Modelos de regresión espacial

Práctica Profesional II

Maximiliano S. Lioi

Verano 2023

Planificación de la Práctica

Práctica Profesional II

- **Carrera**: Ingeniería Civil Matemática [UChile]
- **Requisitos**: MA4702 Programación Lineal Mixta
- **Empleador**: Pedro Donoso
- **Trabajo**: Investigación sobre modelos de regresión espacial, acompañado de simulaciones de predicción de precios de viviendas con datos demográficos del censo de EEUU y comparativas entre estos modelos

Contenidos

Motivación

Predecir el precio de las viviendas en una determinada ubicación utilizando datos con componente espacial trata con el desafío de la **autocorrelación espacial** y **heterogeneidad espacial** inmersa en estos, los cuales afectan suposiciones clásicas que se toman en modelos más simples como regresión lineal

- Residuos i.i.d
- Homocedasticidad

Si no se cumplen las hipótesis, no podemos asegurar la calidad del modelo

Motivación

La investigación se enfoca en el desarrollo de **modelos de regresión espacial**, los cuales adaptan la naturaleza espacial de los datos en sus formulaciones, tomar estas cualidades en cuenta permite obtener modelos más precisos, cuyos resultados pueden ser útiles en la industria, como es el caso de la predicción de precios de vivienda dentro del mercado inmobiliario.

Problema de la predicción espacial

Datos espaciales

En el contexto de datos espaciales, trabajamos con un conjunto de observaciones de datos asociados con una componente espacial

$$\mathcal{C} = \{(x(s_i), y(s_i))|\ i \in \mathbb{N}, 1 \leq i \leq n\}$$

donde $n \in \mathbb{N}$ es la cantidad de muestras, $s_i \in \mathbb{R}^2$ es un vector de coordenadas espaciales, $x(s_i) \in \mathbb{R}^m$ son las variables explicativas del modelo e $y(s_i) \in \mathbb{R}$ es la respuesta a dichas variables.

Notación

Este mismo conjunto también puede ser descrito en formato matricial como $(X, Y) \in \mathbb{R}^{(m+1)\times n}$ donde

$$X = [x(s_1), x(s_2), \ldots, x(s_n)]^T \in \mathbb{R}^{m\times n}$$

$$Y = [y(s_1), y(s_2), \ldots, y(s_n)]^T \in \mathbb{R}^n$$

Además, dejamos implicita la componente espacial asociada a los datos, de manera que $(x(s_i), y(s_i)) = (x_i, y_i)$

Predicción espacial

El problema de la predicción consiste en que, dado un conjunto de muestras de data espacial, buscamos modelar una función f tal que

$$Y = f(X) + \varepsilon$$

donde ε es un término para el error, una vez que el modelo se entrena, minimizando el error, puede usarse para precedir la respuesta en otra locación dada sus variables explicativas.

Predicción espacial

La predicción espacial se diferencia de la predicción usual, puesto que en este último se asume que las muestras son i.i.d, y por lo tanto, dado un modelo entrenado en la función f, podemos usarlo para predecir $y(s) = f(x(s))$ para todo s, sin embargo, la suposición de que los datos son i.i.d no se cumple para datos espaciales, esto debido a la relación implícita que existe entre posiciones cercanas en una región.

- Tobler W.R: "Todo se relaciona con todo lo demás, pero las cosas más cercanas se relacionan más que las cosas distantes"

Desafíos

Respecto de los desafíos presentes asociados a la predicción espacial, abordamos los siguientes:

- *Autocorrelación espacial* → Los datos espaciales no son estadísticamente independientes entre sí, al contrario estos, están correlacionados, y la intensidad de dicha relación depende de la distancia que exista entre estos.
- *Heterogeneidad espacial* → Los datos que conforman cada zona no siguen una distribución idéntica, los valores que representan la relación entre los regresores y la variable dependiente, varía dependiento de la ubicación geográfica.

Importando datos

Se importan datos con *tidycensus* provenientes de la American Community Survey (ACS) (2016-2020), encuesta estadounidense que recoge datos demográficos anualmente.

- Primeramente, se activa una *api-key* para poder descargar datos mediante *tidycensus*.
- Queremos predecir precio de viviendas, suponemos que existe autocorrelación espacial sobre esta variable, pues aquellas viviendas que se encuentran cerca, comparten su entorno, y por tanto una serie de factores relacionados al consumo.
- Se toman datos georreferenciados con nivel geográfico de distritos para la ciudad de NY
- Para ello pedimos datos de los condados de: 'Bronx', 'Kings', 'New York', 'Queens', 'Richmond'
- Se importan alrededor de 2000 distritos censales

Las variables que importamos, para usar en los distintos modelos de predicción, son las siguientes:

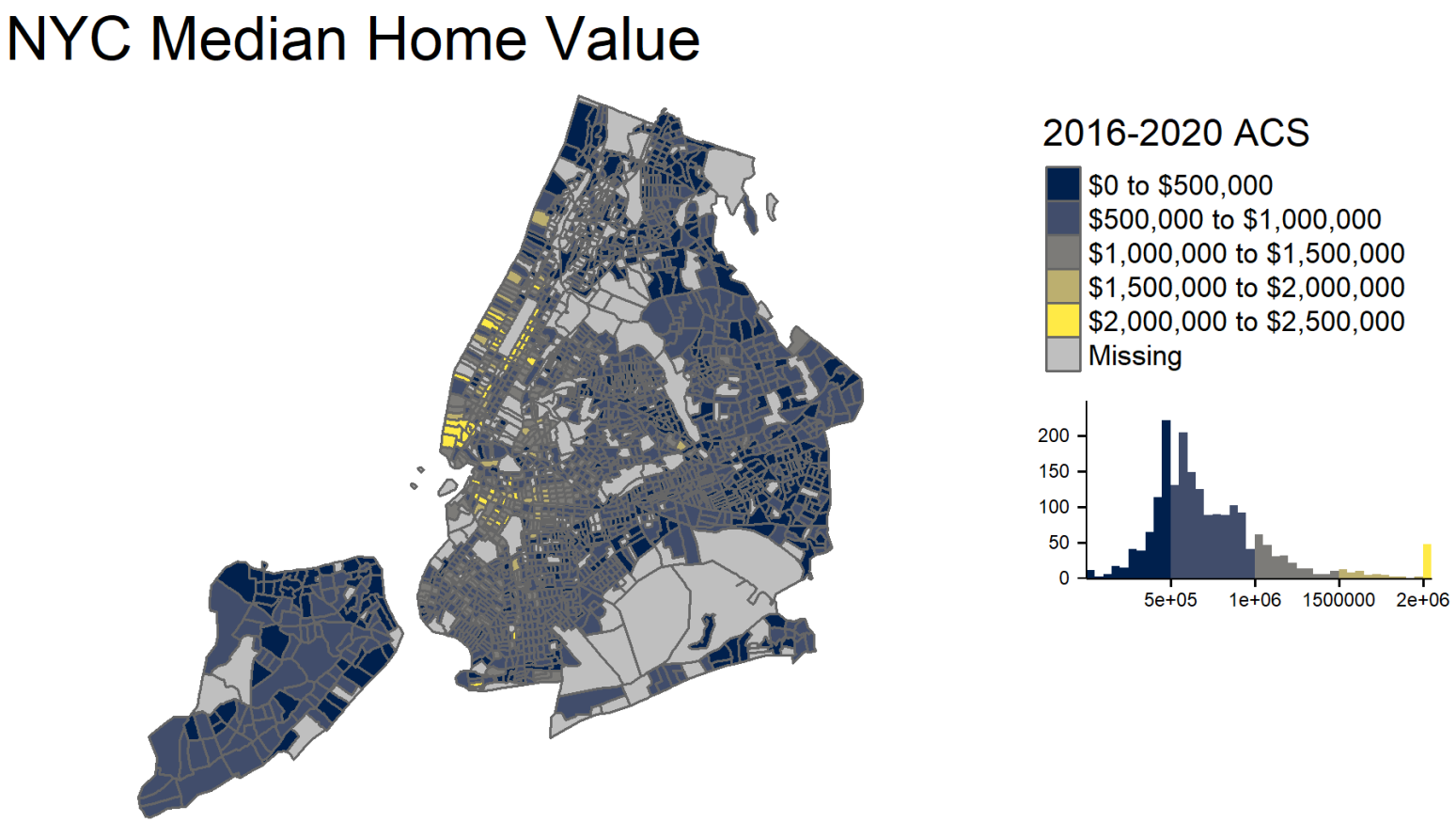
- **median value** : El valor medio de la vivienda del distrito censal, nuestra variable a predecir
- median rooms : Cantidad media de habitaciones por casa en el tramo censal
- total population : Población total en el tramo censal
- median age : Edad media de la población en el tramo censal

- median year built : Mediana del año donde se construyó la vivienda
- median income : Ingreso medio de los hogares en el tramo censal
- pct college : Porcentaje de la población de 25 años o más con un título universitario de cuatro años
- pct foreign born: Porcentaje de la población que nació fuera de EE.UU
- pct white : Porcentaje de la población que se identifica como blanco no-hispano, se sigue la misma lógica con pct black, pct asian, pct hispanic dentro del distrito.

