UNIVERSIDAD TORCUATO DI TELLA

Trabajo Final

Instrumentos de análisis territorial

Maestría en políticas públicas

Universidad Torcuato Di Tella

Profesor: Martín Montané

Alumnos: Andrés Gibson y Clara Rodríguez

Junio 2020

Introducción

El brote de la enfermedad causada por el coronavirus “COVID-19” ha registrado una rápida propagación a escala comunitaria, regional e internacional, con un aumento exponencial del número de casos y muertes. El total de casos confirmados en Argentina al 1 de junio es de 17.415 y en la Ciudad de Buenos Aires, una de las más afectadas, la cifra acumulada es de 8.480 casos.

El presente documento toma una base de datos de uso privado generada desde el comienzo de las medidas de prevención gubernamentales ante la pandemia, el 12 de abril, hasta el 1 de junio. Esta cuenta con un muestreo de 2.395 pacientes contagiados y amplia información sobre los mismos, entre ellas, sus domicilios.

La pregunta que intentamos indagar es si un contexto socioeconómico y habitacional desfavorable incide en la probabilidad de contagio de coronavirus.

Preparación de los datos

Antes de comenzar a trabajar el archivo en R, se borró la identidad de los pacientes para proteger su privacidad y se les asignó un número de ID.

Muchos de los casos provienen de barrios vulnerables, por lo que las direcciones no cuentan con el formato tradicional de calle y altura, sino que tienen una nomenclatura particular (barrio -> manzana -> casa). Para poder georreferenciar dichas direcciones fue preciso utilizar el paquete “RUMBA” (https://github.com/bitsandbricks/RUMBA). Este posee dos tipos de códigos que se ajustan a la necesidad de nuestra base.

A pesar de correr el código para ambos tipos de dirección, un gran porcentaje de estas no pudo ser geolocalizada por lo que fue necesario adaptarlas manualmente a un formato que cumpla con los requerimientos del paquete. Una vez terminado, se realizó el mismo proceso en RUMBA para georreferenciar los datos restantes.

Como resultado se obtuvieron cuatro bases de datos, las que se unieron para conformar el archivo final con el que comenzar el análisis.

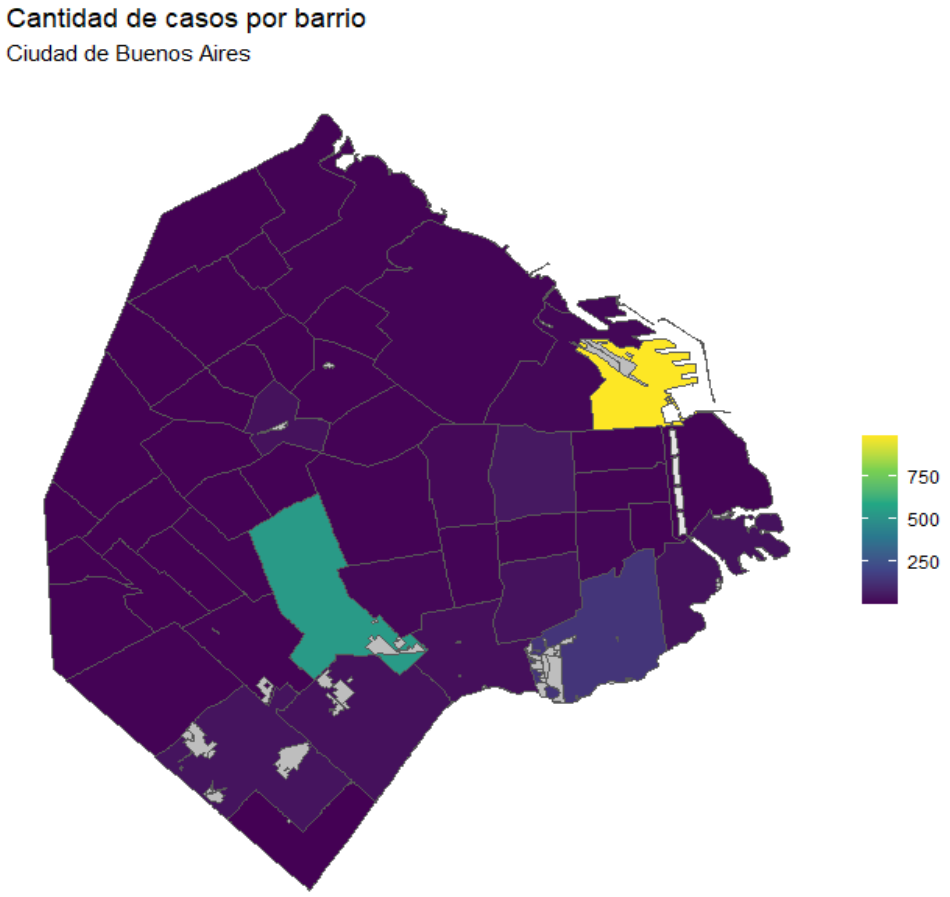
- Geolocalizados: 2157 casos.
- Restantes: 238 casos.

Análisis

Inicialmente, se procedió a convertir el dataframe a sf y obtener archivos espaciales necesarios para generar un primer mapa que brinde una idea aproximada de la distribución de casos en la ciudad. Estos archivos son:

- GeoJSON con los 48 barrios de CABA (https://data.buenosaires.gob.ar/dataset/barrios).
- Shapefile con los polígonos de barrios vulnerables de Argentina (https://datos.gob.ar/dataset/otros-barrios-populares-argentina).

