual mientras que el análisis espacial es anual con horizonte de tiempo de 10 años que va desde 2010 hasta el 2019.

1.4. Marco teórico

1.4.1. Grafo

Es una estructura matemática compuesta de un conjunto de vértices V y uno de aristas E. En el análisis espacial, se usan los grafos para representar la estructura espacial de los datos, siendo los vértices elementos que identifican a cada región y las aristas representan las conexiones de cada región con zonas contiguas.

1.4.2. Matriz de adyacencia

Matriz booleana simétrica en la que las filas y columnas representan los vértices del grafo. Sus componentes son:

$$a_{ij} = \begin{cases} 1 & \text{si los vértices son adyacentes entre sí} \\ 0 & \text{caso contrario} \end{cases}$$

En el análisis espacial esta matriz permite representar la estructura espacial de los datos al mostrar las relaciones existentes entre las divisiones territoriales.

1.4.3. Modelo de regresión lineal

Es un modelo matemático que permite describir una variable respuesta o dependiente como función de otras variables independientes o predictoras. Su ecuación es:

$$\hat{y} = \hat{\beta}X + \varepsilon$$

donde

$\hat{\beta}$: son las estimaciones de los parámetros de la regresión.

ε : representa el término de error en la regresión.

X : es la matriz de variables independientes en la regresión.

1.4.4. Filtrado Espacial de Vectores Propios

Desarrollado por Griffith y Chun, por sus siglas en inglés ESF (Eigenvector Spatial Filtering) es un enfoque que incluye la relación espacial en el modelado. Consiste en seleccionar un subconjunto de vectores propios de la matriz de adyacencia espacial y agregarlos al modelo de regresión como nuevas variables independientes.

Este análisis toma una variable que tiene correlación espacial; es decir, una variable en la que las observaciones dependen o tienen influencia de sus vecinos geográficos; y la transforma en una variable espacialmente independiente al eliminar el patrón espacial incrustado. Este patrón se elimina al dividir la información de la variable en dos partes, un componente espacial filtrado y un componente residual no espacial, de ese modo se puede analizar a la variable sin tener en cuenta la componente espacial [4].

ESF especificado a los modelos de regresión lineal

Esta metodología se aplica a modelos de regresión que tienen correlación espacial. Los residuos de los modelos de regresión que tienen correlación espacial generan problemas en la regresión pues el supuesto de exogeneidad de los residuos no se cumple. En consecuencia, ESF separa los residuos autocorrelacionados espacialmente en su componente espacial y no espacial al tomar un subconjunto parsimonioso de vectores propios y añadirlos al modelo como variables predictoras adicionales escribiendo al modelo como:

$$Y = \beta_X X + \gamma_k E_k + \varepsilon \tag{1.1}$$

donde:

- X es una matriz de dimensión $n \times p+1$ que contiene n observaciones con p covariables y el intercepto.
- β_X es un vector de dimensión $p+1$ que contiene los coeficientes de los parámetros de la regresión.
- E_k es una matriz de dimensión $n \times k$ que contiene los k vectores propios añadidos al modelo.
- γ_k es un vector de dimensión k que contiene los coeficientes de los parámetros espaciales de la regresión.
- $\varepsilon \sim N(0, I\sigma^2)$ es el vector de dimensión n que contiene los errores de la regresión, que se busca sean variables aleatorias normales independientes e idénticamente distribuidas [4].

La principal ventaja de esta técnica, en comparación con los métodos convencionales de máxima verosimilitud y el método de momentos, es que el investigador no necesita especificar explícitamente las partes en que el modelo esta espacialmente correlacionado o estimar los parámetros espaciales correspondientes. Esta característica del ESF resulta bastante útil, pues permite determinar si existe un proceso espacial con la ayuda de una prueba de correlación espacial, sin necesidad de tener conocimiento de la forma original del modelo.

En cambio, el inconveniente que presenta este método ESF, es que la descomposición espectral de la matriz de adyacencia produce n vectores propios, los cuales al ser incluidos en la regresión generan un modelo lineal en el que la cantidad de variables independientes espaciales depende de la división del territorio haciendo infactible o lenta la estimación por mínimos cuadrados ordinarios conforme la división aumenta.

Sin embargo, Griffith [4] argumenta que un subconjunto de los vectores propios es suficiente para eliminar la dependencia espacial en la variable dependiente, entonces lo que interesa es saber qué vectores propios se debe escoger. Para esta selección, se han propuesto algoritmos de selección hacia adelante, que iterativamente agregan los vectores propios hasta alcanzar el R^2 ajustado deseado.

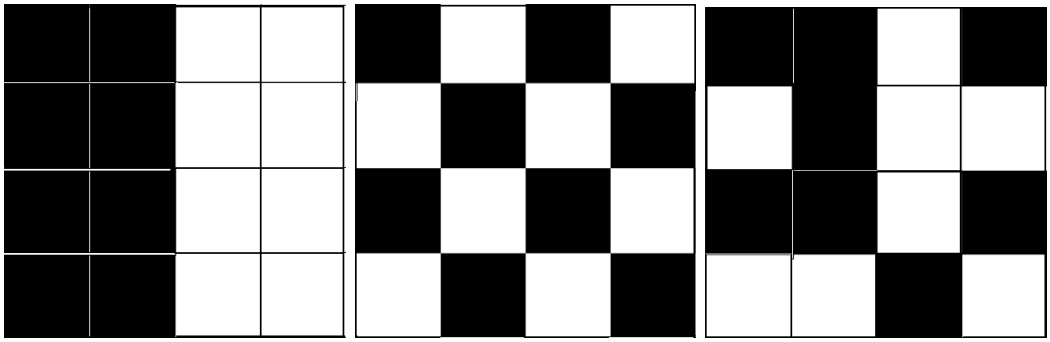
Otro problema mencionado por Seya [7], es que se puede dar el caso de que la mayoría de los vectores propios sean cero. Por tanto, su selección

será escasa. Recomendando utilizar una regresión penalizada con l_1 que justamente es el método Lasso.

1.4.5. I de Moran

Es una medida estadística, desarrollada en 1950 por Pierce Moran, que determina la ausencia o presencia de autocorrelación espacial de una variable [1].

El índice de Moran varía entre valores de -1 a 1, donde: 1 indica que existe autocorrelación positiva perfecta o agrupamiento perfecto de valores similares 1.1a; mientras que -1 indica, autocorrelación negativa o dispersión perfecta 1.1b; y un valor de 0 en cambio, indica la existencia de patrones completamente aleatorios 1.1c en la distribución espacial [3].



(a) Agrupamiento Perfecto (b) Dispersión Perfecta (c) Aleatoriedad Perfecta

Figura 1.1: I de Moran

Este índice se calcula a través de la fórmula:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

donde:

- n es el número de unidades geográficas en el mapa.
- W_{ij} es la matriz de distancia que nos indican si las áreas geográficas i y j son contiguas entre sí o no.

El índice de Moran se ajusta a una prueba estadística de significancia estadística de valores Z, suponiendo que esta sigue una distribución normal. El calculo de esa fórmula da una regla de decisión que valida la hipótesis de investigación de la siguiente forma:

Ho : El índice de Moran es igual a 0, no existe autocorrelación parcial

$H1$: El índice de Moran es distinto de 0, existe autocorrelación espacial.

1.4.6. Relación entre I de Moran y ESF

Según Griffith[4], el método ESF se basa en el estadístico I de Moran para calcular la correlación espacial de los residuos. Asumiendo que los residuos de la regresión $y = X\beta + \varepsilon$ son $M_X y = \widehat{u}$ entonces el I de Moran es:

$$I = \frac{y' M_X W M_X y}{y' M_X y} = \frac{\widehat{u}' W \widehat{u}}{\widehat{u}' \widehat{u}}$$

donde:

- $M_X = I - X(X'X)X'$
- W es una matriz de adyacencia simétrica de valores reales $n \times n$.

Además, Griffith [4] explica que cada uno de los n vectores propios, calculados a partir de la matriz de adyacencia representan patrones espaciales mutuamente ortogonales y solo un subconjunto de vectores propios será relevante para el modelo, es decir, en los modelos de regresión, solo un subconjunto de vectores propios tendrá coeficientes distintos de cero.

1.4.7. Modelo LASSO

El método de regresión LASSO proveniente de las siglas en inglés *Least Absolute Shrinkage and Selection Operator* desarrollado por Robert Tibshirani en 1996 [8], el cual, combina el método de mínimos cuadrados y una restricción de la norma l_1. Por tanto, el método LASSO resuelve el

problema:

$$\min_{\beta_0,\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}\left\|y-X\beta_0-X\beta\right\|_2^2\right\}$$
$$\text{s.a.}\sum_{j=1}^{p}\|\beta\|_1\leq\theta$$

donde

- y es la variable dependiente
- X es la matriz con las variables independientes
- β_0 es el intercepto
- β son los parámetros de la regresión
- θ es un parámetro de ajuste

El parámetro θ entonces permite controlar la cantidad de contracción que se aplica a las estimaciones, así, si $\theta = 0$ entonces la solución de Lasso se reduce a la de mínimos cuadrados ordinarios, mientras que si θ es lo suficientemente grande, el vector de parámetros penalizados se reduce a cero y no se seleccionan valores propios. Por el contrario, valores más moderados de θ hacen que algunos de los parámetros se reduzcan a cero y otros sean cero [2].

1.4.8. Mi Lasso

Rowan Cherodian propone una alternativa para resolver este problema de calibración de los parámetros del método Lasso usando la información proporcionada por el estadístico de i de Moran llamado Moran's i Lasso (Mi Lasso) [2]. Como ESF elimina cualquier patrón de correlación espacial en las regresiones lineales y esta información esta en los residuos del modelo $\widehat{u}$, Cherodian [2] propone usar esta información para determinar una estimación puntual de θ.