Preparando datos

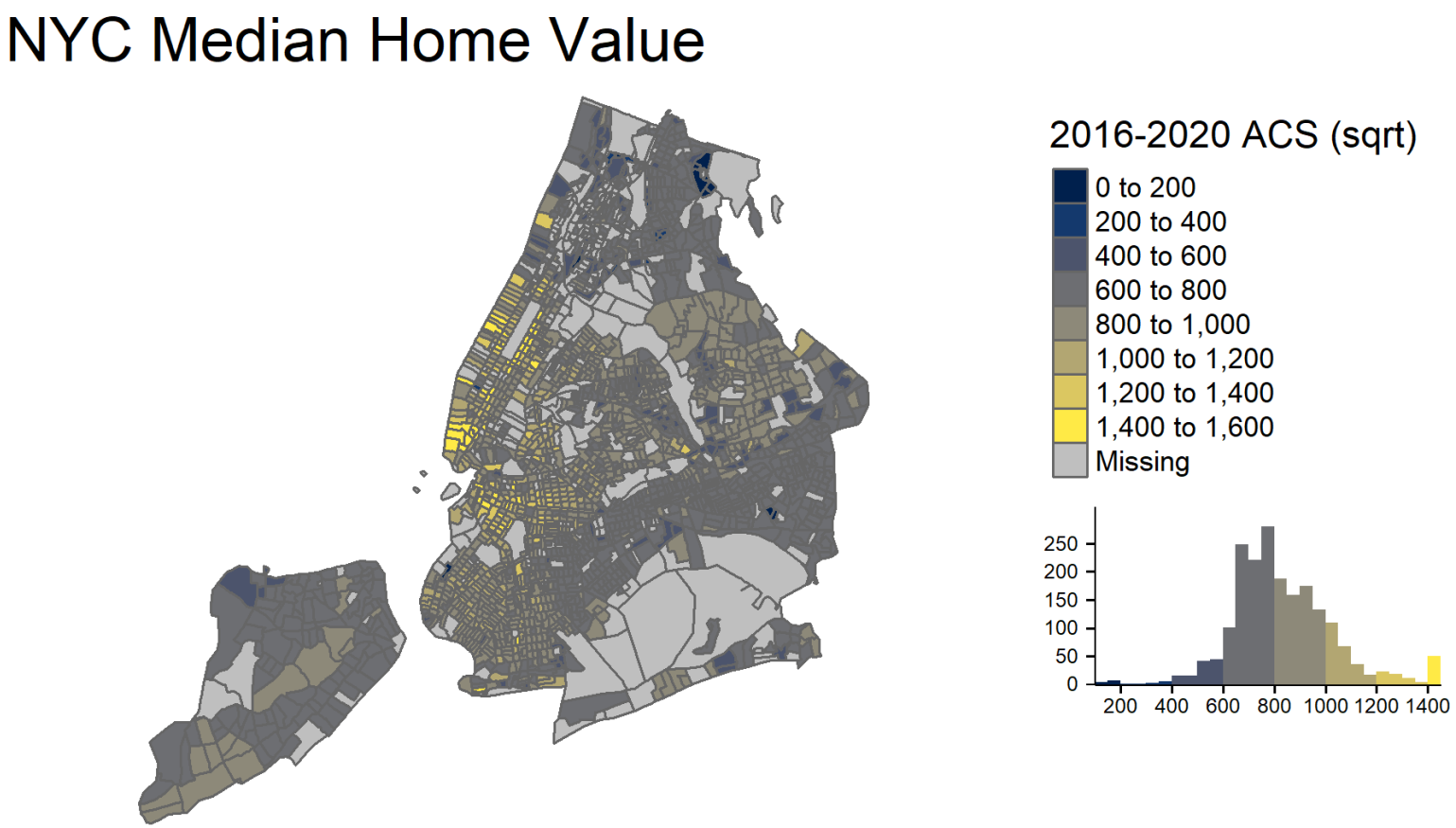
- Se realiza una limpieza de las muestras (data scrubbing), identificando datos incompletos, incorrectos o inexactos.
- Nuestra variable dependiente para los modelos de regresión será el valor promedio de la vivienda en NYC, esta la importamos bajo el nombre de median value.
- Dadas las componentes espaciales, y la cantidad total de personas por distrito, creamos la variable *pop density*, que mide densidad de población en el tramo censal por metros cuadrados.

Visualización: NYC Median Home Value



- El histograma presenta una asimetría a la derecha, i.e, existe una población no menor cuyas viviendas poseen un valor muy alejado de la mediana.
- Se ajustan los datos como una distribución normal, esto para gozar de más propiedades e inducir normalidad en los residuos del modelo.

Visualización: NYC Median Home Value



Autocorrelación espacial: Efectos sobre una regresión lineal

- Para medir la autocorrelación espacial presente en los datos, hacemos uso del índice de Moran I el cual nos entrega una magnitud de la presencia de este fenómeno para n observaciones de una misma variable.

Definimos el índice de Moran como:

$$I = \frac{n}{W} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y(s_i) - \bar{y})(y(s_j) - \bar{y})}{\sum_{i=1}^{n}(y(s_i) - \bar{y})^2}$$

donde $\bar{y} = \sum_{i=1}^{n} y(s_i)/n$ con n el número total de muestras, a partir de ahora, para simplificar la notación, $y(s_i) = y_i$ dejando implicita la componente espacial, además $W = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$, y de forma abreviada, $W = (w_{ij})_{i,j=1}^{n}$, que consideramos como una **matriz de pesos estandarizada**.

Matriz de pesos estandarizada

Entenderemos por **matriz de pesos estandarizada** aquella que cumple las siguientes propiedades

- Supondremos $w_{ij} \geq 0 \quad \forall i, j \in \{1, \ldots, n\}$, pues w_ij es una magnitud de la influencia que hay entre los vecinos i y j
- Podemos suponer que $w_{ii} = 0 \quad \forall i \in \{1, \ldots, n\}$, pues el hecho de que un punto tenga influencia con el mismo puede generar ruido, sin embargo esto dependera del modelo
- Suponemos que la influencia que tiene y_i sobre y_j es la misma que la de y_j sobre y_i, esto es, $w_{ij} = w_{ji} \quad \forall i, j \in \{1, \ldots, n\}$, es decir, W es simétrica

Índice de Moran

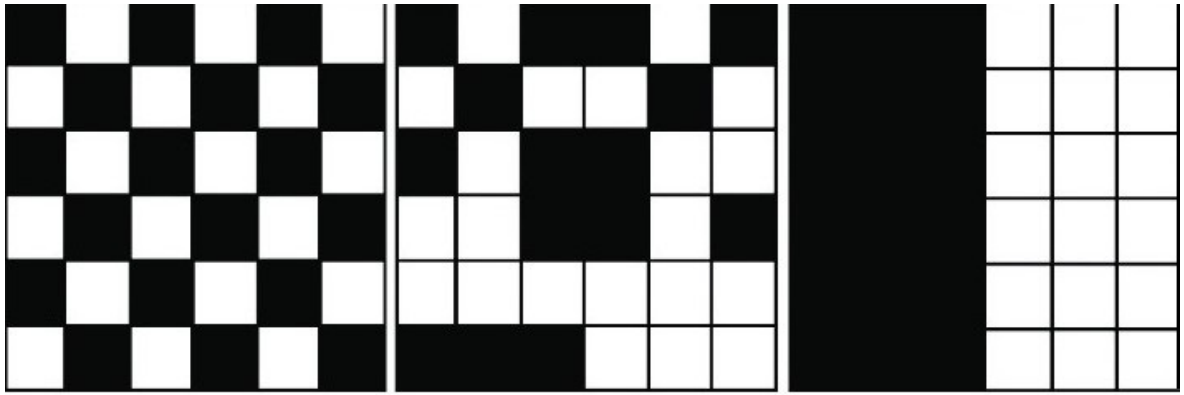
- **Row-normalized**: Para cada $i \in \{1, \ldots, n\}$ se tiene que $\sum_{j=1}^{n} w_{ij} = \alpha$ para algún $\alpha \in \mathbb{R}$ esto pues, entendemos la suma $\sum_{j=1}^{n} w_{ij}$ como la influencia total que existe entre el vecino i y todos sus vecinos j, por lo que pedimos que la influencia total sea la misma para cada vecino.