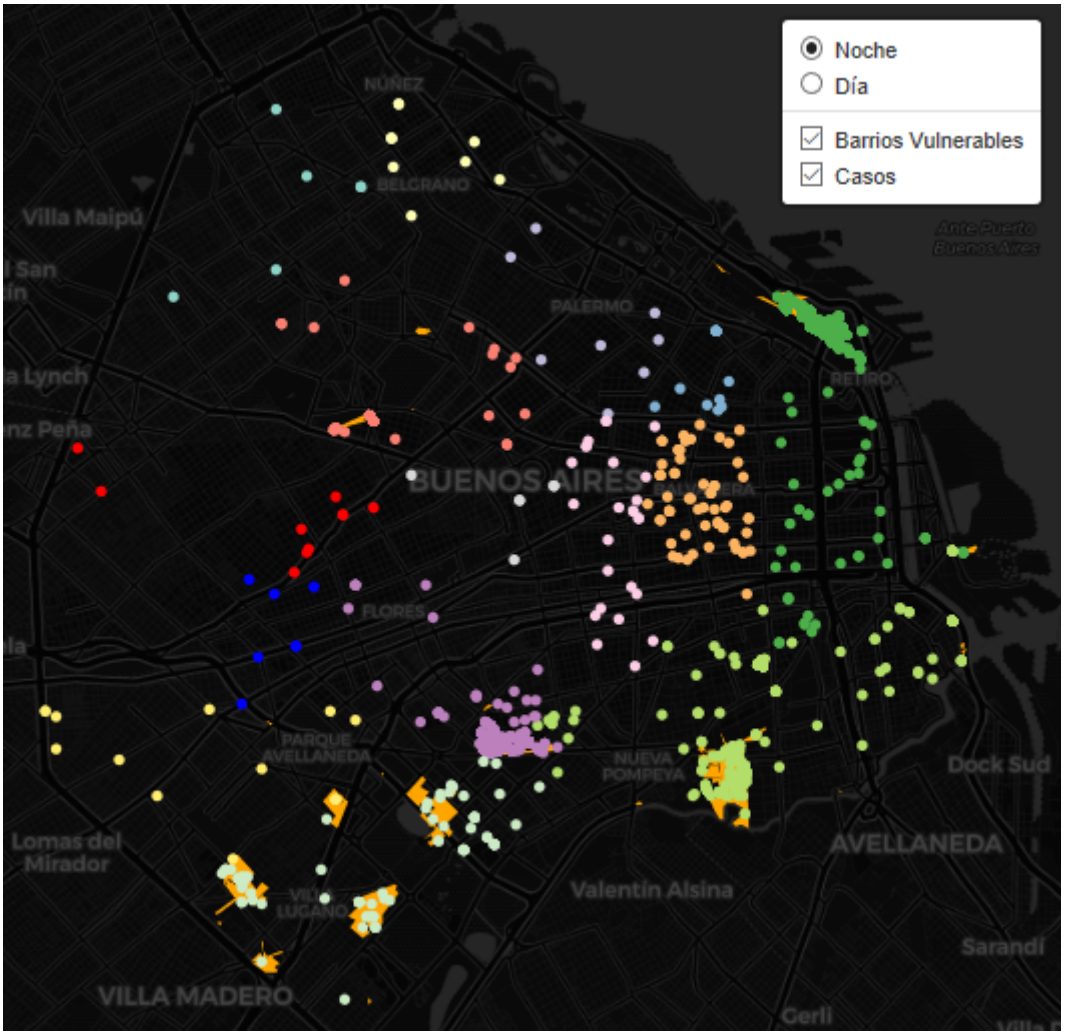
Se filtró el shapefile para obtener solamente los barrios vulnerables de la ciudad y se unieron los dataframes de los barrios de CABA con el archivo final de casos en sf. Esto permite saber dentro de qué polígono de barrio se encuentra cada una de nuestras observaciones. Se agruparon y sumaron la cantidad de casos por barrio, y se confeccionó el siguiente mapa con ggplot:



Podemos observar en el mismo que los barrios más afectados son: Retiro (996), Flores (540) y Barracas (146). Esto se debe a que dentro de estos se encuentran los barrios vulnerables (pintados en gris), que aportan gran número de casos. Flores alberga la villa 1-11-14, Retiro la villa 31 y Barracas la villa 21-24.

Posteriormente, se confeccionó un mapeo de casos por puntos con Leaflet (se incluye una captura de pantalla para facilitar el análisis, pero se recomienda abrir el archivo adjunto enviado desde el navegador, dado que es interactivo). Este contiene los casos coloreados según la comuna a la que

pertenecen y los barrios vulnerables en amarillo, los cuales se pueden activar y desactivar haciendo clic en el recuadro superior derecho. También es posible visualizar los datos en formato día o noche, según la preferencia del usuario, y hacer zoom.



En este mapa podemos visualizar nuevamente que los casos se acumulan en los barrios vulnerables. Es necesario aclarar que muchos de los puntos graficados representan muchas observaciones solapadas, dado que es habitual que los pacientes solamente declaren el barrio donde viven sin hacer referencia a la manzana o casa. En estos casos, el paquete RUMBA asigna un mismo punto a todas las observaciones que digan, por ejemplo, “Villa 31”. Por otro lado, la precisión máxima alcanzada por RUMBA para barrios vulnerables es a nivel manzana.

Para concluir el trabajo en R, se descargó una base de datos que contiene los polígonos de los radios censales de la ciudad, correspondiente al censo del año 2010 (https://data.buenosaires.gob.ar/dataset/informacion-censal-por-radio). Esta

cuenta con muchas variables, siendo la más relevante para nuestro análisis aquella que indica la cantidad de hogares con necesidades básicas insatisfechas (NBI). Se unió esta base a la base de casos y, además, se crearon tres variables que calculan la densidad poblacional, los habitantes promedio por hogar y la cantidad de casos dentro de cada radio censal. Estos datos permiten una mejor comprensión de la situación de vulnerabilidad de cada radio.

A partir de este punto, se continuó el análisis en GeoDa, utilizando el archivo anteriormente mencionado (guardado en formato GeoJSON.

GeoDa

Una vez abierto el archivo GeoJSON en GeoDa, creamos una matriz de pesos que resume el valor y peso de los radios que limitan con el radio de análisis. Esto se realiza usando Queen contiguity de orden 1 (se considera vecino de un radio a cualquier otro radio que lo toque al menos en un punto) para comenzar el análisis.

El objetivo inicial es entender cómo las variables de interés se relacionan entre sí. Para esto, se realizó una regresión lineal usando la matriz de pesos. Como variable dependiente se utilizó la cantidad de casos por radio censal y como variables explicativas los hogares con NBI, la cantidad promedio de personas por hogar y la densidad poblacional.

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  COVID_Radios
Dependent Variable  :       Casos  Number of Observations: 3552
Mean dependent var  :     1.50563  Number of Variables   :    4
S.D. dependent var  :     8.05827  Degrees of Freedom    : 3548

R-squared           :    0.036612  F-statistic           :      44.945
Adjusted R-squared  :    0.035797  Prob(F-statistic)     : 1.68119e-28
Sum squared residual:      222207  Log likelihood        :    -12385.8
Sigma-square        :     62.6289  Akaike info criterion :     24779.6
S.E. of regression  :     7.91384  Schwarz criterion     :     24804.3
Sigma-square ML     :     62.5584
S.E of regression ML:     7.90938