Para el test de correlación espacial, plantea usar la prueba I de Moran estandarizada, pues se la puede usar en muestras pequeñas (Kelejian y

Piras, 2017) y tiene buen poder si se la compara con modelos autorregresivos y diferentes distribuciones residuales (Anselin y Rey, 1991).

Si el nivel de la correlación espacial de los residuos es bajo, entonces es necesario usar un subconjunto pequeño de vectores propios de la matriz de pesos y un alto nivel de regularización. En cambio, si la correlación espacial de los residuos es alta, se requiere un subconjunto grande de vectores propios de la matriz de pesos y un bajo nivel de regularización. Se define entonces el estimador Moran i Lasso (Mi Lasso) como:

$$\theta = \frac{1}{Z^{\alpha}} \tag{1.2}$$

donde:

- $Z = |z| \quad \text{con} \quad z = \left(\frac{m - E(m)}{\sqrt{Var(m)}} \right)$
- $E(m) = \frac{tr(M_X W M_X)}{n-k}$
- $Var(m) = \frac{2\left((n-k)tr((M_X W M_X)^2) - [tr(M_X W M_X]^2\right)}{(n-k)^2(n-k-2)}$
- α es una constante positiva, $\forall z \neq 0$

El valor de z representa la correlación, y si es alta o baja. Por tanto, el método Mi Lasso recomienda usar el inverso del valor absoluto de z de los residuos, como estimación del parámetro θ para el modelo Lasso [2].

Procedimiento del método Mi Lasso

Para realizar el ajuste del modelo econométrico $y = \beta_X X + \gamma_k W_k + \varepsilon$:

1. Verificar si la matriz de adyacencia W es simétrica.

   a) En caso de no ser simétrica, se la obtiene de la siguiente manera:
   $$H = \frac{1}{2}(W \times W^T)$$

   b) Si es simétrica se toma $H = W$.

2. Se calculan los valores y vectores propios de H.

3. Luego se analizan los siguientes casos:

   a) No hay matriz de covariables X: Crear una matriz que contenga los vectores propios de la matriz W, y crear un vector de unos $(1)_{n\times 1}$ como la matriz X.

   b) Si existe matriz de covariables X: Se crea una matriz con las covariables y los vectores propios y se añade la columna de unos $(1)_{n\times 1}$ en la matriz X.

   Para el análisis, se trabaja con el caso en que la matriz de covariables X esta conformada por la media de las distancias registradas entre las coordenadas de cada grupo de observaciones y el establecimiento de interés. Por tanto, a partir de esta matriz se genera otra que contenga a la matriz X y los vectores propios de la matriz de adyacencia obtenidos anteriormente y se añade una columna de unos a la matriz X que corresponden al intercepto en el modelo que se ajustará más adelante.

4. Una vez que ya se tiene modificada la matriz X se calcula la matriz de proyección P_x asociada esta a través de la siguiente fórmula:

$$I - X(X^T X)^{-1} X^T$$

5. Se calcula los residuales de la regresión ajustada y el valor estandarizado z del estadístico de Moran con las fórmulas de (1.2).

6. Se halla el valor del parámetro θ para el método Lasso y se ajusta la regresión.

7. Finalmente se seleccionan los vectores propios que se añadirán como variables independientes al modelo.

   Selección de vectores propios

   Se analizan los siguientes casos:

   a) Si el número de observaciones es igual al de las variables predictoras y no existe matriz X: Se ajusta una regresión lineal simple a la constante con el método de mínimos cuadrados ordinarios.

b) Si el número de observaciones es igual al de las variables predictoras y existe matriz X: Se elimina la columna correspondiente al intercepto en la matriz X y se ajusta una regresión lineal simple a la matriz X con el método de mínimos cuadrados ordinarios.

c) Si el número de observaciones es distinto al de las variables predictoras: Se seleccionan los valores β de la regresión ajustada inicialmente que son distintos de cero y se analiza si se tiene o no la matriz de covariables X para ajustar la regresión lineal a esta matriz X y los vectores propios seleccionados.

Observación: Tener en cuenta que el algoritmo nos devuelve el primer y último modelo ajustado, cuántos vectores no fueron seleccionados, los vectores seleccionados, el estadístico de Moran y el estadístico estandarizado z.

Capítulo 2

Metodología

2.1. Descripción de los datos

La ciudad de Valencia tiene aproximadamente 788 842 habitantes, es un municipio y ciudad de España, capital de la provincia homónima y de la comunidad Valenciana. Se sitúa a orillas del río Turia, en la costa levantina de la península Ibérica, en el centro del golfo de Valencia con una extensión de 134.65 km^2. Por su densidad demográfica es la tercera ciudad más poblada de España, su economía se centra principalmente en actividades de servicios aunque también existen actividades con base industrial y agrícolas.

El Ministerio del Interior Español indica que en 2011 la capital valenciana tuvo una tasa de 69.8 infracciones penales por cada 1000 habitantes. Esta situación fue de gran preocupación para sus habitantes, quienes demandaban medidas inmediatas para reducir estas estadísticas. Las medidas implantadas por las autoridades empiezan a dar resultados para 2013, donde los registros de delincuencia tenían un comportamiento decreciente registrando alrededor de 2000 infracciones penales menos. En 2017, la tendencia nuevamente fue creciente hasta los inicios de la pandemia, donde las restricciones de movilidad cambiaron por completo la dinámica de estos delitos [5].

Para el presente estudio, se dispone de una base de datos que contiene los registros de tres tipos de crímenes de la ciudad de Valencia desde el 2010 hasta el 2020 (Ver figura 2.1), proporcionados por una fuente confiable. Sin embargo, como el año 2020 producto de la pandemia del COVID-19 fue atípico no se incluirá en el análisis. Por tanto, entre 2010 y 2019 se registraron 81406 observaciones, las cuales se clasifican en agresión (55610 casos), sustracción (25342 casos), alarmas mujer (454 casos).

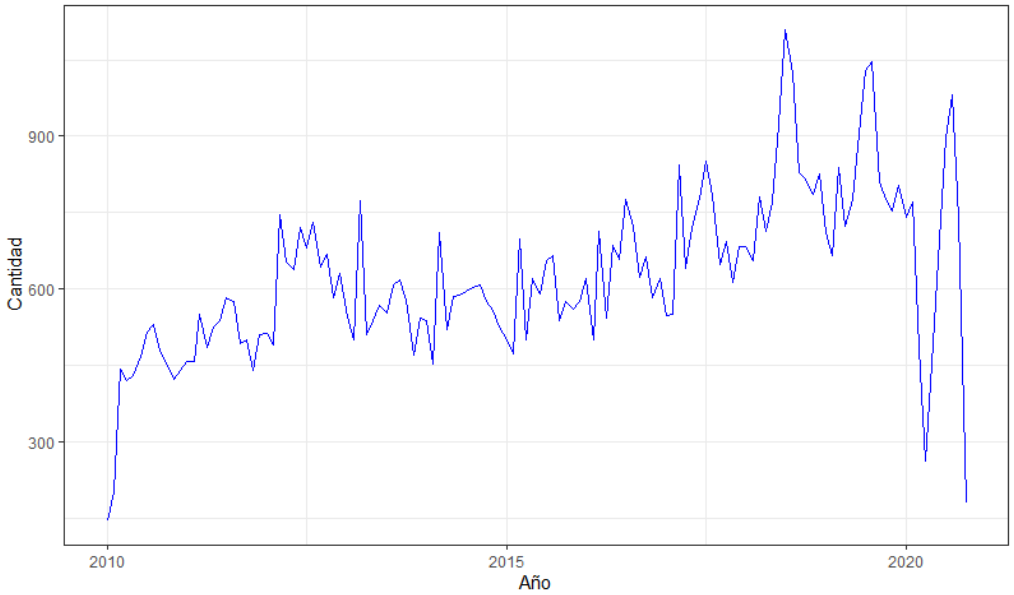


Figura 2.1: Delitos registrados entre 2010-2020

Los registros de las infracciones penales contienen las variables fecha, tipo de crimen, hora y las coordenadas de estos en longitud y latitud. Además, tiene covariables que indican la distancia más corta entre el lugar del crimen y algunos establecimientos. Entre estos tenemos: fábricas; estaciones de policía y paradas de taxi; centros de financiamiento como: bancos y cajeros automáticos; centros de abastecimiento como: cafés, mercados y restaurantes; centros de entretenimiento, que incluyen: bares, discotecas y pubs. Cabe aclarar que los pubs describen establecimientos nocturnos en donde se sirven bebidas alcohólicas y se escucha música y se diferencian de los bares porque ofrecen mayor variedad de cervezas y no hay pista de baile. Para visualizar los datos a analizar se grafica los datos de 2019 que tuvieron el mayor número de infracciones penales (Ver Figura 2.2).

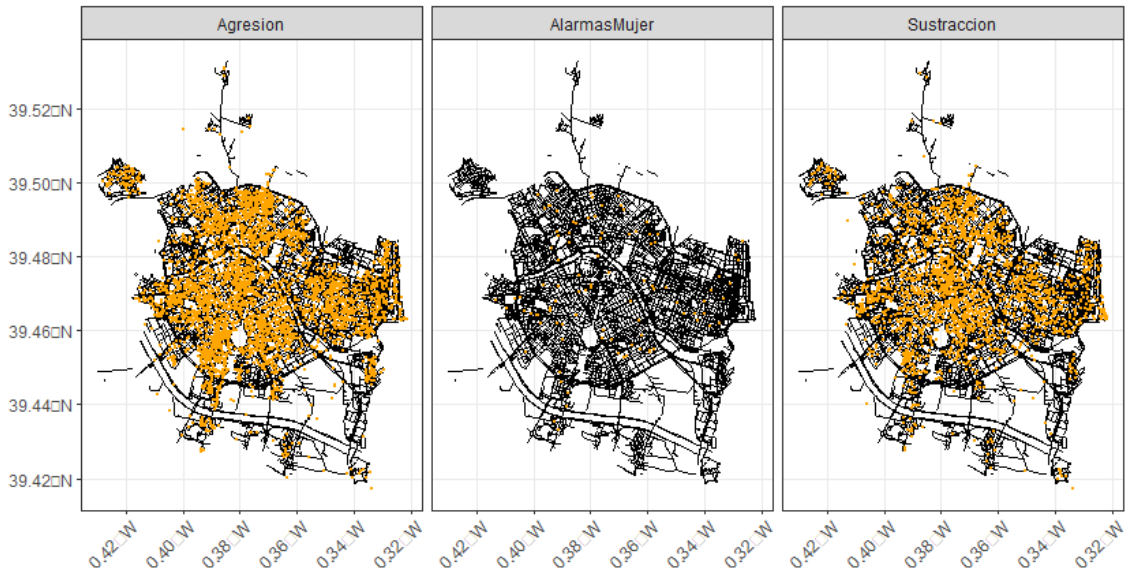


Figura 2.2: Delitos registrados en 2019

La agresión se define como los robos que se cometen luego de golpear a una persona y es el delito que posee mayor cantidad de registros durante el período 2010-2019, además presenta un incremento cercano al $20\,\%$ respecto a períodos anteriores entre el 2012-2018 de acuerdo a la base.