Propiedad

Para $I = \frac{n}{W} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ donde $\bar{y} = \sum_{i=1}^{n} y_i / n$ con W matriz de pesos estandarizada e $y \in \mathbb{R}^n$, se cumple que $I \in [-1, 1]$.

$\rightarrow$ Mide que tan similar es una variable respecto de las variables cercanas, por ejemplo, si este se acerca a -1 nos dice que cada en cada región, esta tiende a tener valores distinto de su entorno, y en caso contrario, si se acerca a 1, cada región es similar a su entorno.

Ejemplo



- $I \approx -1 \rightarrow$ Se clusterizan valores distintos entre sí (dispersión perfecta)
- $I \approx 0 \rightarrow$ No existe una clara relación entre los valores de los datos y sus entornos, es decir, no hay pruebas de autocorrelación espacial.
- $I \approx 1 \rightarrow$ Se clusterizan valores similares entre sí

Hipótesis nula H_0

Además, si consideramos la hipótesis nula:

$$H_0 : \text{No existe autocorrelación espacial}$$

La cual consiste en suponer que, dada una permutación de la muestra, es decir, cambiando el orden en el que vienen los datos $(x_i)_{i=1}^N$, y fijando la matriz de pesos antes del reordenamiento, no se debiese tener un impacto sobre el valor de I, pues estos no muestran relacionarse entre sí, se tiene que

$$\mathbb{E}(I) = -\frac{1}{N-1}$$

Esto es consistente con el modelo, pues para una muestra grande, de no existir autocorrelación espacial, esperamos que $I \approx 0$, podemos usar esta propiedad para inferencias estadísticas por medio del p-valor

Regresión Lineal

Planteamos un modelo de regresión lineal para los datos importados
La formulación es la siguiente:

$$\sqrt{median\ value} = \alpha + \beta_1(median\ rooms) + \beta_2(median\ income)$$

$$+\beta_3(pct\ college) + \beta_4(pct\ foreign\ born) + \beta_5(pct\ white) +$$

$$+\beta_6(median\ age) + \beta_7(percent\ ooh)$$

$$+\beta_8(pop\ density) + \beta_9(total\ population) + \varepsilon$$

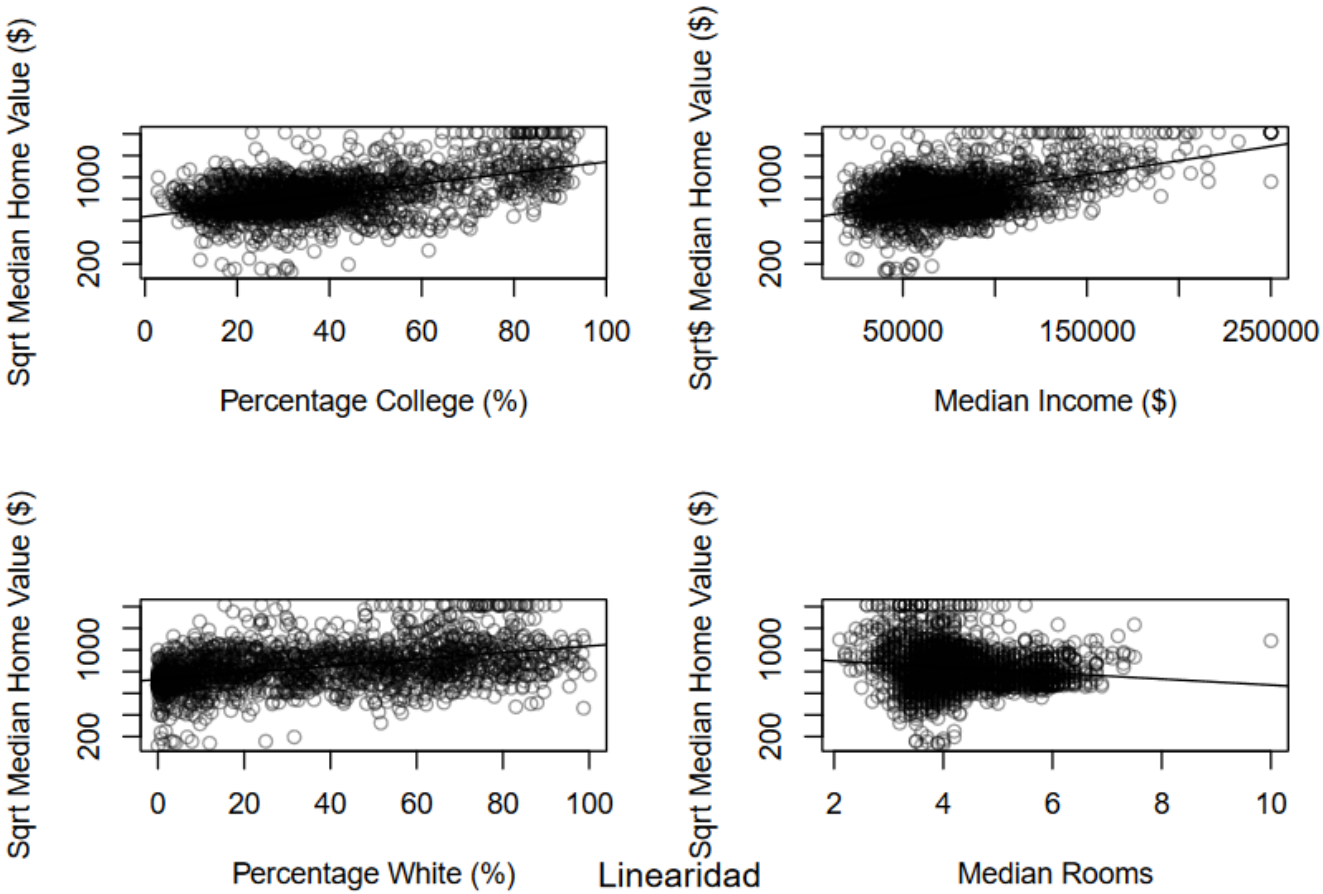
De donde, a partir de n observaciones de la variable dependiente junto a sus regresores, buscamos estimar los valores α y β_i, mediante mínimos cuadrados, es decir, tenemos para cada observación un error asociado

Hipótesis de Gauss-Márkov

Un modelo de regresión lineal requiere de una serie de hipótesis para asegurar la calidad de las estimaciones de β (supuestos de Gauss-Márkov) tales como:

- **Linearidad**: Los regresores poseen una relación lineal con la variable de respuesta
- **Independencia**: Las observaciones representan una muestra aleatoria distribuida identica e independiente, de manera que es generalizable para el total de la población
- **Error con media cero**: Consideramos el error de cada observación ε_i variable aleatoria tal que $\mathbb{E}(\varepsilon_i) = 0$
- **Normalidad**: Suponemos que ε_i sigue una distribución normal de la forma $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **Homocedasticidad e incorrelación**: Los errores del observaciones del modelo poseen la misma varianza σ^2 y son tales que $Cov(\varepsilon_i, \varepsilon_j)$.

Linearidad



Linearidad

Calibración del modelo

Para la calibración de los parámetros β y α del modelo de regresión lineal, podemos usar la función *lm* que viene incluida con los paquetes básicos del lenguaje R.