------------------------------------------------------------------------------
      Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
         CONSTANT      -2.22758       0.699501       -3.18453     0.00146
       HOGARES_NBI     0.0361168     0.00424349        8.51112     0.00000
         Hab_Hogar      0.773161       0.244744        3.15906     0.00160
          Densidad    3.8931e-05    7.22427e-06        5.38892     0.00000
------------------------------------------------------------------------------
```

A partir del resultado, se observa que las tres independientes se relacionan positivamente con la variable dependiente, por lo que un aumento en cualquiera de ellas, genera un aumento en la cantidad de casos, y que además son significativas a la hora de explicar la relación. Puntualmente para cada variable podemos interpretar las siguientes conclusiones:

- Si se sumara un hogar más con necesidades básicas insatisfechas, esto generaría un aumento de 0.036 en la cantidad de casos.
- Un habitante más por hogar generaría un incremento de 0.77 casos.
- Un aumento de 1 punto en la densidad del radio generaría un aumento de 0,39 casos.

Es importante aclarar que esta primera regresión no tiene en cuenta aspectos espaciales a la hora de ver la relación entre las variables. Existen dos modelos que permiten mejorar las estimaciones, el Spatial Autoregressive Model (SAR) y el Spatial Error Model (SEM). El primero asume que existe interacción entre las unidades de análisis próximas entre sí, por lo que lo que sucede con una, afecta a las demás y viceversa. El segundo asume que existen variables relevantes que se podrían usar para controlar que no se pudieron medir, por lo que agrega los residuos de la regresión a la regresión.

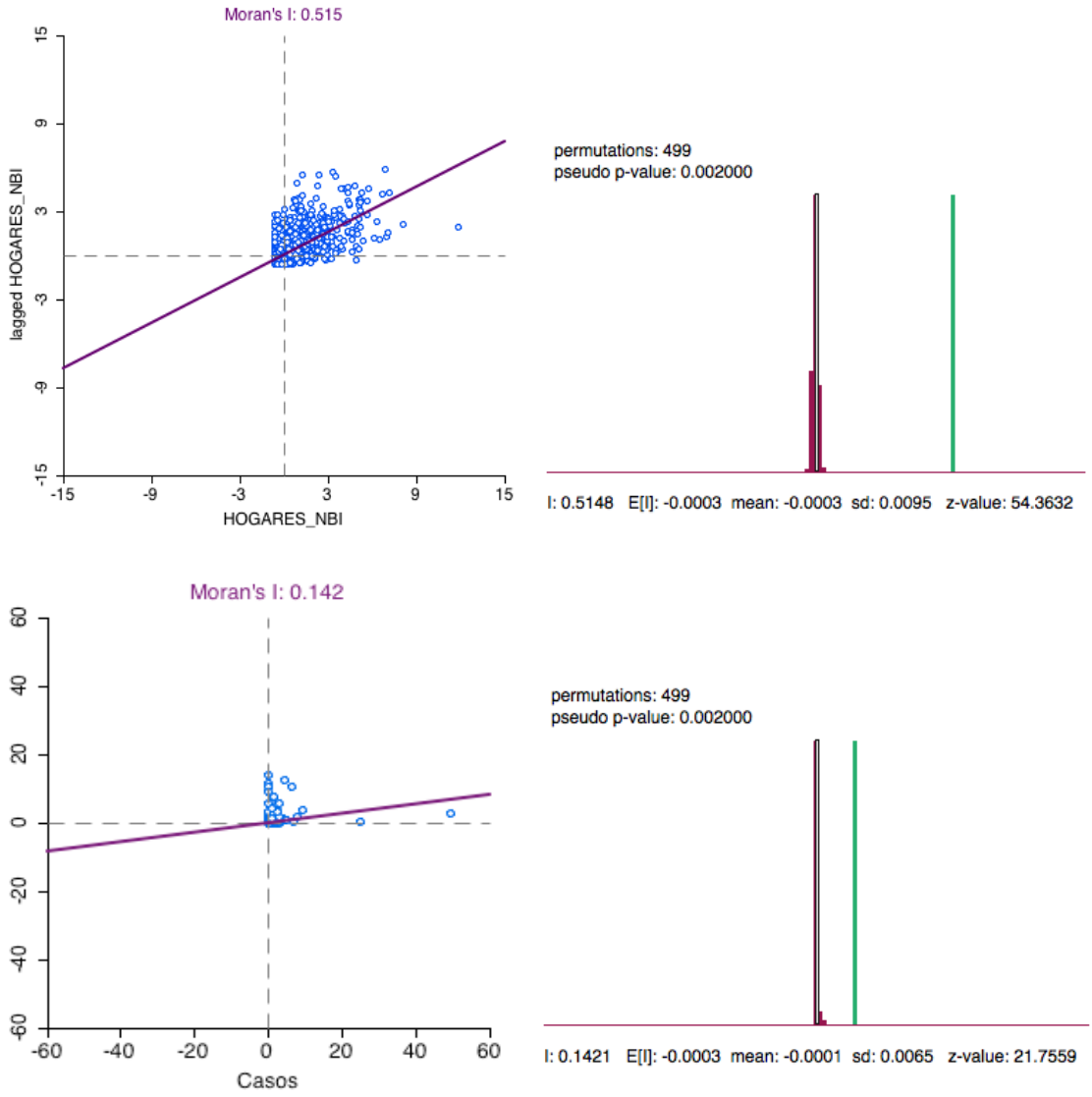
En nuestro caso, desde el punto de vista conceptual, el modelo más apropiado a utilizar es el SAR, ya que la enfermedad se nutre de la interacción entre personas y es altamente contagiosa. Por otro lado, también es posible constatar que este modelo se ajusta mejor a los datos dado que en los criterios Akaike y Schwarz se observan valores más bajos que en la regresión lineal y el modelo SEM (24704,4 y 24729,1 respectivamente para SEM).

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : COVID_Radios
Spatial Weight      : COVID_Radios_weight
Dependent Variable  :       Casos  Number of Observations: 3552
Mean dependent var  :     1.50563  Number of Variables   :    5
S.D. dependent var  :     8.05827  Degrees of Freedom    : 3547
Lag coeff.   (Rho)  :    0.217819

R-squared           :    0.066684  Log likelihood        :   -12343.1
Sq. Correlation     : -            Akaike info criterion :    24696.3
Sigma-square        :     60.6056  Schwarz criterion     :    24727.1
S.E of regression   :     7.78496

------------------------------------------------------------------------------
      Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
          W_Casos      0.217819      0.0267504       8.14264      0.00000
         CONSTANT     -1.23526       0.688146       -1.79506      0.07264
       HOGARES_NBI     0.030812     0.00421211       7.31511      0.00000
         Hab_Hogar     0.443661       0.240916       1.84156      0.06554
          Densidad   2.4207e-05    7.12295e-06       3.39845      0.00068
------------------------------------------------------------------------------
```

Los resultados del SAR minimizan un poco la relación vista en la primera regresión:

- Si se sumara un hogar más con necesidades básicas insatisfechas, esto generaría un aumento de 0.031 en la cantidad de casos.
- Un habitante más por hogar generaría un incremento de 0.44 casos (esta variable es menos significativa que antes).
- Un aumento de 1 punto en la densidad del radio generaría un aumento de 0,24 casos.