La sustracción se define como hurtos en los que no se usó la fuerza, los cuales también tuvieron un crecimiento en el período 2010-2019 registrando alrededor de 3000 denuncias por año a partir del 2016.

Alarmas mujer se definen como el robo a una mujer haciendo uso de la fuerza. A lo largo de los 10 años son aquellos que no han presentado mayores variaciones y poseen un registro menor en comparación a los delitos ya mencionados manteniendo su registro en alrededor de 100 denuncias por año a partir del 2017.

Considerando que la cantidad de datos influye en la precisión del análisis, se omitirán los casos de Alarmas mujer. Además, como las categorías de sustracción y agresión se refieren a robos no se realizará distinción alguna entre ellas y se las agrupará en adelante en una sola categoría notada robos.

2.2. Análisis Temporal

Para la estimación de la componente temporal, se utiliza modelos ARIMA y se analiza sus características para seleccionar el modelo que proporcione la mejor estimación.

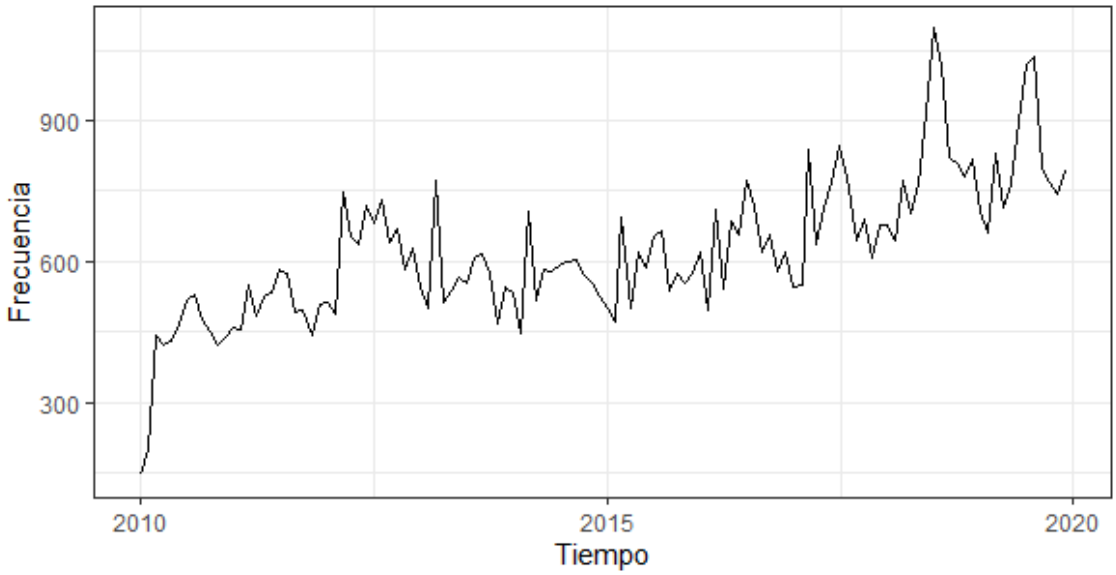


Figura 2.3: Serie temporal delitos en Valencia

En la gráfica de la serie de tiempo se observa que tanto la media como la varianza de los datos no son constantes y si se descompone la serie, vea la figura 2.4 existe tendencia y estacionalidad que son indicios de que la serie no es estacionaria necesitando ser diferenciada.

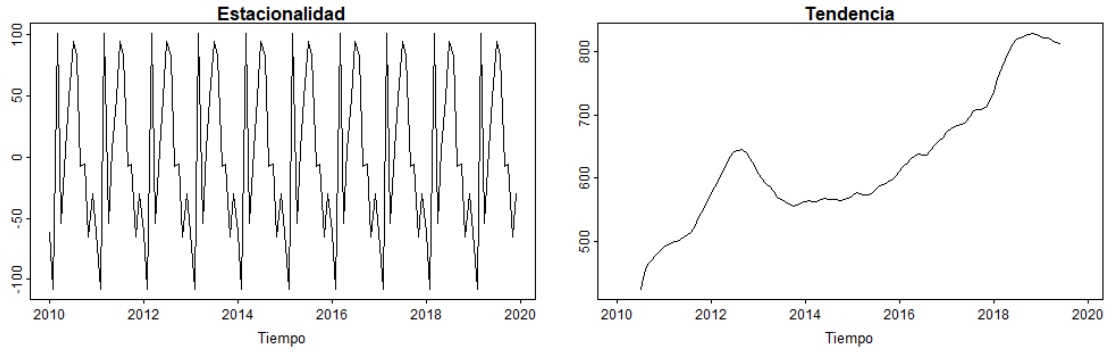


Figura 2.4: Descomposición de la serie temporal

Por ello se diferencia la serie estacionariamente y se aplica nuevamente el test de Dickey Fuller, para indagar sobre la estacionariedad.

Estadístico de prueba	Valor crítico del estadístico	p valor
-5.3745	-1.95	$< 0{,}01$

Cuadro 2.1: Test de Dickey Fuller Aumentado a la serie diferenciada

Aquí el $valor\ p$ de la prueba es menor al $5\,\%$, por lo que se rechaza la hipótesis nula, concluyendo que la serie es estacionaria.

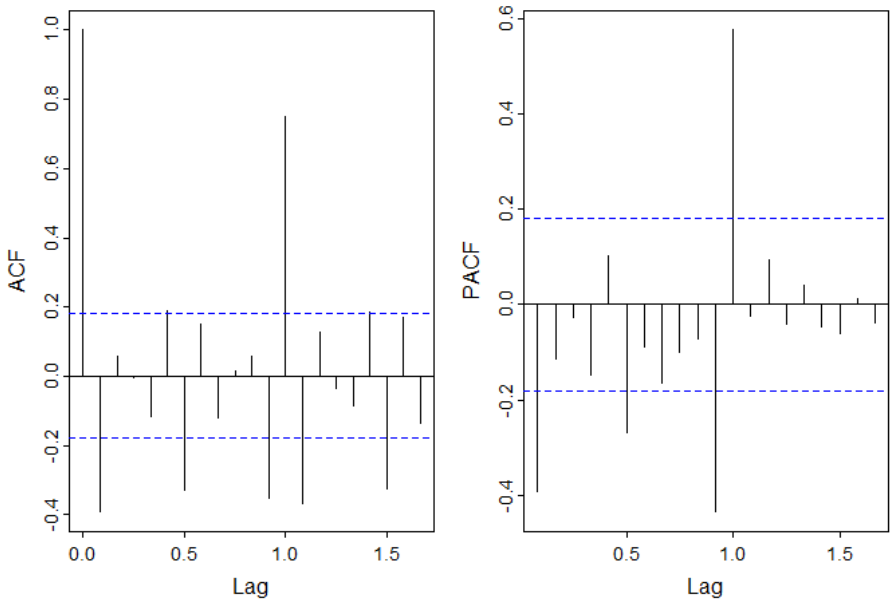


Figura 2.5: Autocorrelogramas residuales

Tomando en cuenta la figura 2.5, se propone entonces un modelo $SARIMA(0,1,1)(1,1,0)_{12}$.

Los coeficientes de los parámetros del modelo son los siguientes:

	Estimado	**Desviación Estándar**	**Estadístico t**	**p valor**
MA 1	-0.3489	0.0977	-3.5706	5e-04
SAR 1	-0.3850	0.0931	-4.1377	1e-04

Cuadro 2.2: Coeficientes del modelo SARIMA

Se observa que el $valor\ p$ de los coeficientes es menor al $5\,\%$ por lo que los parámetros son significativos.

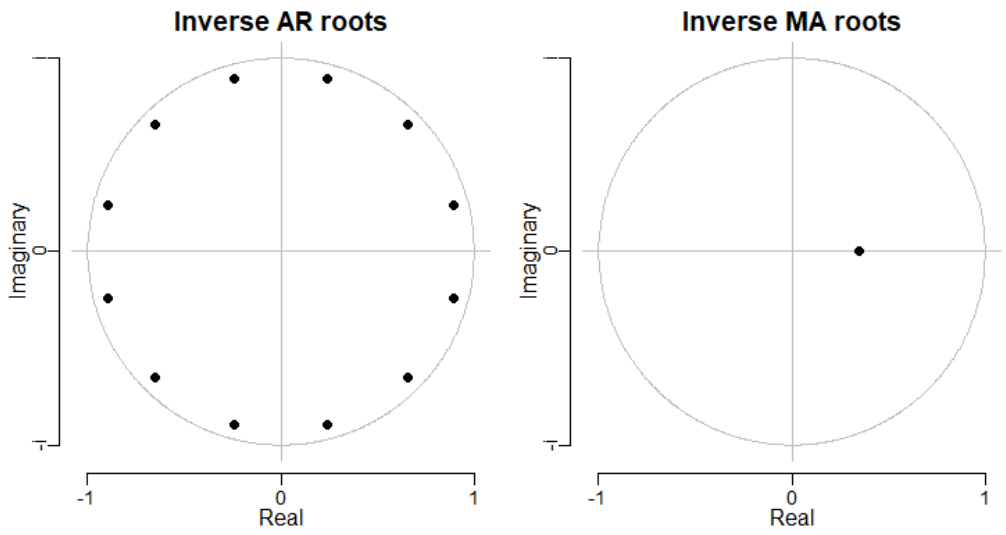


Figura 2.6: Raíces invertidas del modelo SARIMA

Al analizar las raíces invertidas del polinomio para la serie ajustada, y se obtiene que todas se encuentran dentro del círculo de raíz unitaria, por lo que, el proceso ajustado no tiene raíces unitarias y es estacionario.