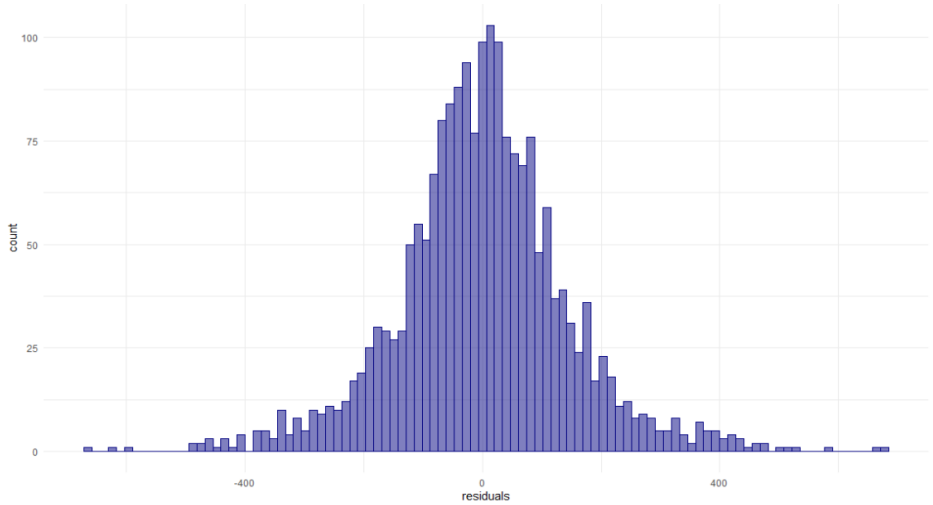
```
Call:
lm(formula = formula, data = nyc_data_prepped)

Residuals:
    Min      1Q  Median      3Q     Max
-657.75  -78.17    0.32   77.44  683.88

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.048e+02  3.974e+01  12.704  < 2e-16 ***
median_rooms     6.209e+01  6.465e+00   9.605  < 2e-16 ***
median_income    2.091e-03  1.704e-04  12.272  < 2e-16 ***
pct_college      1.696e+00  3.624e-01   4.680 3.06e-06 ***
pct_foreign_born -6.281e-01  3.245e-01  -1.936 0.053039 .
pct_white        2.421e+00  2.376e-01  10.192  < 2e-16 ***
pct_black        1.003e+00  2.074e-01   4.838 1.42e-06 ***
pct_hispanic     -1.399e-02  2.127e-03  -6.577 6.16e-11 ***
pct_asian        3.108e+00  3.000e-01  10.360  < 2e-16 ***
median_age       -2.166e+00  5.984e-01  -3.620 0.000302 ***
percent_ooh      -4.513e+00  2.807e-01 -16.077  < 2e-16 ***
pop_density      1.046e-03  3.401e-04   3.075 0.002133 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 146.8 on 1957 degrees of freedom
Multiple R-squared:  0.4569,    Adjusted R-squared:  0.4539
F-statistic: 149.7 on 11 and 1957 DF,  p-value: < 2.2e-16
```

Valor de R^2

→ Observando el valor R^2, coeficiente que mide el porcentaje de varianza de la variable dependiente que es explicado con las variables del modelo, un $R^2 = 0.4569$ nos revela que **el modelo es deficiente para la predicción**, una de las razones de esto es debido a la autocorrelación espacial presente en los datos.
→ Para hacer cuenta de esto último, se estudian los residuos del modelo.

Residuos

Mediante un histograma vemos su distribución:



Normalizar la variable dependiente ayuda con la distribución de los residuos, sin embargo, la suposición de independencia de los residuos comúnmente se viola en los modelos que utilizan datos espaciales. Esto último por la autocorrelación espacial presente en el término de error, lo que significa que el rendimiento del modelo en sí mismo depende de la ubicación geográfica.

Matriz de pesos

Podemos evaluar esto utilizando el índice de Moran, buscamos armar una matriz de pesos, para medir la interacción entre cada tramo censal, para ello generamos una "neighborhood list" con estructura Queen en R con el paquete **spdep**, que luego damos estructura matricial. Tenemos la siguiente estructura

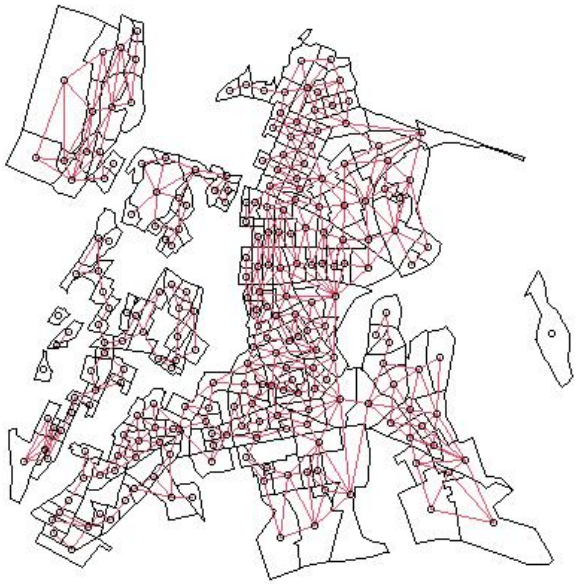
```
Neighbour list object:
Number of regions: 1967
Number of nonzero links: 10700
Percentage nonzero weights: 0.2765509
Average number of links: 5.439756
5 regions with no links:
86 90 242 1727 1767
```

- Tenemos 1967 distritos censales en NYC (aquellos con valor NA se omiten)
- El porcentaje de las entradas $w_{i,j}$ tales que $w_{i,j} \neq 0$ es 0.276
- El número promedio de conexiones entre distritos censales, el promedio de cuantas conexiones posee cada distrito es 5.44
- Existen 5 distritos que no se conectan con nadie.

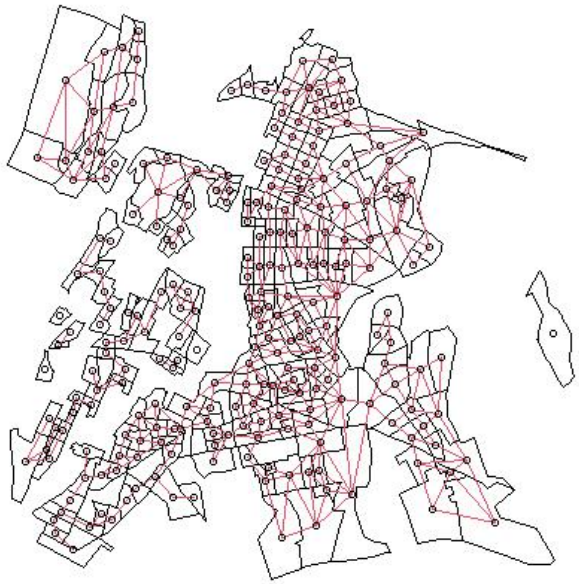
Visualización: Matriz de pesos en Bronx

A modo de ejemplo, vemos la estructura que genera la matriz para la zona de Bronx

Queen



Rook



Test de Índice de Moran

Con la matriz de pesos podemos realizar un test de índice de Moran sobre los residuos del modelo lineal ya calibrado

Definición: p-valor

Se define como la probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula cierta.

$$\text{p-value} = \mathbb{P}(\text{valor tan extremo o más} | H_0)$$

Si el p-valor cumple con la condición de ser menor que un nivel de significancia impuesto arbitrariamente, este se considera como un resultado estadísticamente significativo y, por lo tanto, permite rechazar la hipótesis nula.
El p-valor está basado en la presunción de que una hipótesis nula es cierta, y es por tanto una medida de significación estadística.