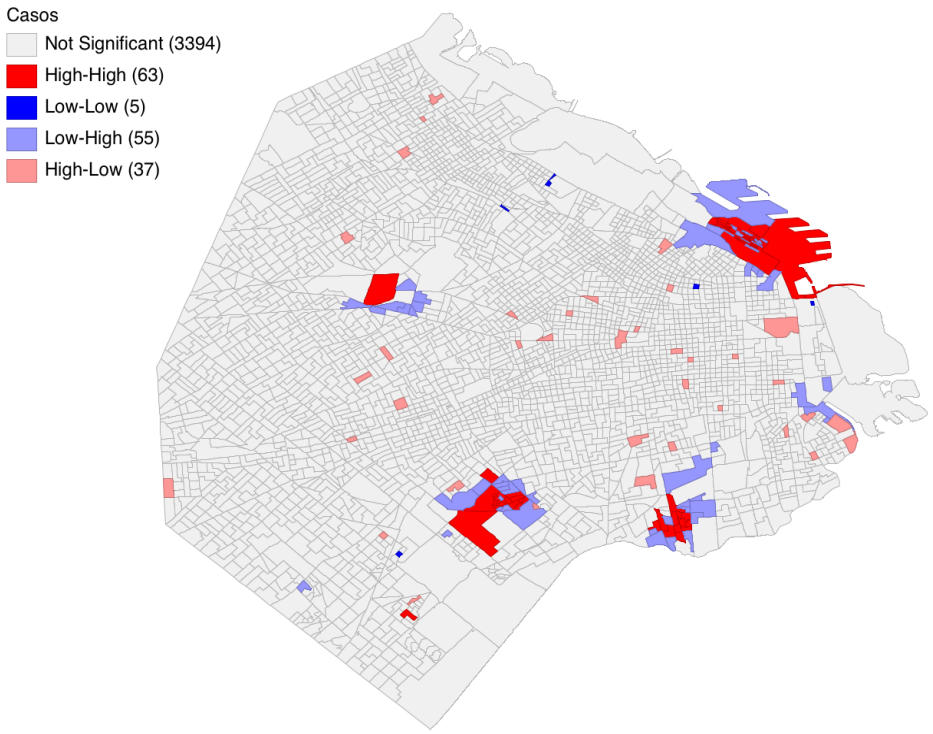
Otro test a realizar es el Test local de Moran “I”, que permite saber si hay relación espacial entre los radios. Para esto, se tomaron como referencia las variables hogares con NBI y casos por radio. Los resultados observados son los siguientes:



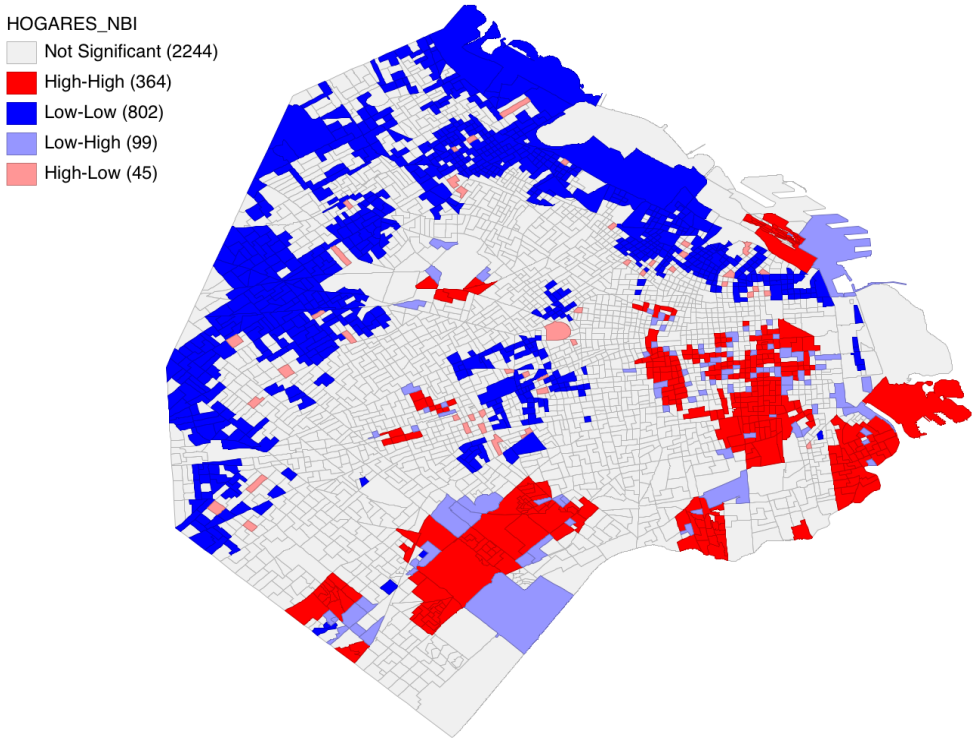
Para ambas variables existe una relación positiva entre las observaciones y el promedio de las observaciones cercanas. Esto se hace más notorio en el caso de los hogares con NBI, ya que la cantidad de casos existen dos outliers que inclinan la curva (estos corresponden a radios dentro de la villa 31 y la villa 1-11-14). Además, gracias a los gráficos de la derecha, podemos confirmar que

existe autocorrelación espacial, ya que se randomizaron los datos 499 veces (cambiar los valores de los vecinos por los de vecinos en otras áreas de la ciudad) y el resultado obtenido la primera vez (marcado en verde) es único respecto al resultado del resto, concentrados a su izquierda.

Estos datos también pueden ser mostrados espacialmente a través de cluster maps:

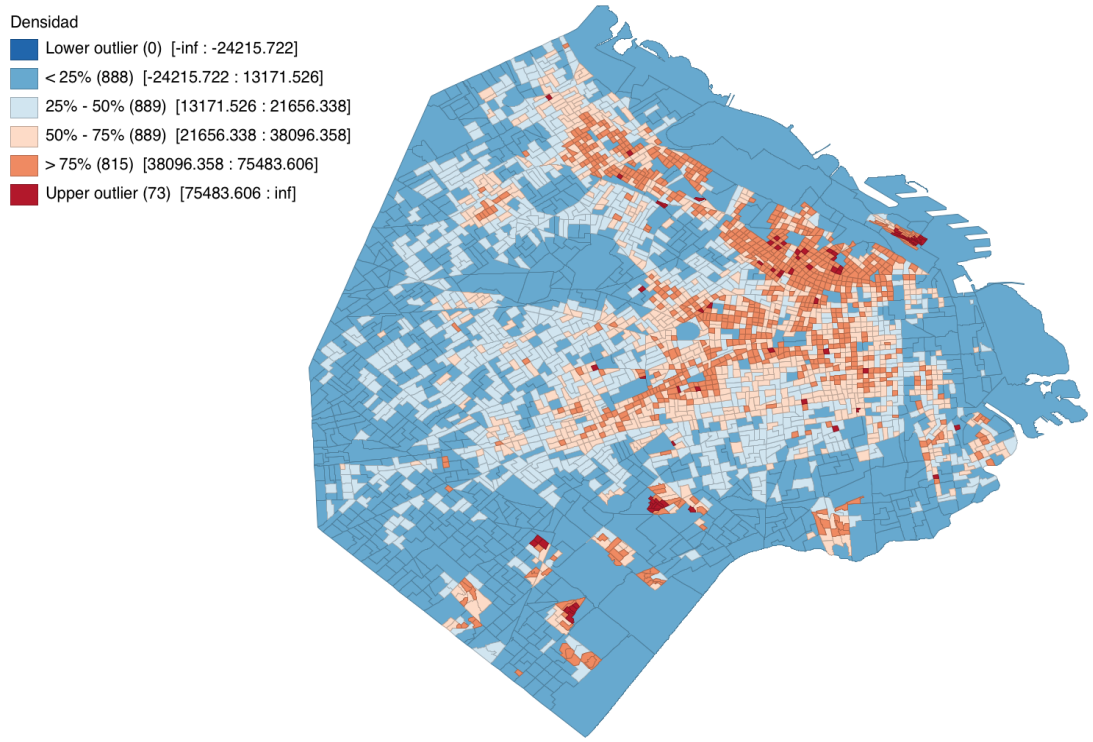


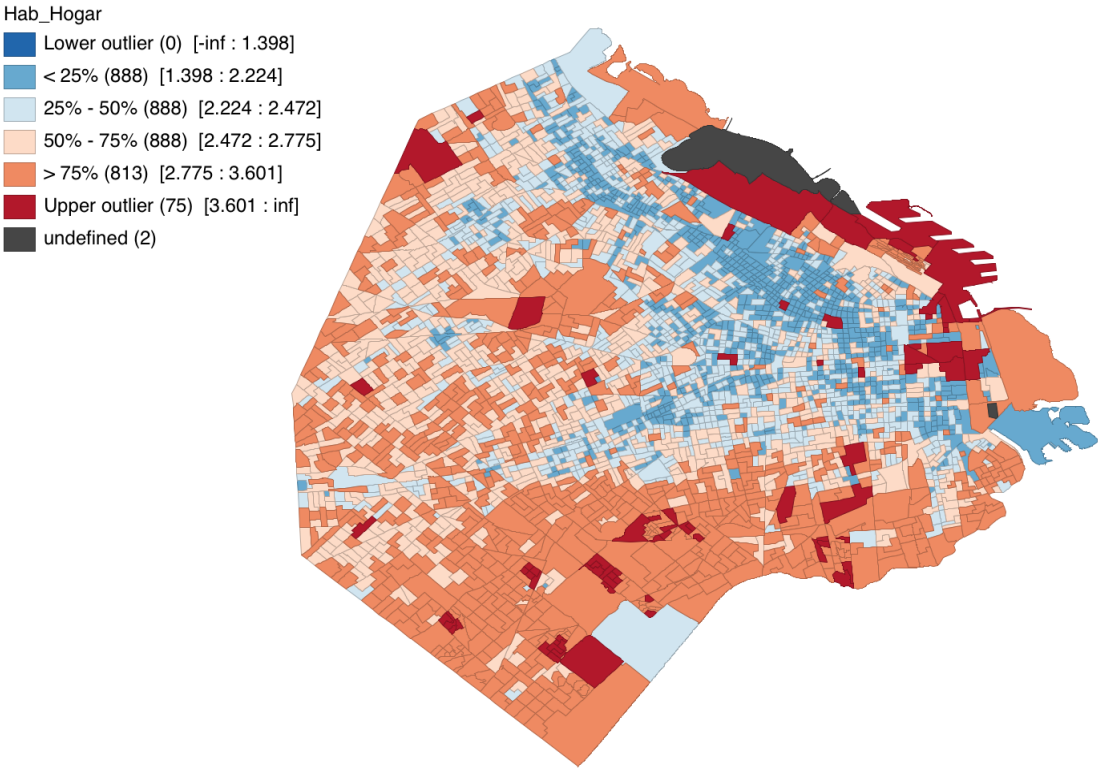
En este mapa vemos en rojo los radios que presentan un alto número de casos y están rodeados de radios con la misma característica. En azul se muestra lo mismo, pero para números bajos de casos. Las cuatro áreas en rojo pertenecen, además, a villas de la ciudad (31, 1-11-14, 21-24 y Carbonilla).



En este mapa se observan en rojo aquellos radios con muchos hogares con NBI y se encuentran rodeados de radios que presentan la misma característica. En azul se muestra lo mismo, pero para números bajos de hogares con NBI. En este caso también se observa una relación entre altos niveles de hogares con NBI y los barrios vulnerables, y la desigualdad norte-sur en la ciudad.

Por último, se muestran dos boxmaps con la densidad poblacional en la ciudad y los habitantes promedio por hogar. Si bien por sí solas estas dos variables no son sinónimo de vulnerabilidad, valores altos en ambas contribuyen a la propagación de la enfermedad y son características observadas en zonas carenciadas.


Densidad
Lower outlier (0) [-inf : -24215.722]
< 25% (888) [-24215.722 : 13171.526]
25% - 50% (889) [13171.526 : 21656.338]
50% - 75% (889) [21656.338 : 38096.358]
> 75% (815) [38096.358 : 75483.606]
Upper outlier (73) [75483.606 : inf]


Hab_Hogar
Lower outlier (0) [-inf : 1.398]
< 25% (888) [1.398 : 2.224]
25% - 50% (888) [2.224 : 2.472]
50% - 75% (888) [2.472 : 2.775]
> 75% (813) [2.775 : 3.601]
Upper outlier (75) [3.601 : inf]
undefined (2)

Conclusiones

Luego de los análisis realizados en base al muestreo de datos de infectados por coronavirus en la ciudad, **es posible afirmar que existe una relación significativa entre habitar en un contexto socioeconómico y habitacional desfavorable y la probabilidad de contagio de coronavirus.**

A través de la cantidad de hogares con necesidades básicas insatisfechas, se logró observar cuales son las zonas más vulnerables de la ciudad. Esta variable parece incidir significativamente en la cantidad de casos observados por radio censal y muestra valores altos en los barrios considerados vulnerables. Lo mismo se realizó con los habitantes promedio por hogar y la densidad poblacional, que no necesariamente indican vulnerabilidad, pero sí son variables que mostraron altos valores en zonas carenciadas y también inciden significativamente en la cantidad de casos.

Introduction

The outbreak of the disease caused by the coronavirus "COVID-19" has rapidly spread at the community, regional, and international levels, with an exponential increase in the number of cases and deaths. As of June 1st, Argentina has reported a total of 17,415 confirmed cases, with 8,480 cases accumulated in the City of Buenos Aires, one of the most affected regions.

This document uses a privately generated database from the beginning of the government's preventive measures against the pandemic, starting from April 12th until June 1st. The database includes a sample of 2,395 infected patients and comprehensive information, including their addresses.

The main question this study aims to investigate is whether an unfavorable socioeconomic and housing context influences the likelihood of contracting the coronavirus.

Data Preparation

Before working with the data in R, patient identities were removed to protect their privacy, and each was assigned an ID number. Many cases come from vulnerable neighborhoods, so their addresses do not follow the traditional street and house number format but have a particular nomenclature (neighborhood -> block -> house). To georeference these addresses, the "RUMBA" package was used (https://github.com/bitsandbricks/RUMBA). It offers two types of codes that fit the needs of our database. Despite running the code for both types of addresses, a significant percentage could not be geolocated, requiring manual adaptation to a format that complies with the package requirements. Once completed, the RUMBA package was used again to georeference the remaining data.

As a result, four databases were obtained, which were merged to form the final file for analysis:

Geolocated: 2,157 cases.

Remaining: 238 cases.