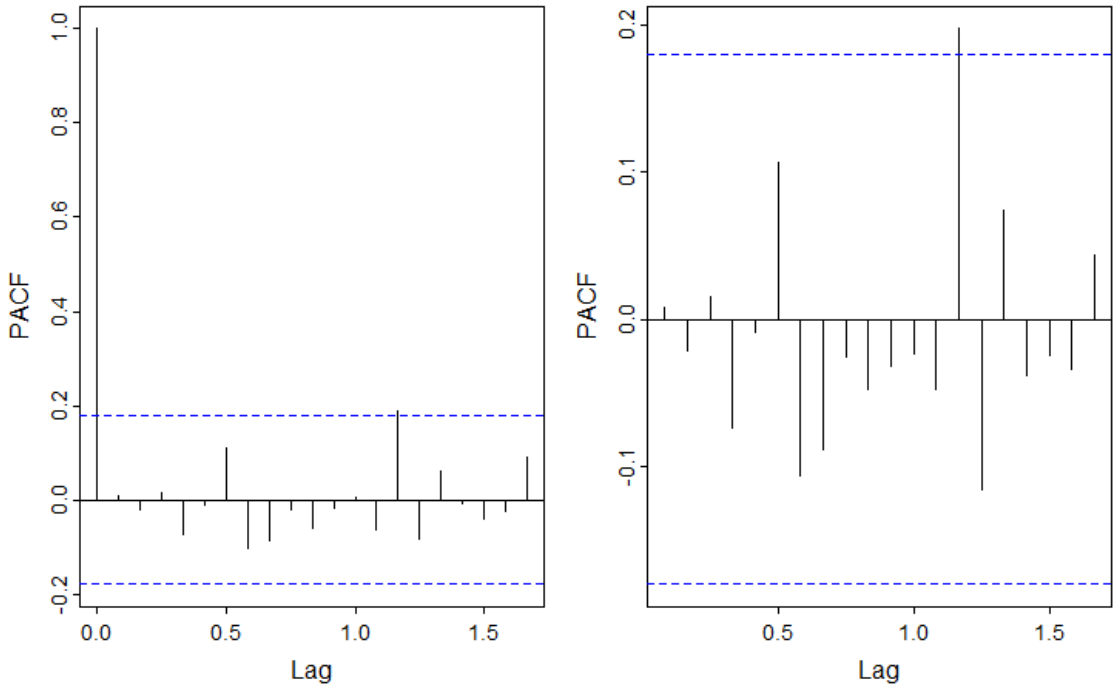


Figura 2.7: Autocorrelogramas residuales modelo SARIMA

Las correlaciones del autocorrelograma y autocorrelograma parcial de los residuos se encuentran dentro de las bandas de confianza, por lo que los residuos no tienen correlaciones significativas y se comportan como un ruido blanco.

Se realiza también la prueba de Ljung-Box que tiene por hipótesis.

Ho : Los datos se distribuyen de forma independiente

Ha : Los datos no se distribuyen de forma independiente

Estadístico de prueba	**Grados de libertad**	**p valor**
0.0079222	1	0.9291

Cuadro 2.3: Test de Ljung-Box del modelo

Como el $valor\ p$ es mayor al nivel de significancia $5\,\%$, las autocorrelaciones son cero o no significativas.

2.3. Análisis Espacial

2.3.1. Tratamiento de los datos

Para poder aplicar el algoritmo del método Mi Lasso a la base de datos es necesario calcular el conteo de robos por celda, la matriz de adyacencia de los datos y la matriz de covariables.

Para ello, se inicia construyendo una grilla e intersecándola con el mapa de tal manera que el conteo en cada celda sea representativo y evitando que se llene de ceros. El criterio para la elección del número de celdas en la grilla, considera que el procentaje de ceros sea menor al $30\,\%$. Bajo este criterio el número de celdas elegido es de 17×18 celdas que tienen un $30{,}12\,\%$ de valores nulos.

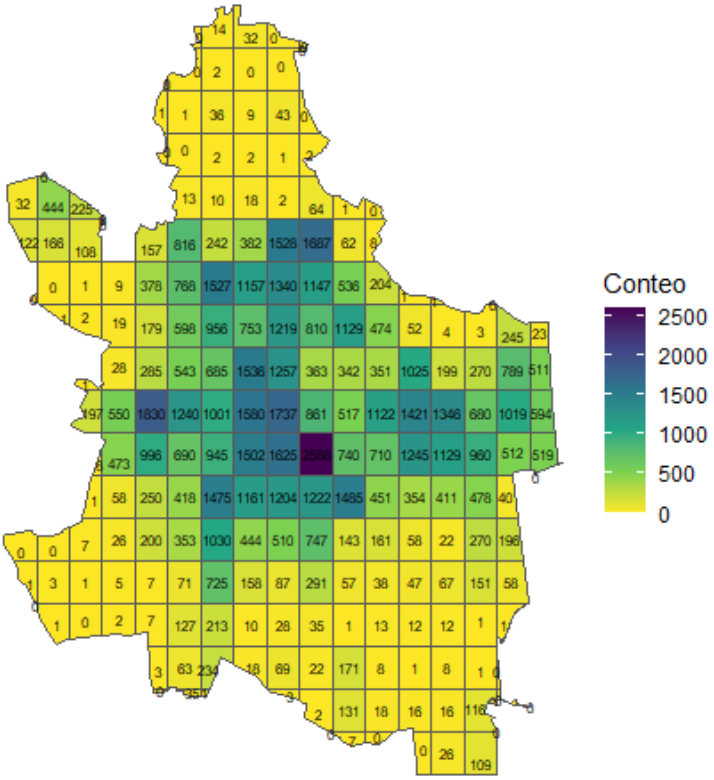


Figura 2.8: Elección de la grilla

En vista de que con esa división se visualiza que la mayor cantidad de robos se concentra en la zona centro de Valencia se plantea tomar únicamente las celdas que tienen un conteo mayor a dos con la finalidad de mejorar la estimación del modelo Mi Lasso. Por tanto, se escoge un grilla de 40×40 y se obtiene la gráfica 2.9.

Se construye el grafo a partir de los vecinos contigüos en la grilla construida, asumiendo que los vecinos de una celda específica son los vértices de las otras celdas que tienen al menos un límite en común (Ver figura 2.10), denominados vecinos del tipo Reina [6]. La matriz de adyacencia se obtiene a partir de este grafo identificando los vecinos existentes en cada celda de la grilla sin valores nulos.