Realizando un test de índice de Moran

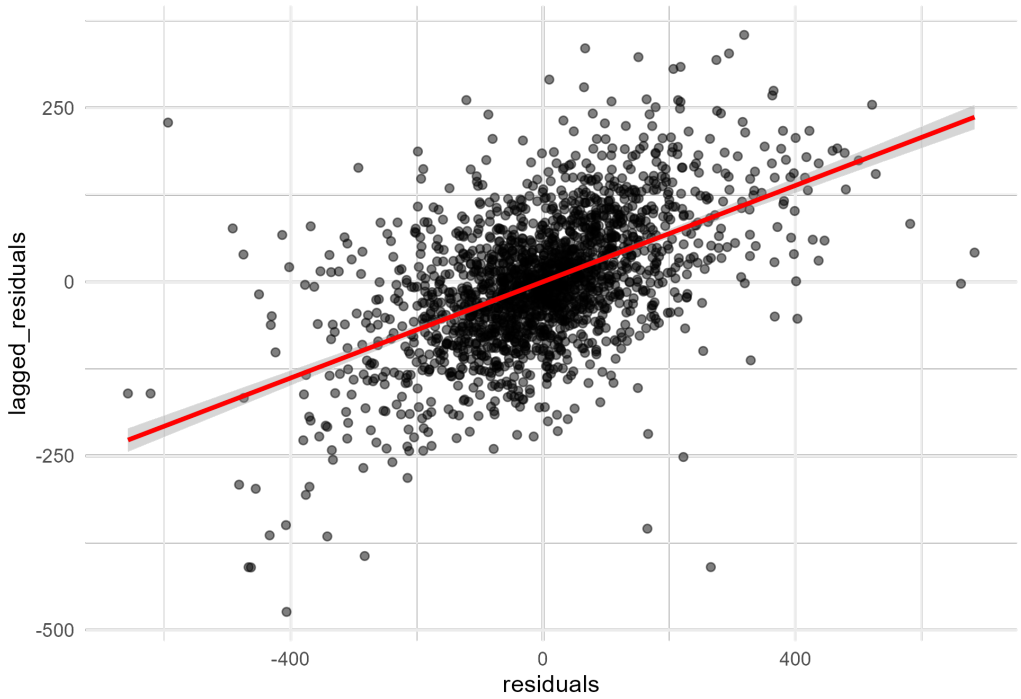
```
        Moran I test under randomisation

data:  nyc_census_data_prepped$residuals
weights: wts  n reduced by no-neighbour observations


Moran I statistic standard deviate = 24.186, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.3456912899     -0.0005099439      0.0002048902
```

- El valor esperado de I bajo la hipótesis nula H_0 de no autocorrelación es de -0.00051, por lo que un valor de $I = 0.345$ es estadísticamente significativo, el p-valor que toma el estadístico es $< 2.2e-16$ lo que nos habla de la significancia que posee en el modelo.
- El test de índice de Moran nos lleva a rechazar la hipótesis nula, por lo demás, vemos que el índice de Moran tiene un valor positivo, lo que nos dice que en el mapa, distritos censales con atributos similares tienden a agruparse (clusters)

Moran Scatterplot

Consta de un gráfico de datos espaciales comparando los valores de una variable de respuesta Y con sus valores *lagged*, que denotamos $Y_{lag} = WY$, el cual es una versión ponderada de las observaciones vecinas de cada punto.

→ Realizando un gráfico entre los residuos con su versión *lagged* visualizamos una correlación positiva



Spatial Lag Model

El modelo de *lag* espacial tiene en cuenta la dependencia espacial al incluir una variable de retraso sobre el resultado del modelo. Al hacerlo, se tienen en cuenta los efectos de dependencia espacial, es decir, la posibilidad de que los valores en áreas vecinas influyan en los valores en una ubicación determinada.
El modelo tiene la formulación:

$$Y = \alpha + \rho WY + X\beta + \varepsilon$$

Donde formulamos el valor de la respuesta Y como una versión ponderada de los valores que toma la respuesta en los alrededores de cada dato, sumado a una dependencia lineal que tenga la variable Y con los regresores X, tomando W una matriz de pesos estandarizada y ρ refleja la intensidad de la dependencia espacial.

Spatial Lag Model

Para visualizarlo más claramente, vemos para cada fila tenemos

$$Y_i = \alpha + \rho Y_{lag-i} + \sum_k \beta_k X_{ik} + \varepsilon_i \quad \forall i \in \{1, ..., n\}$$

donde

$$Y_{lag-i} = \sum_j w_{ij} Y_j$$

Por lo que, la variable Y_{lag-i} representa una versión ponderada de los valores que toman las respuestas Y_j que son vecinas de la respuesta Y_i, es decir, aquellos Y_j tales que $w_{ij} \neq 0$.

Spatial Lag Model

Podemos ver este modelo como una versión ponderada por por el término $(I - \rho W)^{-1}$ sobre el valor esperado $\alpha + X\beta$ y el residuo del modelo de una regresión lineal clásica, basta notar que

$$
\begin{aligned}
Y &= \alpha + \rho WY + X\beta + \varepsilon \\
\implies Y - \rho WY &= \alpha + X\beta + \varepsilon \\
\implies (I - \rho W)Y &= \alpha + X\beta + \varepsilon \\
\implies Y &= \alpha' + (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon
\end{aligned}
$$

Para estimar los parámetros de este modelo, podemos usar el método de máxima verosimilitud, usaremos en este caso funciones incluidas en el paquete *spatialreg*.

Calibración: Lag Model

```
Call:spatialreg::lagsarlm(formula = formula, data = nyc_data_prepped,
listw = wts, zero.policy = TRUE)

Residuals:
      Min        1Q    Median        3Q       Max
-634.1228  -62.9577   -1.1571   60.0997  659.2782

Type: lag
Regions with no neighbours included:
 86 90 242 1728 1768
Coefficients: (asymptotic standard errors)
                   Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)       1.3959e+02  3.5798e+01   3.8995 9.638e-05
median_rooms      5.3798e+01  5.3775e+00  10.0042 < 2.2e-16
median_income     1.3189e-03  1.4271e-04   9.2419 < 2.2e-16
pct_college       8.6781e-01  3.0321e-01   2.8620  0.004209
pct_foreign_born -1.4747e-01  2.6966e-01  -0.5469  0.584463
pct_white         1.0503e+00  1.9994e-01   5.2529 1.497e-07
pct_black         3.4867e-01  1.7197e-01   2.0275  0.042610
pct_hispanic     -8.7095e-03  1.7668e-03  -4.9296 8.242e-07
pct_asian         1.4127e+00  2.5211e-01   5.6036 2.099e-08
median_age       -9.4528e-01  4.9688e-01  -1.9024  0.057114
percent_ooh      -3.1191e+00  2.3721e-01 -13.1492 < 2.2e-16
pop_density       2.0319e-04  2.8196e-04   0.7207  0.471119

Rho: 0.54771, LR test value: 589.69, p-value: < 2.22e-16
Asymptotic standard error: 0.020508
    z-value: 26.707, p-value: < 2.22e-16
Wald statistic: 713.26, p-value: < 2.22e-16

Log likelihood: -12316.63 for lag model
ML residual variance (sigma squared): 14799, (sigma: 121.65)
Nagelkerke pseudo-R-squared: 0.59749
Number of observations: 1969
Number of parameters estimated: 14
AIC: 24661, (AIC for lm: 25249)
LM test for residual autocorrelation
test value: 1.4002, p-value: 0.23669
```

- Al igual que el modelo de regresión lineal se nos entregan los valores de α (intercept) y los β_i asociados a cada regresor, junto a sus p-valores, lo que nos permite hacer un análisis de la significancia que tiene cada regresor en el modelo
- Además, el modelo calibra el valor de ρ, que refleja la intensidad de la matriz W, este toma un valor positivo y estadísticamente significativo, pues su p-valor es del orden de 2.22×10^{-16}
- Entrega un valor de pseudo-R cuadrado, que es mayor al valor correspondiente para el modelo de regresión lineal, con un valor de 0.59.

Comentario

Los valores de pseudo R-cuadrado no son directamente comparables al R-cuadrado de los modelos de mínimos cuadrados, tampoco se pueden interpretar como la proporción de la variabilidad en la variable dependiente que es explicada por el modelo, más bien, las medidas de pseudo R-cuadrado son medidas relativas entre modelos lineales similares que indican qué tan bien el modelo explica los datos.
El R^2 de **Nagelkere** es una adaptación del R^2 de Cox y Snell adaptado para tomar los valores entre 0 y 1

Spatial Error Model

En contraste con el modelo de Lag, los modelos de error espacial incluyen un *lag* en el término de error del modelo. Esto está diseñado para capturar procesos espaciales latentes que actualmente no se están teniendo en cuenta en la estimación del modelo y, a su vez, aparecen en los residuos del modelo
El modelo de error espacial se puede escribir de la siguiente manera:

$$Y = \alpha + X\beta + u$$

Donde u es un término autorregresivo de la forma

$$u = \lambda u_{lag} + \varepsilon$$

Y

$$u_{lag} = Wu \implies u = \lambda Wu + \varepsilon$$

Spatial Error Model

Es decir, para cada fila

$$Y_i = \alpha + \sum_k \beta_k X_{ik} + \lambda u_{lag-i} + \varepsilon_i$$

Donde

$$u_i = \lambda u_{lag-i} + \varepsilon_i \ , \ u_{lag-i} = \sum_j w_{ij} u_j$$

Por lo que buscamos calibrar λ y u con un modelo de la forma

$$Y_i = \alpha + \sum_k \beta_k X_{ik} + \lambda \sum_j w_{ij} u_j + \varepsilon_i$$

Notemos que llegamos a una formulación similar al caso del modelo Lag, sin embargo buscamos calibrar u que se incluye en el término del error al considerarlo de la forma $u = \lambda W u + \varepsilon$.