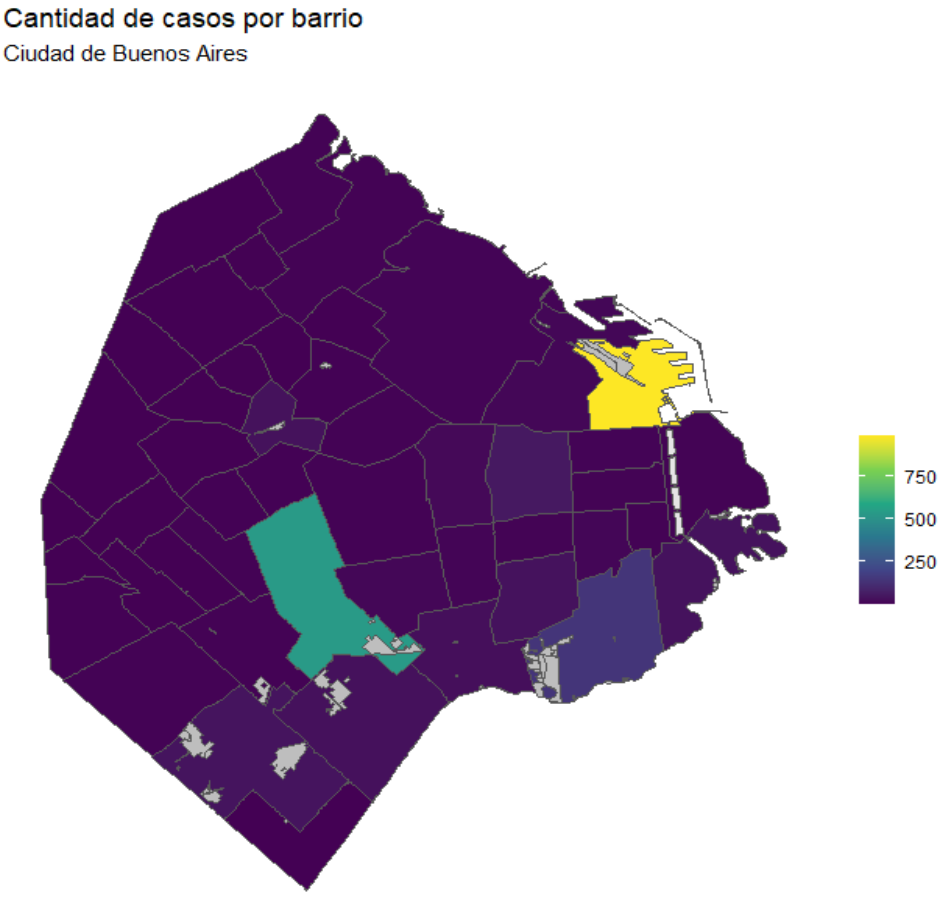
Analysis

Initially, the data frame was converted to an "sf" object, and spatial files needed for generating an initial map representing the distribution of cases in the city were obtained. These files include:

- GeoJSON with the 48 neighborhoods of CABA (https://data.buenosaires.gob.ar/dataset/barrios).
- Shapefile with polygons of vulnerable neighborhoods in Argentina (https://datos.gob.ar/dataset/otros-barrios-populares-argentina).

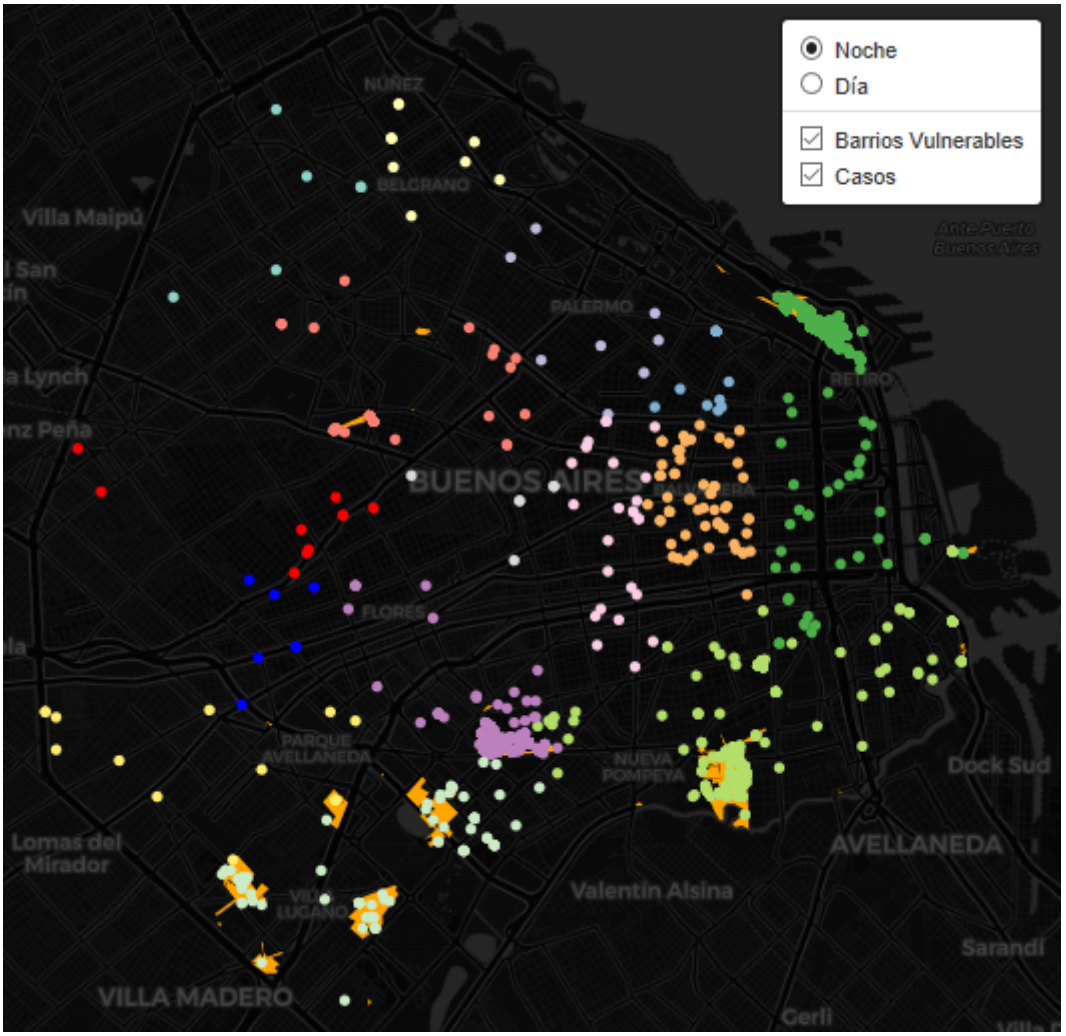
The shapefile was filtered to obtain only the vulnerable neighborhoods in the city, and the data frames of CABA neighborhoods were joined with the final case data in "sf" format. This allows us to know which neighborhood polygon

each observation falls within. The cases were grouped and summed by neighborhood, resulting in a map generated with ggplot.



The map shows that the most affected neighborhoods are Retiro (996 cases), Flores (540 cases), and Barracas (146 cases). This is due to these neighborhoods containing vulnerable areas (highlighted in gray) that contribute a significant number of cases. Flores is home to the "Villa 1-11-14," Retiro to "Villa 31," and Barracas to "Villa 21-24."