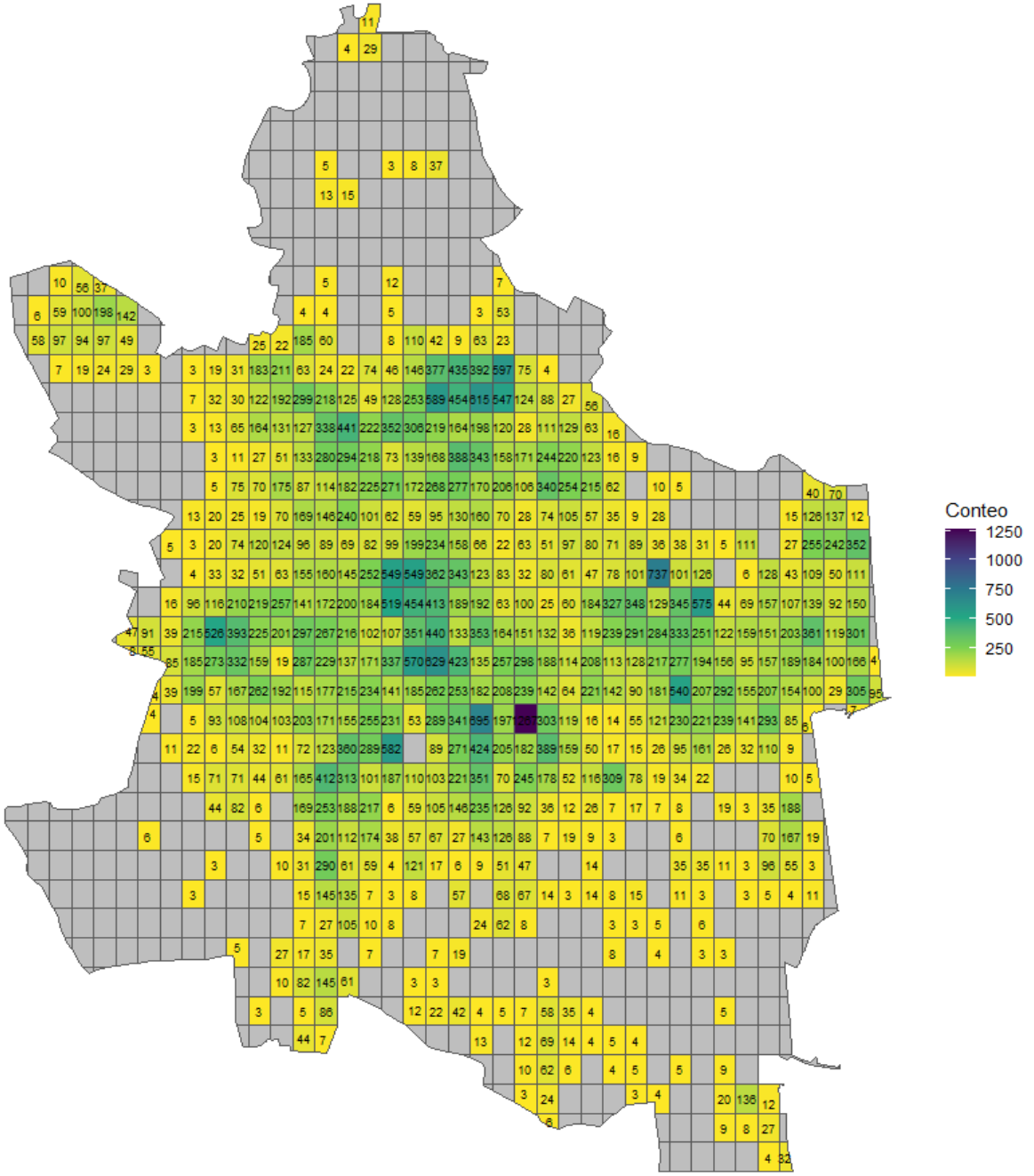


Figura 2.9: Grilla sin ceros

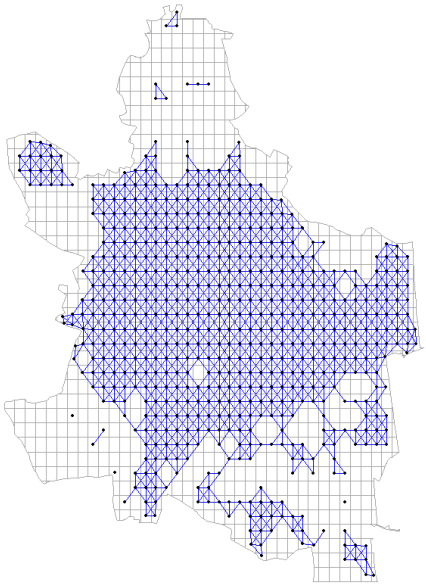


Figura 2.10: Grafo del mapa basado en contigüidad

La figura muestra zonas disjuntas en las que incluso no existen vecinos, hecho que representa que no todas las zonas de la ciudad de Valencia presentan interacciones o comportamientos delictivos similares o con influencia de las zonas cercanas. De hecho, se puede decir que los datos de la zona centro de Valencia es aquella que alberga la mayor interacción espacial y en la que los delitos se pueden asociar o relacionarse entre zonas contiguas.

Para construir la matriz de covariables se requiere establecer una medida representativa para cada zona establecida pudiendo ser por ejemplo la media o median e incluso si se quiere ser más exactos se definiría un centro en cada celda y a partir de este se obtendría las distancias a los establecimientos de interés. En el presente análisis, dado que la base de datos tiene las distancias entre cada punto de robo y los establecimientos especificados se decide tomar como medida representativa la media de las distancias registradas por cada celda.

2.3.2. Estimación del Modelo Mi Lasso

Las entradas para el modelo econométrico Mi Lasso son: y el conteo de robos por cada celda, X la matriz de covariables y E la matriz de adyacencia. Para realizar el análisis y ver la utilidad del método se inicia ajustando un modelo de regresión lineal simple con la matriz de covariables obteniendo:

Variable	Parámetro	Error estándar	t valor	p valor
Intercepto	226.608311	10.395431	21.799	<2e-16 ***
atm_dist	-0.027865	0.009130	-3.052	0.00237 **
bank_dist	-0.070760	0.014898	-4.750	2.55e-06 ***
bar_dist	-0.012392	0.018226	-0.680	0.49682
cafe_dist	0.014705	0.020011	0.735	0.46274
industrial_dist	0.006986	0.010024	0.697	0.48614
market_dist	-0.028229	0.009861	-2.863	0.00435 **
nightclub_dist	-0.026020	0.012209	-2.131	0.03347 *
police_dist	-0.011927	0.010947	-1.090	0.27634
pub_dist	0.015221	0.014232	1.070	0.28525
restaurant_dist	0.018039	0.015685	1.150	0.25059
taxi_dist	0.032211	0.013965	2.306	0.02142 *

Cuadro 2.4: Coeficientes del modelo de regresión lineal simple

La tabla 2.4 indica que las variables que son significativas son la distancia a: cajeros automáticos *atm_dist*, bancos *bank_dist*, supermercados *market_dist*, discotecas *nightclub_dist* y estaciones de taxis *taxi_dist*; pues tiene un *valor p* menor al nivel de significancia del $5\,\%$.

Error estándar residual: 118.8 con 599 grados de libertad	
R^2: 0.2533	R^2 Ajustado: 0.2396
Estadístico F: 18.47 con 11 y 559 grados de libertad, p valor: <2.2e-16	

Cuadro 2.5: Resumen del modelo de regresión lineal simple

El cuadro 2.5, presenta un error estándar residual de $118{,}80$ indicando que, en promedio, las observaciones en el modelo difieren de sus valores predichos por $118{,}80$ unidades aproximadamente.

El R^2 de 0.2533 muestra que el modelo explica el $25{,}33\,\%$ de variabilidad de los datos del conteo (variable dependiente), lo cual es insuficiente. Además, el R^2 ajustado tiene un valor de $-0{,}2396$ medida semejante al R^2, con la diferencia de que tiene en cuenta el número de variables explicativas que forman parte del modelo penalizando, el aumento que hay en el R^2 debido a la inclusión de variables que no mejoran de manera significativa la capacidad predictiva del modelo. El valor bajo de este indicador resulta muy interesante pues nos indica que el modelo no se ajusta bien a los datos o que a su vez las variables incluidas no aportan significativamente para explicar el conteo de delitos. También puede ser un indicio de sobre ajuste, problemas de multicolinealidad, bondad de ajuste, entre otros, volviéndose un modelo no confiable para realizar predicciones.

En vista de que el modelo debe ser corregido, se plantea la hipótesis de que el problema de un R^2 ajustado bajo se debe a la exclusión de la correlación espacial existente entre las variables. Para probar esta hipótesis se calcula el test de Moran:

Ho : No hay correlación espacial en los residuos

Ha : Existe correlación espacial en los residuos

Obteniendo:

Estadístico de prueba	**p valor**
0.3151	0.001

Cuadro 2.6: Test de Moran del modelo de regresión lineal simple

Dado que el $valor\ p$ es menor al nivel de significancia de la prueba $(5\,\%)$, se concluye que no existe suficiente evidencia para aceptar la hipótesis nula y por tanto existe correlación espacial en los residuos haciéndose necesaria la aplicación del método Mi Lasso.

Aplicando el método Mi Lasso descrito con anterioridad se obtienen los resultados de la tabla 2.7. Se visualiza las 15 primeras variables, de una dimensión total de 566 variables.

Variable	Parámetro	Error estándar	t valor	p valor
(Intercept)	1625	4.534	35.829	<2e-16 ***
Xatm_dist	-1.313e-02	2.223e-03	-5.907	1.60e-06 ***
Xbank_dist	-6.115e-02	1.675e-03	-36.517	<2e-16 ***
Xbar_dist	-3.192e-02	3.471e-03	-9.196	2.28e-10 ***
Xcafe_dist	2.206e-02	3.755e-03	5.874	1.76e-06 ***
Xindustrial_dist	-1.703e-02	1.286e-03	-13.248	2.63e-14 ***
Xmarket_dist	8.806e-03	2.102e-03	4.190	0.000215 ***
Xnightclub_dist	-1.144e-02	2.136e-03	-5.357	7.72e-06 ***
Xpolice_dist	-2.790e-02	2.079e-03	-13.425	1.85e-14 ***
Xpub_dist	4.587e-02	2.590e-03	17.709	<2e-16 ***
Xrestaurant_dist	-2.472e-02	5.724e-03	-4.319	0.000150 ***
selV1	-121.8	47.60	-25.597	<2e-16 ***
selV3	133.5	24.46	5.460	5.75e-06 ***
selV4	71.97	19.21	3.745	0.000737 ***
selV5	-9.307	10.63	-0.876	0.387850

Cuadro 2.7: Coeficientes del modelo Mi LASSO

El *valor p* de las variables, son en su mayoría significativos al ser menores al nivel de significancia $5\,\%$. A diferencia de la regresión lineal simple, se incrementa la cantidad de variables predictoras significativas del modelo.

Error estándar residual: 3.886 con 31 grados de libertad	
R^2: 1	R^2 Ajustado: 0.9992
Estadístico F: 1294 con 579 y 31 grados de libertad, p valor: <2.2e-16	

Cuadro 2.8: Resumen del modelo Mi Lasso

Se observa en el cuadro 2.8 que el error estándar residual se reduce significativamente, por lo que los errores son menores a los del primer modelo. El R^2 alcanzado se visualiza como 1, sin embargo, el que un modelo represente el $100\,\%$ de variabilidad de los datos no es esperado o probable. Se reviso el nivel de precisión de la tabla de resumen que nos da el *RStudio* el cual se encontraba en 4 decimales. En vista de que se supera este nivel, el programa nos da el valor de 1. Verificado esto

se accede con otra función al valor del R^2 real obteniendo en el modelo obteniendo $0{,}9999586$ por lo que el modelo explica alrededor del $99{,}996\,\%$ de variabilidad de los datos. Si se penaliza la cantidad de variables incluidas en el modelo, se tiene que el R^2 ajustado es $0{,}999186$. Ambos valores nos indican la adecuación y apropiado ajuste del modelo.