Calibración: Spatial Error Model

Para la calibración de parámetros hacemos uso del paquete *spatialreg*

```
Call:spatialreg::errorsarlm(formula = formula, data = nyc_data_prepped,
    listw = wts, zero.policy = TRUE)

Residuals:
      Min        1Q    Median        3Q       Max
-615.6737  -63.4504   -3.1502   59.4877  647.2780

Type: error
Regions with no neighbours included:
 86 90 242 1728 1768
Coefficients: (asymptotic standard errors)
                  Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)       5.9647e+02  4.0882e+01  14.5901 < 2.2e-16
median_rooms      5.9986e+01  6.1124e+00   9.8138 < 2.2e-16
median_income     1.1220e-03  1.6129e-04   6.9563 3.492e-12
pct_college       1.5031e+00  3.7370e-01   4.0223 5.764e-05
pct_foreign_born -4.7362e-01  3.7505e-01  -1.2628  0.206662
pct_white         9.5976e-01  3.1394e-01   3.0572  0.002235
pct_black        -4.9166e-01  3.0898e-01  -1.5912  0.111561
pct_hispanic     -5.9840e-03  1.9980e-03  -2.9950  0.002744
pct_asian         1.2736e+00  3.9161e-01   3.2523  0.001145
median_age       -8.8651e-01  5.6659e-01  -1.5647  0.117665
percent_ooh      -3.1476e+00  2.5688e-01 -12.2535 < 2.2e-16
pop_density      -5.3587e-04  3.5238e-04  -1.5207  0.128333

Lambda: 0.66716, LR test value: 552.3, p-value: < 2.22e-16
Asymptotic standard error: 0.019925
    z-value: 33.484, p-value: < 2.22e-16
Wald statistic: 1121.2, p-value: < 2.22e-16

Log likelihood: -12335.32 for error model
ML residual variance (sigma squared): 14451, (sigma: 120.21)
Nagelkerke pseudo-R-squared: 0.58977
Number of observations: 1969
Number of parameters estimated: 14
AIC: 24699, (AIC for lm: 25249)
```

- La calibración entrega un modelo cuyo pseudo-R^2 es de 0.589, el cual es ligeramente menor al pseudo-R^2 del modelo Lag, por lo demás, *spatialreg* entrega los parámetros del modelo en el mismo formato.
- Vemos que el valor que toma λ es estadísticamente significativo, con una magnitud de $\lambda = 0.667$ y un bajo p-valor, esto da cuenta nuevamente de la importancia de considerar la autocorrelación espacial del modelo

Los modelos de lag espacial y de error espacial ofrecen enfoques alternativos para tener en cuenta los procesos de autocorrelación espacial al ajustar modelos

Comentario

Pregunta: ¿Qué modelo usar?

Según Walker K, respecto de cual de los dos modelos usar, debe considerarse el contexto del tema en estudio, por ejemplo, si los efectos de dependencia espacial están relacionados con las hipótesis que el analista evalúa, como es el caso del efecto de los valores de las viviendas vecinas en torno al valor de una vivienda, se puede preferir un modelo Lag, por otro lado, si hay factores autocorrelacionados espacialmente y que probablemente influyen en la variable de respuesta Y pero son difíciles de medir cuantitativamente, como puede ser la discriminación o el sesgo racial en el mercado de la vivienda, podría ser preferible un modelo de error espacial.

Test de índice de Moran: Lag Model

Para complementar, podemos aplicar un test de índice de Moran sobre los residuos de ambos modelos de regresión espacial, para ver si resuelven el problema de la dependencia espacial de los errores.

```
        Moran I test under randomisation

data:  lag_model$residuals
weights: wts  n reduced by no-neighbour observations


Moran I statistic standard deviate = -0.64445, p-value = 0.7404
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
    -0.0097224085     -0.0005094244      0.0002043731
```

Test de índice de Moran: Error Model

```
        Moran I test under randomisation

data:  error_model$residuals
weights: wts  n reduced by no-neighbour observations


Moran I statistic standard deviate = -3.3594, p-value = 0.9996
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
    -0.0485344554     -0.0005094244      0.0002043707
```

- Ambos modelos reducen el índice de Moran, acercandose a su valor esperando bajo la hipótesis nula H_0 de no autocorrelación, sin embargo, el modelo de error espacial hace un mejor trabajo sobre la autocorrelación espacial en los residuos
- Esto pues, si bien los modelos toman valores similares de I, el modelo de error espacial asigna un p-valor de 0.99 al estadístico mientras que en el modelo Lag toma un p-valor de 0.7404, eliminando por completo la dependencia espacial en el error.

Comparativa: Regresión Lineal, Lag Model, Error Model

Realizamos una comparativa entre el primer modelo de regresión lineal y los dos modelos de regresión espacial que hemos introducido

	Model 1	Model 2	Model 3
(Intercept)	504.79 ***	139.59 ***	596.47 ***
	(39.74)	(35.80)	(40.88)
median_rooms	62.09 ***	53.80 ***	59.99 ***
	(6.46)	(5.38)	(6.11)
median_income	0.00 ***	0.00 ***	0.00 ***
	(0.00)	(0.00)	(0.00)
pct_college	1.70 ***	0.87 **	1.50 ***
	(0.36)	(0.30)	(0.37)
pct_foreign_born	-0.63	-0.15	-0.47
	(0.32)	(0.27)	(0.38)
pct_white	2.42 ***	1.05 ***	0.96 **
	(0.24)	(0.20)	(0.31)
pct_black	1.00 ***	0.35 *	-0.49
	(0.21)	(0.17)	(0.31)
pct_hispanic	-0.01 ***	-0.01 ***	-0.01 **
	(0.00)	(0.00)	(0.00)
pct_asian	3.11 ***	1.41 ***	1.27 **
	(0.30)	(0.25)	(0.39)
median_age	-2.17 ***	-0.95	-0.89
	(0.60)	(0.50)	(0.57)
percent_ooh	-4.51 ***	-3.12 ***	-3.15 ***
	(0.28)	(0.24)	(0.26)
pop_density	0.00 **	0.00	-0.00
	(0.00)	(0.00)	(0.00)
rho		0.55 ***	
		(0.02)	
lambda			0.67 ***
			(0.02)
N	1969	1969	1969
R2	0.46	0.63	0.65

*** p < 0.001; ** p < 0.01; * p < 0.05.

Respecto de la calidad de predicción según R^2, tenemos en orden de menor a mayor desempeño

1. Regresión Lineal - lm() $\rightarrow R^2 = 0.46$
2. Spatial Lag Model - spatialreg::lagsarlm() $\rightarrow R^2 = 0.63$
3. Spatial Error Model - spatialreg::errorsarlm() $\rightarrow R^2 = 0.65$

Por lo que, ambos modelos de regresión espacial, logran una amplia mejoría respecto del modelo de regresión lineal simple.

Geographically Weighted Regression

Motivación

Los modelos abordados, estiman relaciones globales entre la variable dependiente y sus regresores $(\beta_i)_{i=1}^{m}$, es decir, a pesar de que estos modelos dan cuenta de la autocorrelación espacial, asumen que los coeficientes β son iguales para toda ubicación, por lo tanto, no abordan la **heterogeneidad espacial**, ya que los datos que conforman un sector no siguen una distribución por lo que es razonable suponer que la relación entre los regresores y la variable dependiente que se observa para toda la región, varíe significativamente entre sectores.

Geographically Weighted Regression

El siguiente modelo, aborda la heterogeneidad espacial aprendiendo un conjunto de parámetros β_{ik} asociado a los regresores del modelo en cada ubicación s_i.