Subsequently, a point-based mapping of cases was created using Leaflet (a screenshot is provided for analysis, but it is recommended to open the interactive file attached in the browser). The map colors the cases based on the commune they belong to, with vulnerable neighborhoods highlighted in yellow. Users can activate or deactivate these layers by clicking on the top-right checkbox. The map also offers a choice between day and night mode and allows zooming.



This map reinforces the observation that cases accumulate in vulnerable neighborhoods. It is worth noting that many points on the map represent multiple overlapping observations because patients often only report their neighborhood without specifying the block or house. In such cases, the RUMBA package assigns the same point to all observations stating, for example, "Villa 31." Additionally, RUMBA's maximum precision for vulnerable neighborhoods is at the block level.

The analysis continued by downloading a database containing census tract polygons for the city, corresponding to the 2010 census (https://data.buenosaires.gob.ar/dataset/informacion-censal-por-radio). This database includes various variables, with the most relevant one for this analysis being the number of households with unmet basic needs (NBI). This database was joined with the case database, and three variables were created to calculate population density, average inhabitants per household, and the number of cases within each census tract. These data provide a better understanding of the vulnerability situation in each census tract.

The analysis was further continued in GeoDa, using the previously mentioned GeoJSON file.

GeoDa Analysis

Upon opening the GeoJSON file in GeoDa, a spatial weights matrix was created summarizing the value and weight of neighboring census tracts. Queen contiguity of order 1 was used for this analysis (considering any touching census tract as a neighbor). The initial objective was to understand how the variables of interest relate to each other. A linear regression was performed using the spatial weights matrix, with the number of cases per census tract as the dependent variable and the number of households with NBI, average inhabitants per household, and population density as explanatory variables.

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  COVID_Radios
Dependent Variable  :       Casos  Number of Observations: 3552
Mean dependent var  :     1.50563  Number of Variables   :    4
S.D. dependent var  :     8.05827  Degrees of Freedom    : 3548

R-squared           :    0.036612  F-statistic           :      44.945
Adjusted R-squared  :    0.035797  Prob(F-statistic)     : 1.68119e-28
Sum squared residual:      222207  Log likelihood        :    -12385.8
Sigma-square        :     62.6289  Akaike info criterion :     24779.6
S.E. of regression  :     7.91384  Schwarz criterion     :     24804.3
Sigma-square ML     :     62.5584
S.E of regression ML:     7.90938

------------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
          CONSTANT      -2.22758       0.699501       -3.18453     0.00146
       HOGARES_NBI     0.0361168      0.00424349        8.51112     0.00000
         Hab_Hogar      0.773161       0.244744        3.15906     0.00160
          Densidad    3.8931e-05    7.22427e-06        5.38892     0.00000
------------------------------------------------------------------------------
```

The results indicate that all three independent variables are positively related to the dependent variable, meaning that an increase in any of them leads to an increase in the number of cases. These relationships are also statistically significant. Specifically:

- Adding one more household with unmet basic needs would result in an increase of 0.036 in the number of cases.
- Adding one more inhabitant per household would result in an increase of 0.77 cases.
- An increase of 1 point in population density would lead to an increase of 0.39 cases.

It is essential to clarify that this initial regression does not account for spatial aspects in the relationship between variables. There are two models that can improve the estimates: the Spatial Autoregressive Model (SAR) and the

Spatial Error Model (SEM). The SAR assumes that there is interaction between nearby analysis units, meaning what happens in one unit affects others and vice versa. The SEM assumes that there are relevant variables that could be used to control for unmeasured factors, so it adds the residuals of the regression to the model.

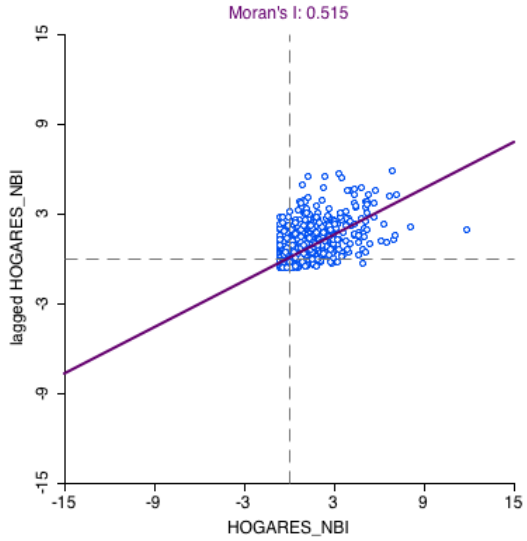
Conceptually, the SAR model is more appropriate for our study as the disease spreads through interactions between people and is highly contagious. Furthermore, this model fits the data better as indicated by the lower values in Akaike and Schwarz criteria compared to the linear regression and SEM (24704.4 and 24729.1, respectively, for SEM).

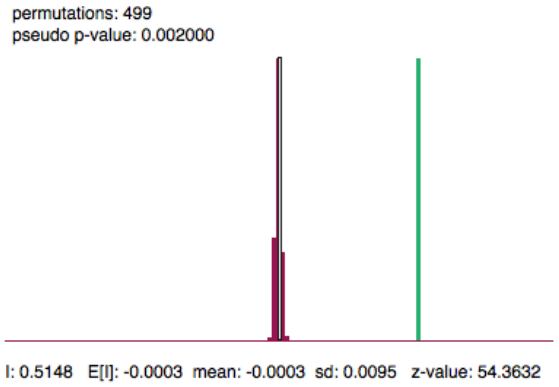
```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : COVID_Radios
Spatial Weight      : COVID_Radios_weight
Dependent Variable  :       Casos  Number of Observations: 3552
Mean dependent var  :     1.50563  Number of Variables   :    5
S.D. dependent var  :     8.05827  Degrees of Freedom    : 3547
Lag coeff.   (Rho)  :    0.217819

R-squared           :    0.066684  Log likelihood        :   -12343.1
Sq. Correlation     : -           Akaike info criterion :    24696.3
Sigma-square        :     60.6056  Schwarz criterion     :    24727.1
S.E of regression   :     7.78496