2.3.3. Valor del exponente del parámetro de ajuste

Se analiza el valor del exponente α que se va a elegir y como este influye en el valor del R^2 ajustado obteniendo la siguiente información:

Alfa	R^2 **ajustado**	**Coeficientes significativos**
0.10	0.9936	493
0.22	0.9965	519
0.34	0.9984	548
0.46	0.9990	562
0.58	0.9995	571
0.70	0.9998888	574
0.82	0.9998895	571
0.93	0.9999982	541

Cuadro 2.9: Valores de α en el modelo Mi Lasso

Recuerde que el valor α es el exponente del denominador del parámetro θ del método LASSO (Ver ecuación (1.2)). Dado que la correlación se representa por z y para el modelo analizado es mayor a 1, $(16{,}10417)$. Se tiene entonces una relación inversa entre el parámetro α y θ en donde conforme aumenta el valor de α disminuye el valor de θ. Así por ejemplo, un valor de $\alpha = 1{,}1$ resulta en $\theta = 0{,}04702922$ valor cercano a cero que como se menciono antes es semejante a ajustar un modelo de regresión simple e incluso complica la selección de variables por ser casi nulo. En cambio, un valor $\alpha = 0{,}5$ que es el ajustado en el modelo 2.7, le corresponde un valor $\theta = 0{,}2491901$, seleccionando 566 variables como significativas. Por otro lado, si $\alpha = 0{,}1$, entonces $\theta = 0{,}7573666$ seleccionando 491 variables como significativas.

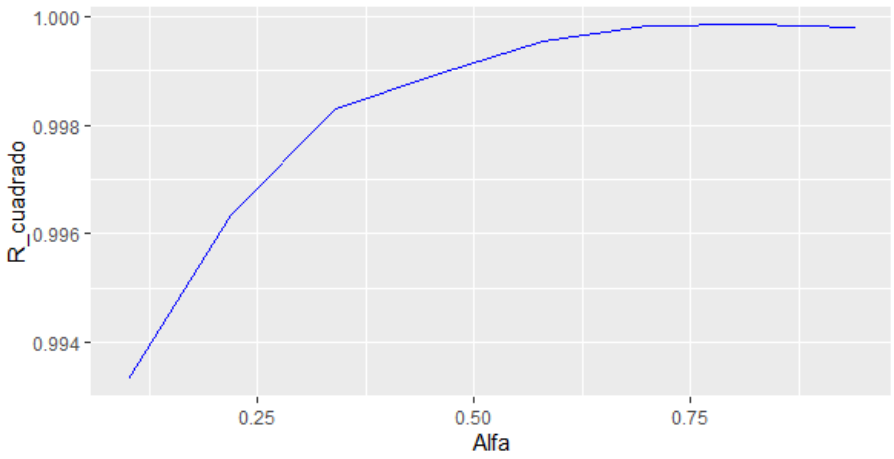


Figura 2.11: Valor del α vs R^2 ajustado

En la figura 2.11, se observa entonces que a medida que aumenta el valor α, el R^2 ajustado también aumenta. Si se restringe el valor del α a la cantidad de parámetros significativos 2.9, buscando minimizar el número de coeficientes sin afectar el valor del R^2, se obtendría que el modelo óptimo es aquel que tiene un valor $\alpha = 0{,}94$.

Capítulo 3

Resultados, conclusiones y recomendaciones

3.1. Resultados

3.1.1. Evaluación de la serie temporal

El modelo temporal elegido fue un $SARIMA(0,0,1)(1,1,0)_{12}$, para evaluar la calidad del ajuste del modelo, se realiza el análisis residual. Para ello, se aplica la prueba de Jarque Bera cuyas hipótesis son:

Ho : Los residuos se distribuyen normalmente

Ha : Los residuos no se distribuyen normalmente

Estadístico de prueba	**Grados de libertad**	**p valor**
3.4984	2	0.1739

Cuadro 3.1: Test de Jarque Bera del modelo

El $valor\ p$ de la prueba de Jarque Bera sobre los residuos es de $0{,}1739$ por lo que no se rechaza la hipótesis nula y se concluye que los residuos se distribuyen normalmente. Lo cual se sustenta en la figura 3.1.

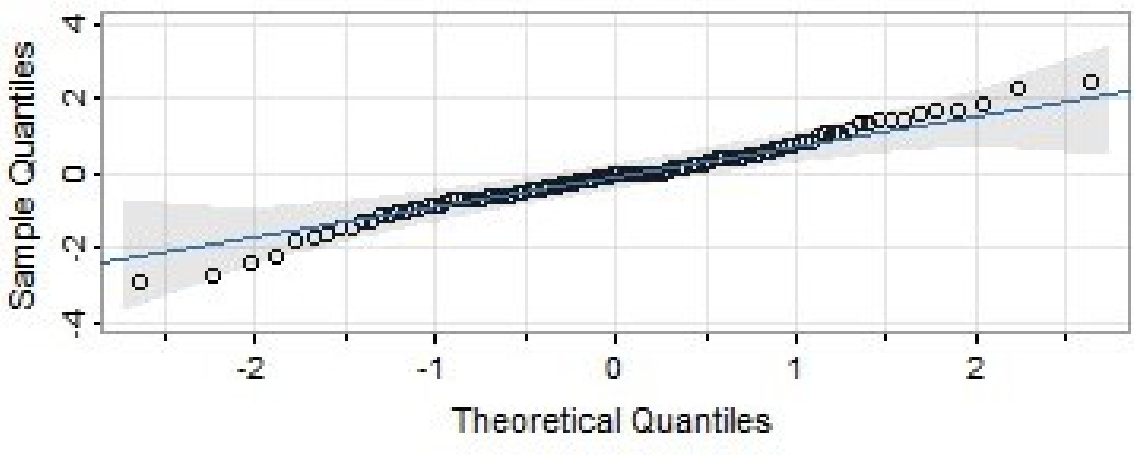


Figura 3.1: Gráfico Q-Q plot de la serie temporal

Además, se verifica la heterocedasticidad mediante la prueba del multiplicador de Lagrange (LM) para la heterocedasticidad condicional autorregresiva (ARCH) cuyas hipótesis son:

Ho : No existe presencia de efectos ARCH

Ha : Existe presencia de efectos ARCH

Estadístico de prueba	**p valor**
1.3369	0.2476

Cuadro 3.2: Test de heterocedasticidad del modelo

Dado que el $valor\ p$ de la prueba es mayor al nivel de significancia $5\,\%$ no se rechaza la hipótesis nula y se concluye que no hay presencia de efectos de ARCH. Concluyendo entonces que la serie de robos de Valencia tiene la ecuación:

$$(1 - B^{12})(1 - B)x_t = (1 - 0{,}3850B^{12})(1 - 0{,}3489B)w_t$$

Y su pronóstico se muestra a continuación:

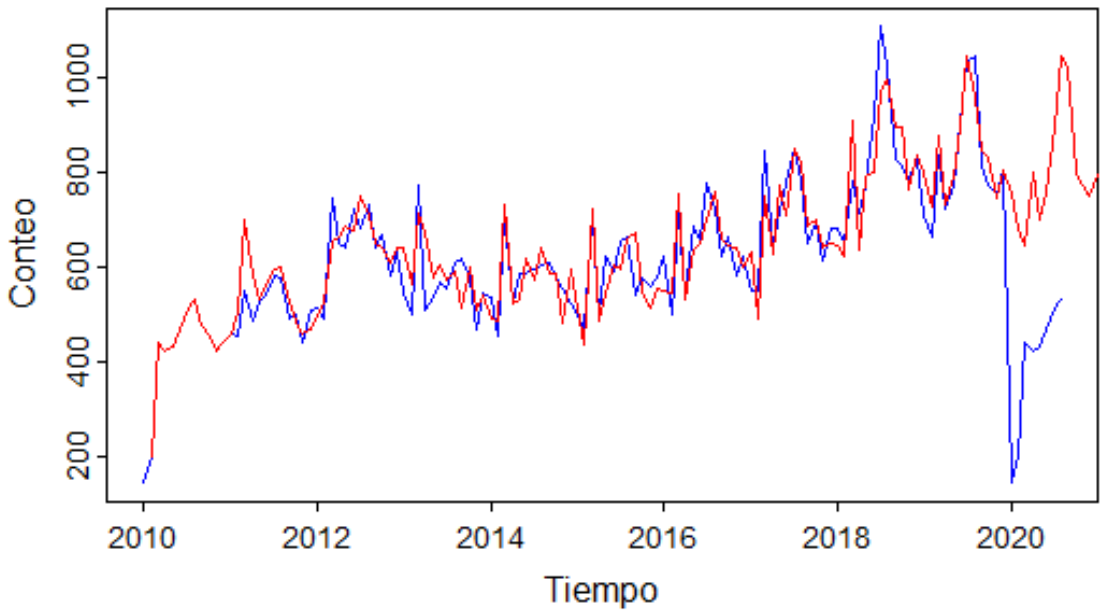


Figura 3.2: Predicción del modelo SARIMA

En el gráfico 3.2 se observa los valores ajustados en rojo y los valores observados en azul. El modelo temporal tiene un buen ajuste pues en promedio los residuos del modelo son de 5 delitos por cada mes. Además, se realiza la predicción para el año 2020 y se la compara con los datos observados. La comparación nos permite identificar una diferencia entre los observados y predichos mayor a 200 robos, diferencia muy grande que se debe a la pandemia del COVID 19, la cual obligó a los países a implementar restricciones de movilidad que disminuyeron alrededor de un $50\,\%$ los robos durante los primeros meses.

3.1.2. Evaluación del modelo espacial

En cuanto a la parte espacial, el modelo que mejor se ajusta a los datos es el modelo econométrico espacial de parámetro $\alpha = 0{,}90$ con las variables de la tabla 3.3, puesto que maximiza el R^2 y a su vez minimiza la cantidad de coeficientes del modelo.