Mientras que la formulación de un modelo de regresión lineal, el cual tiene la forma

$$Y_i = \alpha + \sum_{k=1}^{m} \beta_k X_{ik} + \varepsilon_i$$

Con los β_k idénticos para cada locación s_i, la formulación del modelo de regresión geográficamente ponderado (GWR) para una ubicación s_i dada, se escribe como

$$Y_i = \alpha + \sum_{k=1}^{m} \beta_{ik} X_{ik} + \varepsilon_i$$

Donde el intercepto α_i , los regresores $(X_{ik})_{k=1}^{m}$, y los errores ε_i están todos a locación s_i.
Por lo demás, los parámetros β_{ik} serán coeficientes locales para el regresor X_k con ubicación s_i
El coeficiente $\beta(s_h)$ en locación s_h puede entrenarse vía método de mínimos cuadrados ponderados, donde el peso que se le asigna a cada muestra depende de su distancia a s_h, es decir, buscamos encontrar $\beta(s_h)$ tal que

$$\beta(s_h) = \arg\min_{\beta(s_h)} \sum_i w(s_i, s_h)(y(s_i) - x(s_i)^T \beta(s_h))^2$$

donde $y(s_i)$ y $x(s_i)$ son la respuesta y los regresores en locación (s_i) respectivamente, además $w(s_i, s_h)$ es una función que decae con la distancia entre s_i y s_h, por ejemplo, podemos tomar $w(s_i, s_h) = \exp(-\frac{1}{2}||s_i - s_h||_2^2)$ (Jiang, Z 2019)

Para n datos, considerando la notación matricial antes usada, tomando $W = (w(s_i, s_h))_{i,h}^n$ como matriz de pesos, y manteniendo implícita la componente espacial, escribimos la calibración de β_h vía mínimos cuadrados ponderados como

$$\beta(s_h) = \arg\min_{\beta(s_h)} \sum_i w_{ih}(Y_i - \sum_{i=1}^{n} X_{ik}\beta_{ih})^2$$

Podemos calibrar los parametros haciendo uso de los paquetes *GWmodel* y *spgwr*, el modelo se basa en el concepto de ancho de banda de kernel para computar un modelo de regresión local en cada ubicación, el cual se basa en un tipo de kernel, fijo o adaptativo, y asignandole una estructura a w_{ij} que decaiga con la distancia.

Calibración: Geographically Weighted Regression

Para las simulaciones mediante *GWModel*, usamos un kernel adaptativo y para la función de pesos, se calcula una función de decaimiento 'bisquare' de la forma

$$w_{ij} = 1 - (\frac{d_{ij}^2}{h^2})^2$$

donde d_{ij} es la distancia entre las observaciones en locación s_i y sus vecinos en locación s_j, como tomamos un kernel de tipo adaptativo, el valor de h varía, y tomará la distancia entre la ubicación s_i y el vecino más lejano a dicha ubicación.
Por lo demás, *spgwr* trae una configuración estandarizada

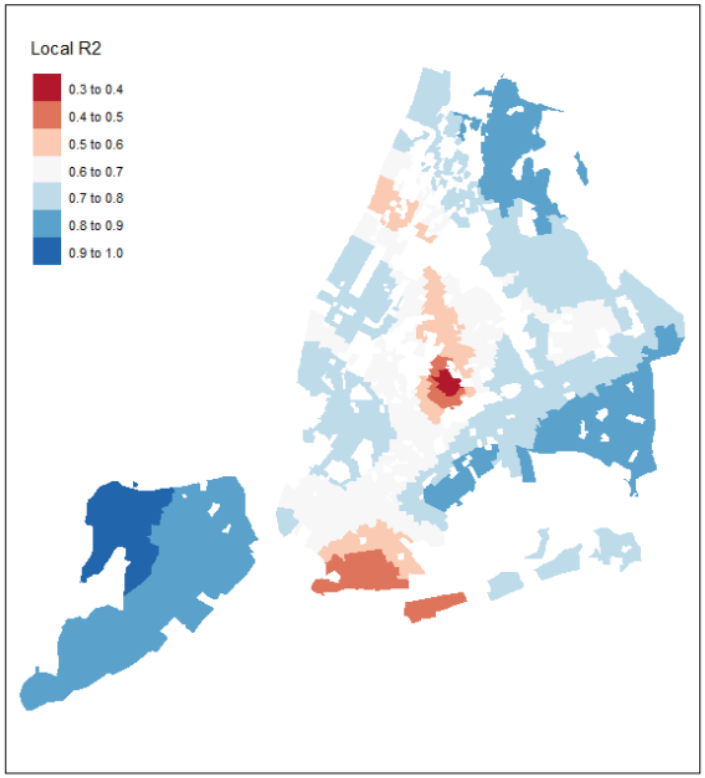
Calibración: Geographically Weighted Regression

```
*********************************************************************************
*          Results of Geographically Weighted Regression                        *
*********************************************************************************

**********************Model calibration information**********************
Kernel function: bisquare
Adaptive bandwidth: 228 (number of nearest neighbours)
Regression points: the same locations as observations are used.
Distance metric: Euclidean distance metric is used.

****************Summary of GWR coefficient estimates:******************
                          Min.     1st Qu.      Median     3rd Qu.       Max.
Intercept         -1.3915e+02  3.5281e+02  5.2017e+02  6.9042e+02 1107.2631
median_rooms      -5.6062e+01  6.3320e+01  8.8646e+01  1.1650e+02  225.2997
median_income     -1.3865e-03  1.0219e-04  5.7385e-04  1.2785e-03    0.0036
pct_college       -4.5722e+00 -9.8358e-01  5.5569e-01  2.2916e+00    5.4054
pct_foreign_born -7.8207e+00 -2.1959e+00 -1.3727e-01  1.8348e+00    5.0736
pct_white         -4.0832e+00  2.1527e-01  1.2922e+00  2.4962e+00    8.8601
pct_black         -8.7637e+00 -1.4388e+00 -3.5948e-01  6.4875e-01    4.1924
pct_hispanic      -3.1966e-02 -1.0257e-02 -5.3383e-03  6.4731e-04    0.0248
pct_asian         -3.8100e+00  3.5358e-01  1.1029e+00  2.3325e+00   15.3143
median_age        -1.1680e+01 -3.6273e+00 -8.2516e-01  2.0362e+00    7.3398
percent_ooh       -6.7411e+00 -4.2751e+00 -3.1605e+00 -2.3080e+00    1.3729
pop_density       -1.1661e-02 -3.3134e-03 -1.1423e-03  4.5817e-04    0.0088
************************Diagnostic information*************************
Number of data points: 1969
Effective number of parameters (2trace(S) - trace(S'S)): 303.8232
Effective degrees of freedom (n-2trace(S) + trace(S'S)): 1665.177
AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): 24436.11
AIC (GWR book, Fotheringham, et al. 2002,GWR p. 96, eq. 4.22): 24140.34
BIC (GWR book, Fotheringham, et al. 2002,GWR p. 61, eq. 2.34): 23694.62
Residual sum of squares: 21641574
R-square value:  0.7213625
Adjusted R-square value:  0.6704926
```

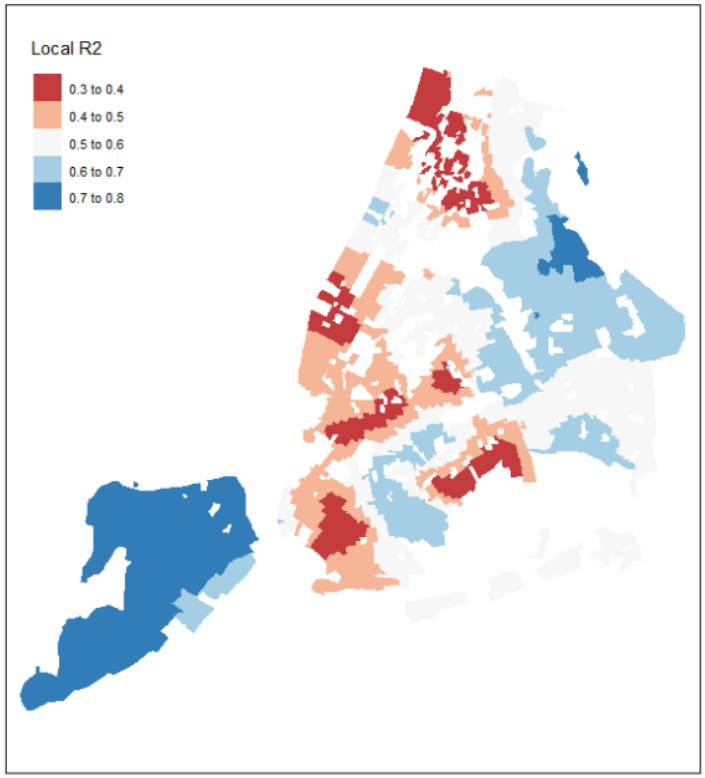
Visualización: Geographically Weighted Random Forest

Debido a que los coeficientes calibrados tienen asociados una posición geográfica, podemos observar una medida local del R^2 respecto del desempeño del modelo por distritos, y poder visualizar las zonas en donde predice mejor el modelo.

GWR [GWModel] Local R2



GWR [spgwr] Local R2



Geographically Weighted Random Forest

Implementamos el siguiente modelo aborda la heterogeneidad espacial mediante un conjunto de modelos de bosques aleatorios (RF) localmente calibrados

La ecuación para la calibración se formula como

$$Y_i = a(s_i)x(s_i) + \varepsilon$$

Donde s_i referencia la posición geográfica y $x(s_i)$ son los datos de entrenamiento

La diferencia entre un GWR, con una formulación lineal y un GRF, es que podemos modelar la no-estacionariedad de los datos junto con un modelo no lineal, lo que vuelve menos restríctiva la función que queremos modelar.

Para entrenar el modelo hacemos uso del paquete *SpatialML*.

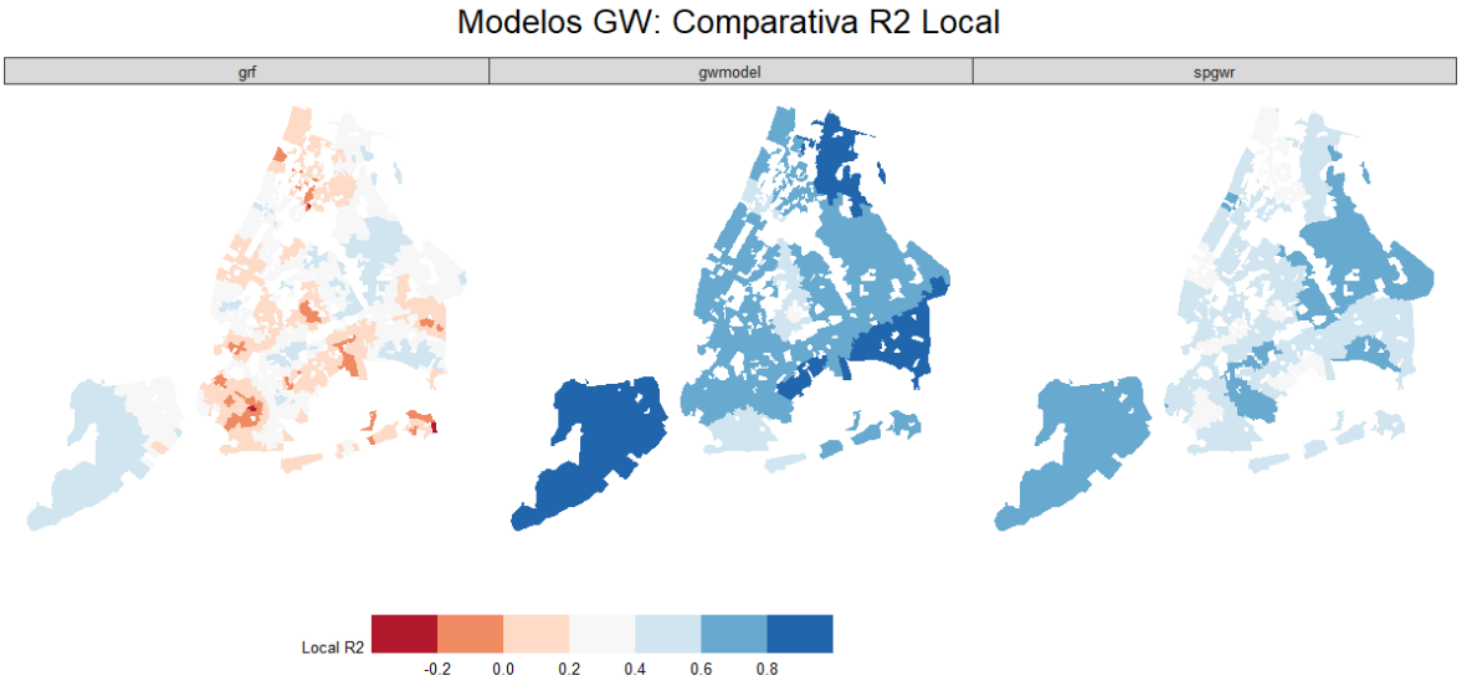
Calibración: Geographically Weighted Random Forest

Al entrenar el modelo, el paquete entrega el siguiente resumen.

```
Call:
 ranger(formula_grf, data = nyc_grf_prepped, num.trees = 500,      mtry = 2, importance =
"impurity", num.threads = NULL)

Type:                             Regression
Number of trees:                  500
Sample size:                      1969
Number of independent variables:  11
Mtry:                             2
Target node size:                 5
Variable importance mode:         impurity
Splitrule:                        variance
OOB prediction error (MSE):       17707.03
R squared (OOB):                  0.5513357
    median_rooms    median_income      pct_college pct_foreign_born        pct_white
pct_black
         4606936         10389873         12857798          6507840          8863042
7202205
    pct_hispanic        pct_asian       median_age      percent_ooh      pop_density
         3932309          4609792          4978187          5044124          5189713
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-757.5622  -54.9174   -0.1554   -3.0302   54.9276  596.4478
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-146.36310   -4.93200    0.02363   -0.09956    4.74542   98.33541
```

Entrega una medida de $R^2 = 0.551$, que es notoriamente inferior a los modelos clásicos de regresión geográficamente ponderada.

Comparativa: Modelos GW R^2 Local



Podemos ver que el modelo *GWR* del paquete *GWmodel* tiene, con diferencia, las puntuaciones R^2 locales más altas, seguidas por el paquete *spgwr* (GWR) y finalmente *SpatialML* (GRF)

Conclusiones

- Dadas las problemáticas asociadas al análisis de datos georreferenciados, tales como la autocorrelación espacial y la heterogeneidad espacial inmersa en los datos debido a su naturaleza demográfica, se reformula una serie de modelos para tratar este problema.
- Comparativamente, se nota una mejoría con respecto a la precisión de modelos de regresión clásicos. Estos resultados pueden ser útiles para diversos ámbitos de la industria, en cuyas tomas de decisiones intervengan datos que están asociados a una componente geográfica, como es el caso de la predicción de precios de vivienda dentro del mercado inmobiliario.

Gracias por su atención

Referencias

- Chen, Yanguang. 2013. "New Approaches for Calculating Moran's Index of Spatial Autocorrelation." Edited by Guy J.-P. Schumann. PLoS ONE 8 (7): e68336. https://doi.org/10.1371/journal.pone.0068336.
- 2021. "An Analytical Process of Spatial Autocorrelation Functions Based on Moran's Index." Edited by Bailang Yu. PLOS ONE 16 (4): e0249589. https://doi.org/10.1371/journal.pone.0249589.
- 2022. "Deriving Two Sets of Bounds of Moran's Index by Conditional Extremum Method." https://doi.org/10.48550/ARXIV.2209.08562.

Referencias

- Georganos, Stefanos, and Stamatis Kalogirou. 2022. "A Forest of Forests: A Spatially Weighted and Computationally Efficient Formulation of Geographical Random Forests." ISPRS International Journal of Geo-Information 11 (9): 471. https://doi.org/10.3390/ijgi11090471.
- Ikramov, Khakim D. 1994. "A Simple Proof of the Generalized Schur Inequality." Linear Algebra and Its Applications 199 (March): 143–49. https://doi.org/10.1016/0024-3795(94)90346-8.
- Jiang, Zhe. 2019. "A Survey on Spatial Prediction Methods." IEEETransactions on Knowledge and Data Engineering 31 (9): 1645–64. https://doi.org/10.1109/TKDE.2018.2866809.

Referencias

- Nikparvar, Behnam, and Jean-Claude Thill. 2021. "Machine Learning of Spatial Data." ISPRS International Journal of Geo-Information 10 (9):600. https://doi.org/10.3390/ijgi10090600.
- Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." Economic Geography 46 (June): 234. https://doi.org/10.2307/143141.
- Walker, Kyle. 2023. "Analyzing US Census Data," January. https://doi.org/10.1201/9780203711415.