Variable	Parámetro	Error estándar	t valor	p valor
(Intercept)	155.00906	1.97369	78.53772	4.45212e-14
Xatm_dist	-0.00445	0.00139	-3.20912	0.01067
Xbank_dist	-0.03748	0.00263	-14.25489	1.75459e-7
Xindustrial_dist	-0.01342	0.00115	-11.67295	9.73993e-7
Xnightclub_dist	-0.00932	0.00129	-7.21199	5.01777e-5
Xpolice_dist	-0.01474	7.43793e-4	-19.81917	9.83662e-9
selV1	-1283.14636	25.70142	-49.92512	2.60372e-12
selV3	83.04372	13.22221	6.28062	0001.44270e-4
selV4	57.20081	6.57956	8.69371	1.13235e-5
selV5	-144.94896	10.44142	-13.88212	2.20623e-7

Cuadro 3.3: Coeficientes del modelo econométrico espacial

Se verifica que todos los coeficientes son significativos pues el $valor\ p$ es menor al nivel de significancia 5% y se evalúa las características del modelo. Primero se analiza la estructura espacial del modelo:

Estadístico de prueba	p valor
-0.22019	0.999

Cuadro 3.4: Test de Moran del modelo econométrico espacial

Como el $valor\ p$ es mayor al nivel de significancia 5% no existe suficiente evidencia para rechazar la hipótesis nula, concluyendo que los residuos ya no tienen autocorrelación espacial.

Se realiza también la prueba de bondad de ajuste de Hosmer y Lemeshow que tiene por hipótesis:

Ho : El modelo se ajusta bien a los datos

Ha : El modelo no se ajusta bien a los datos

Estadístico de prueba	Grados de libertad	p valor
-0.00000058129	8	1

Cuadro 3.5: Test de Hosmer Lemeshow

Donde dado que la precisión de los decimales en el $RStudio$ solo admite hasta 16 cifras se tiene que el valor redondeado es de 1 y como el $valor\ p$ es mayor al nivel de significancia de la prueba $5\,\%$ se acepta la hipótesis nula y se concluye que el modelo se ajusta bien a los datos.

La prueba de White cuyas hipótesis son:

Ho : Homocedasticidad de los residuos

Ha : Heterocedasticidad de los residuos

Estadístico de prueba	**p valor**
1.72	0.423258

Cuadro 3.6: Test de White

El $valor\ p$ es mayor al nivel de significancia $5\,\%$ se acepta la hipótesis nula y se concluye que el modelo tiene residuos homocedásticos, es decir su varianza es constante y se distribuyen de manera homogénea.

Prueba de Lilliefors cuyas hipótesis son:

Ho : Los residuo provienen de una distribución normal

Ha : Los residuo no provienen de una distribución normal

Estadístico de prueba	**p valor**
0.03609	0.05619

Cuadro 3.7: Test de Lilliefors

Dado que el $valor\ p$ es mayor al nivel de significancia $5\,\%$ se concluye que los residuos siguen una distribución normal y se lo sustenta con la gráfica $Q-Q$ plot.

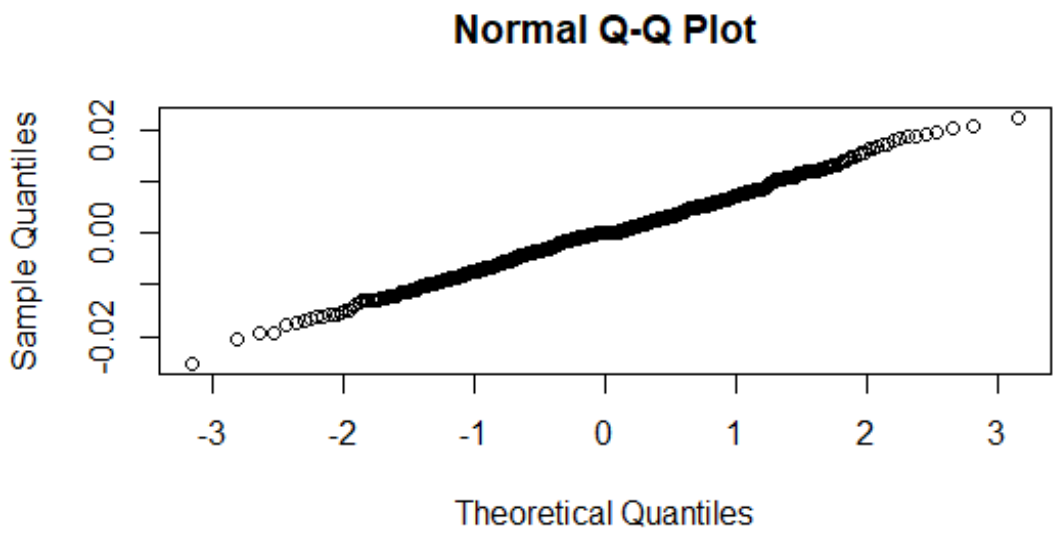


Figura 3.3: Gráfico Q-Q plot del modelo espacial

Error estándar residual: 1.157 con 9 grados de libertad	
R^2: 1	R^2 Ajustado: 0.9999
Estadístico F: 1.406e+04 con 601 y 9 grados de libertad, p valor: <2.2e-16	

Cuadro 3.8: Resumen del modelo econométrico espacial

Se observa en el cuadro 3.8 que el R^2 ajustado es cercano a 1 por lo que el ajuste es bueno y si comparamos los valores ajustados con los observados se tiene que las diferencias son menores a $0{,}50$. Concluyendo que es el mejor modelo econométrico espacial ajustado a los datos.

3.2. Conclusiones

1. El ajuste de series temporales para estudiar cómo se modifican eventos como robos, suele ser de gran utilidad, puesto que permite analizar si los eventos pasados influyen en los futuros. Sin embargo, este tipo de modelos presentan complicaciones cuando se presentan cambios drásticos, que pueden cambiar por completo el comportamiento de la serie.

2. La inclusión espacial en modelos econométricos a través de la matriz

de adyacencia y la selección de vectores propios significativos, con el filtrado de vectores espaciales, mejora significativamente la estimación de modelos que tienen autocorrelación espacial. Estos modelos tienen un buen poder predictivo, pues capturan aproximadamente el $100\,\%$ de variabilidad de los datos. Permiten explicar, con mayor precisión, el comportamiento de variables que tienen influencia del espacio y la interacción entre vecinos.

3. Dada la importancia de la selección de vectores propios, se analizó las implicaciones que tiene variar el parámetro α del método Mi Lasso en la cantidad de vectores seleccionados. Se concluyó que el parámetro se encuentra en el intervalo $(0, 1)$ y este depende del número de celdas de la grilla, así como de la cantidad de datos. La relación entre el parámetro α y la cantidad de vectores propios seleccionados es directamente proporcional. Al cambiar esta relación se determina el valor óptimo de α que maximiza el R^2 y que minimiza la cantidad de vectores significativos seleccionados.

4. El método Mi LASSO mejora las estimaciones realizadas por el método LASSO al proporcionarnos valores eficientes para su parámetro de ajuste θ. Al limitar los valores de este parámetro se proporcionan las mejores estimaciones y se reduce el tiempo que conlleva la selección del modelo econométrico espacial adecuado.

3.3. Recomendaciones

1. Al trabajar con datos espaciales se debe tener en cuenta que existe muchas formas de hacerlo, empezando por la forma en que se construye la división del territorio. Analizar la partición que puede ser irregular, política o regular.El tamaño de la grilla se considera con respecto al área o número de habitantes, todo en función del objetivo de la investigación y la información que se vaya a analizar.

2. La forma en que se determinan los vecinos contiguos, al trabajar con datos espaciales, también es variada. Se puede ser más estricto y exigir que no se tenga uno, sino dos vértices en común, para que

sean vecinos. Incluso se pueden definir vecinos de primer y segundo grado. Estas opciones se podrían analizar en el futuro.

3. En la matriz de adyacencia, se puede utilizar una matriz de pesos que, como su nombre lo dice, tiene componentes con valores distintos al cero o al uno, que buscan capturar la influencia que tienen las observaciones en el análisis.

Capítulo A

Título anexo

A.1. Pseudocodigo del algoritmo del método Mi LASSO

1. Verificar que la matriz W sea simétrica, caso contrario transformarla y hallar sus vectores propios.
2. Verificar si existe matriz de covariables
   a) Si existe: Unir a la matriz de covariables con los vectores propios en una sola matriz.
   b) Si no existe: Conservar la matriz de vectores propios.

   Observación: En ambos casos se añade el vector de unos correspondiente al intercepto en la matriz.
3. Hallar la matriz espectral a partir de la matriz de covariables que tiene el intercepto.
4. Calcular y estandarizar el estadístico de Moran.
5. Establecer el parámetro θ con la ecuación 1.2.
6. Ajustar el modelo LASSO con el párametro de ajuste determinado por θ.

7. Seleccionar los vectores propios distintos de cero significativos.

1. Verificar que la matriz W sea simétrica, caso contrario transformarla y hallar sus vectores propios.
2. Verificar si existe matriz de covariables
   a) Si existe: Juntar la matriz de covariables con los vectores propios en una sola matriz
   b) Si no existe: Conservar la matriz de vectores propios

   Observación: En ambos casos se añade el vector de unos correspondiente al intercepto en la matriz
3. Hallar la matriz espectral a partir de la matriz de covariables que tiene el intercepto
4. Calcular y estandarizar el estadístico de Moran
5. Establecer el parámetro θ con la ecuación 1.2
6. Ajustar el modelo LASSO con el párametro de ajuste determinado por θ
7. Seleccionar los vectores propios distintos de cero significativos.

A.2. Código en R

```
library(tidyverse)
library(sf)
library(dplyr)
library(spatstat)
#Importacion de datos y mapa

base_de_datos_Valencia <- read.csv("Base/base_de_datos_Valencia.csv")
#base de datos eliminada el anio de la pandemia
datos_Valencia<-base_de_datos_Valencia%>%filter(year<2020 & crime_type!="Otros")
#convierte un caracter a fecha
datos_Valencia$crime_date <- as.Date(datos_Valencia$crime_date,format="%m/%d/%Y")
setwd("C:/Users/Usuario/Documents/Katheryn/Universidad/Octavo Semestre/TIC/shape files")
road.sf<-st_read("valencia_outline.shp")
#transformacion de coordenadas
road.sf<-road.sf%>%st_transform(4269)
df.sf <- st_as_sf(datos_Valencia, coords = c("crime_lon", "crime_lat") )%>%st_set_crs(4269)

#Construccion de la grilla

border <- st_union(road.sf)

grid.border <- border %>%
```

```
  st_make_grid(n = c(40,40),crs=4269) %>%
  st_intersection(border) %>%
  st_cast("POLYGON") %>%
  st_sf()

grid <- grid.border%>%
  mutate(cellid = row_number())

# Interseccion con las coordenadas

datos_Valencia <- datos_Valencia %>% filter(crime_type!="AlarmasMujer")
df.sf <- st_as_sf(datos_robos[,c("crime_type","crime_lon", "crime_lat")], coords = c("crime_lon", "crime_lat") )%>%st_set_crs(4269)
points.sf<-df.sf

analisis1 <- grid %>%
  st_join(points.sf) %>%
  mutate()%>%
  group_by(cellid) %>%
  summarize(conteo=n())

analisis2 <- grid %>%
  st_join(points.sf)

analisis3 <- analisis2 %>% mutate(conteo=(!is.na(analisis2$crime_type))*1)
analisis3 <- analisis3 %>% group_by(cellid) %>% summarise(across(conteo, ~ sum(.x, na.rm = TRUE)))

#Grafica
ggplot(analisis3) +
  geom_sf(data = grid, fill = "gray", size = 1) +
  geom_sf(aes(fill = conteo)) +
  geom_sf_text(aes(label=conteo),size=2,color="black")+
  scale_x_continuous(breaks = c(-84)) +
  labs(fill = "Count")+scale_fill_viridis_c(option = "viridis",begin=0,direction = -1)+theme_bw()

regular.y <- analisis3$conteo1

regular.y

# grilla que toma las celdas cuyo conto es mayor a 2

grid_non_zero <- grid$geometry[which(regular.y>2)]

grid.border1 <- grid_non_zero %>%
  st_make_grid(n = c(1,1),crs=4269) %>%
  st_intersection(grid_non_zero) %>%
  st_cast("POLYGON") %>%
  st_sf()

grid1 <- grid.border1%>%
  mutate(cellid = row_number())

datos_Valencia <- datos_Valencia %>% filter(crime_type!="AlarmasMujer")
df.sf <- st_as_sf(datos_Valencia[,c("crime_type","crime_lon", "crime_lat")], coords = c("crime_lon", "crime_lat") )%>%st_set_crs(4269)
points.sf<-df.sf

analisis1 <- grid1 %>%
  st_join(points.sf) %>%
  mutate()%>%
  group_by(cellid) %>%
  summarize(conteo=n())

analisis2 <- grid1 %>%
  st_join(points.sf)

analisis3 <- analisis2 %>% mutate(conteo1=(!is.na(analisis2$crime_type))*1)
analisis3 <- analisis3 %>% group_by(cellid) %>% summarise(across(conteo1, ~ sum(.x, na.rm = TRUE)))

regular.y <- analisis3$conteo1

regular.y

# Construccion del grafo

library(spdep)
```

```r
library(sp)

regular.graph <- poly2nb(grid.border1, queen=TRUE)

adjancecymatrix.regular <- nb2mat(regular.graph,style="B", zero.policy=TRUE)

#Construccion  de la matriz de distancias de las covariables

datos_Valencia <- datos_Valencia %>% filter(crime_type!="AlarmasMujer")
df.sf <- st_as_sf(datos_Valencia, coords = c("crime_lon", "crime_lat") ) %>%st_set_crs(4269)
cov.sf<-df.sf

X <- analisis3 %>% st_join(cov.sf, left=TRUE)%>% mutate()%>% group_by(cellid) %>%
  summarise(across(names(cov.sf)[13:23], ~ mean(.x, na.rm = TRUE))) %>%
  st_drop_geometry()  %>% select(-1) %>%  as.matrix()

#Funcion Mi LASSO

MiLasso <- function(y, X=NA, W, A=2){

  #arguments
  # y <-Y # dependant variable
  # X <- NA # Matrix of covariats
  # W <- W # SWM
  # A <- a # exponent on tuning paramter

  # install.packages("glmnet") #lasso package
  library(glmnet)

  GetMoranStat <- function(MSM, degfree) {
    #MSM    : M %*% S %*% M matrix
    #         M : projection matrix
    #         S : coded symmetric spatial link matrix
    #degfree: degrees of freedom

    MSM <- as.matrix(MSM)
    t1 <- sum(diag(MSM))
    t2 <- sum(diag(MSM %*% MSM))

    E <- t1 / degfree #equ 8 tg07
    Va <- 2 * (degfree * t2 - t1 * t1)/(degfree * degfree * (degfree + 2)) # equ 9 tg07
    return(list(Mean=E,Var=Va))
  }

  if(isSymmetric(W)==FALSE) print("W forced symmetric")
  V <- (W + t(W)) / 2 # make SWM symeteric
  EigenV <- eigen(V, symmetric = TRUE)
  E <- as.matrix(EigenV$vectors)
  colnames(E) <- seq(1,ncol(E))

  no_X <- is.na(X[1])

  if(no_X ==1){ # no X's
    XE <- E # design matrix
    X <- as.matrix(rep(1,length(y)))
    pen <- rep(1,ncol(E))
  }else{
    XE <- cbind(X,E) # design matrix
    pen <- c(rep(0,ncol(X)),rep(1,ncol(E)))
    X <- cbind(X,1) # add constant to matrix of covariates
  }


  M_X <- diag(1, nrow(X)) - X %*% solve(crossprod(X))%*%t(X) #projection matrix
  MStat <- GetMoranStat(MSM = M_X%*%V%*%M_X, degfree = nrow(X) - ncol(X)) #Moran E and Var
  res <- y - X %*% solve(crossprod(X),crossprod(X,y)) # residuals = y - xhat*b
  mI <- (crossprod(res, V) %*% res) / crossprod(res) #mi = (res'S)res/res'res
  zI <- (mI - MStat[["Mean"]]) / sqrt(MStat[["Var"]])

  theta <- abs(zI)^{-A} # absolute vlaue of the inverse of the morans I
  fit = glmnet(XE, y, lambda = theta, penalty.factor =pen) # Milasso

  ###
  #find selected evecs for Miplasso
```

```
  if((fit$df - ncol(X)+1)==0){ # no eigenvectors selected
    if(no_X==1){
      selected <- rep(0,ncol(E))
      names(selected)<- seq(1:ncol(E))
      Miplasso <- lm(y ~ 1) # simple ols model
    }else{
      X <- X[,-ncol(X)]
      selected <- rep(0,ncol(E))
      names(selected)<- seq(1:ncol(E))
      Miplasso <- lm(y ~ X) # simple ols model
    }
  } else {

    Temp <- as.matrix(cbind(fit[["beta"]],seq(1,nrow(fit[["beta"]])),0))
    Temp[,3] <- (Temp[,1] > 0 | Temp[,1] < 0)  # asign non-zero evecs value 1
    sel_no  <-  Temp[,2]*Temp[,3]
    Ext = subset(sel_no,sel_no!=0)  #extracted eigenvectors
    Exx  = XE[,Ext]
    if(no_X==1){

      selV  = Exx
      selected <- Temp[,3]
      names(selected)<- seq(1:ncol(E))
      Miplasso <- lm(y ~ selV) # post lasso
    }else{
      print("aqui")
      X <- X[,-ncol(X)]
      selV  = Exx[,-(1:ncol(X))]
      selected <- Temp[,3][-(1:ncol(X))]
      names(selected)<- seq(1:ncol(E))
      Miplasso <- lm(y ~ X + selV) # post lasso
    }


  }


  return(list(MipLasso=Miplasso, MiLasso=fit, no_selected=(fit$df - ncol(X)), selected=selected, mI= mI, zI=zI))
}

# Aplicacion del metodo Mi LASSO

y=regular.y #variable dependiente
W=adjancecymatrix.regular #matriz de adyacencia
A =0.5  # exponente del parametro de ajuste en el parametro lasso tetha
X #matriz de covariables
Modelo_Espana_1 <- MiLasso(y=regular.y,X=X, W=W, A=0.5)

# Test de Moran aplicado al modelo de regresion lineal simple
y <- regular.y
modelo_inicial <- lm(y~X)

library(ape)
Moran.I(modelo_inicial$residuals,W,na.rm=T)

#Como el p valor es menor a 0.05 se rechaza la H0 que nos indica que los residuos no tienen correlacion espacial

# Seleccion de los parametros significativos

modelo_analisis1 <- Modelo_Espana_1[["MipLasso"]]
Resumen <- summary(modelo_analisis1)
Coeficientes <- data.frame(Resumen$coefficients)
Coeficientes_significativos <- rownames(Coeficientes)[which(Coeficientes$Pr...t..<0.05)]
```

Referencias bibliográficas

[1] Y. Chasco. *Métodos gráficos del análisis exploratorio de datos espaciales*. Anales de Economía Aplicada, Asociación Española de Economía Aplicada, 2003.

[2] R. Cherodian. *Eigenvector Spatial Filtering and Lasso: Theory and Applications*. PhD thesis, School of Economics, University of Kent, 2023.

[3] S. Glen. Moran's I: Definition, Examples. https://www.statisticshowto.com/morans-i/.

[4] Y. Griffith, D. & Chun. Spatial autocorrelation and spatial filtering. *Handbook of Regional Science, Springer-Verlag Berlin Heidelberg*, 2014.

[5] Nieto J. "¿Qué pasa con la seguridad en Valencia? Ya supera a Madrid en delincuencia y se acerca a Barcelona". *Diario digital El Español*, 2022.

[6] P. Moraga. *Spatial Statistics for Data Science: Theory and Practice with R*. Chapman & Hall/CRC, 2023.

[7] H. Seya et al. Application of LASSO to the eigenvector selection problem in Eigenvector-based Spatial Filtering. *Geographical Analysis 47(3), 284–299*, 2014.

[8] R Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288.*, 1996.