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value     Probability
------------------------------------------------------------------------------
          W_Casos      0.217819      0.0267504       8.14264      0.00000
         CONSTANT      -1.23526       0.688146      -1.79506      0.07264
       HOGARES_NBI     0.030812     0.00421211       7.31511      0.00000
         Hab_Hogar     0.443661       0.240916       1.84156      0.06554
          Densidad   2.4207e-05   7.12295e-06       3.39845      0.00068
------------------------------------------------------------------------------
```

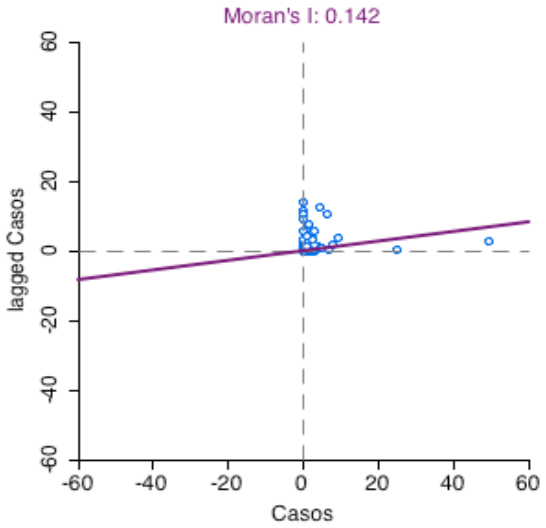
The results from the SAR model slightly attenuate the relationships observed in the initial regression:

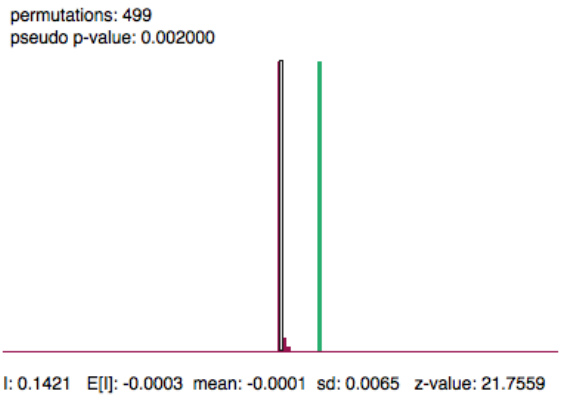
- Adding one more household with unmet basic needs would result in an increase of 0.031 in the number of cases.
- Adding one more inhabitant per household would result in an increase of 0.44 cases (this variable is less significant than before).
- An increase of 1 point in population density would lead to an increase of 0.24 cases.

Another test conducted is the Local Moran's I test, which helps determine if there is spatial autocorrelation between census tracts. The variables "number of households with NBI" and "number of cases per census tract" were used for this analysis.



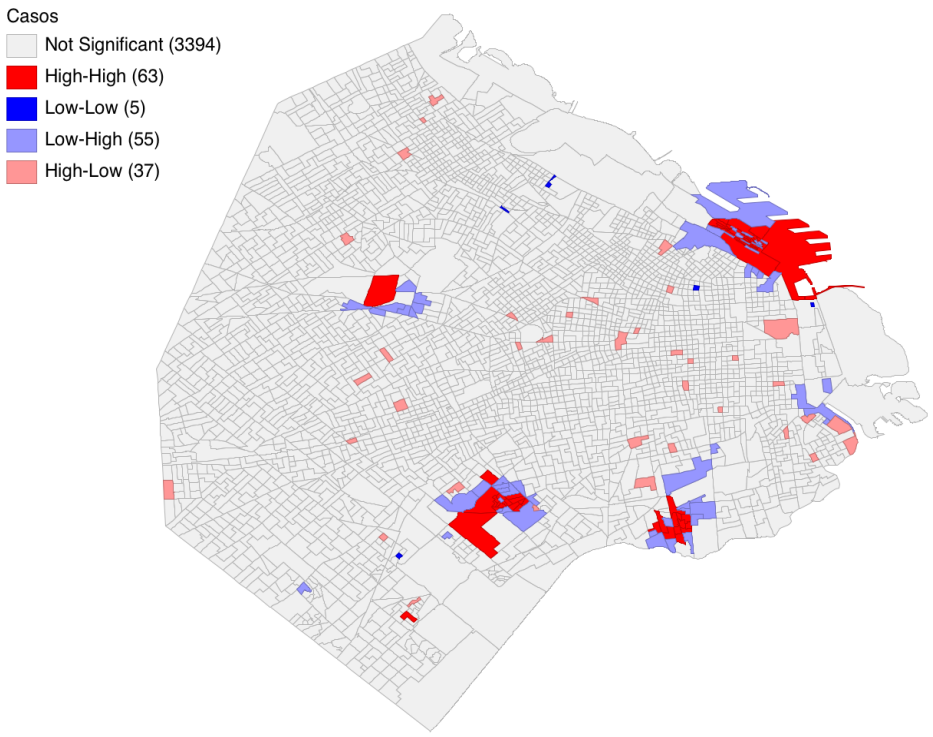






The results indicate a positive relationship between observations and the average of neighboring observations for both variables. This relationship is more noticeable for the variable "number of households with NBI," where two outliers in the curve correspond to census tracts within "Villa 31" and "Villa 1-11-14." Additionally, the graphs on the right confirm the existence of spatial autocorrelation, as the data was randomized 499 times (changing neighbor values with values from neighbors in other areas of the city), and the initial result (marked in green) is unique compared to the rest of the results, concentrated to its left.

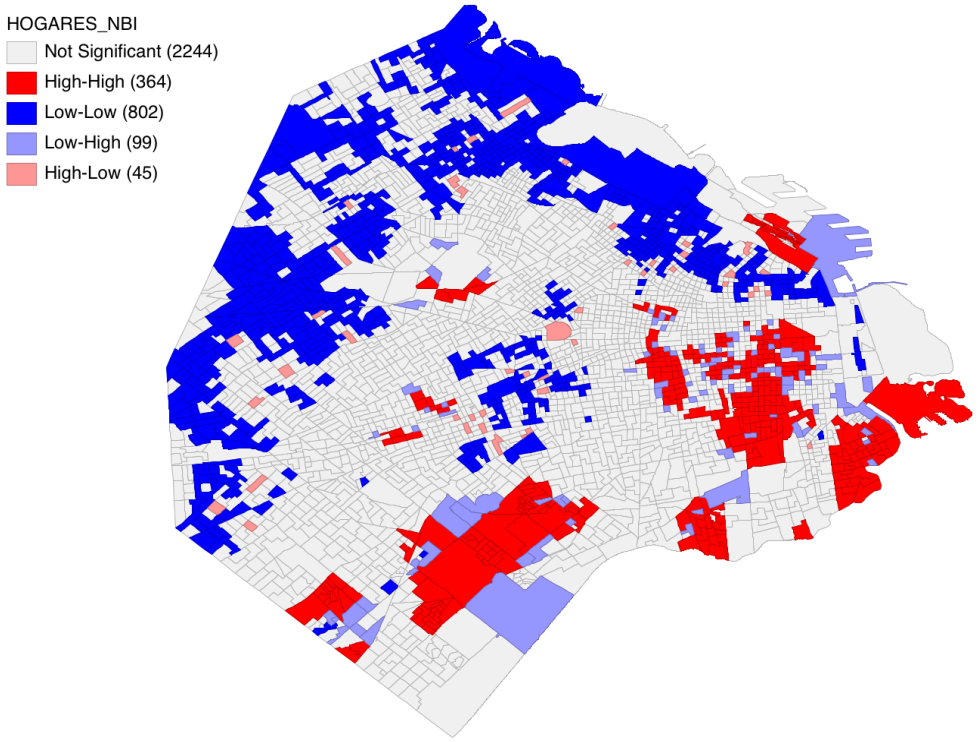
These data can also be displayed spatially through cluster maps:



The red areas represent census tracts with a high number of cases, surrounded by census tracts with a similar characteristic.

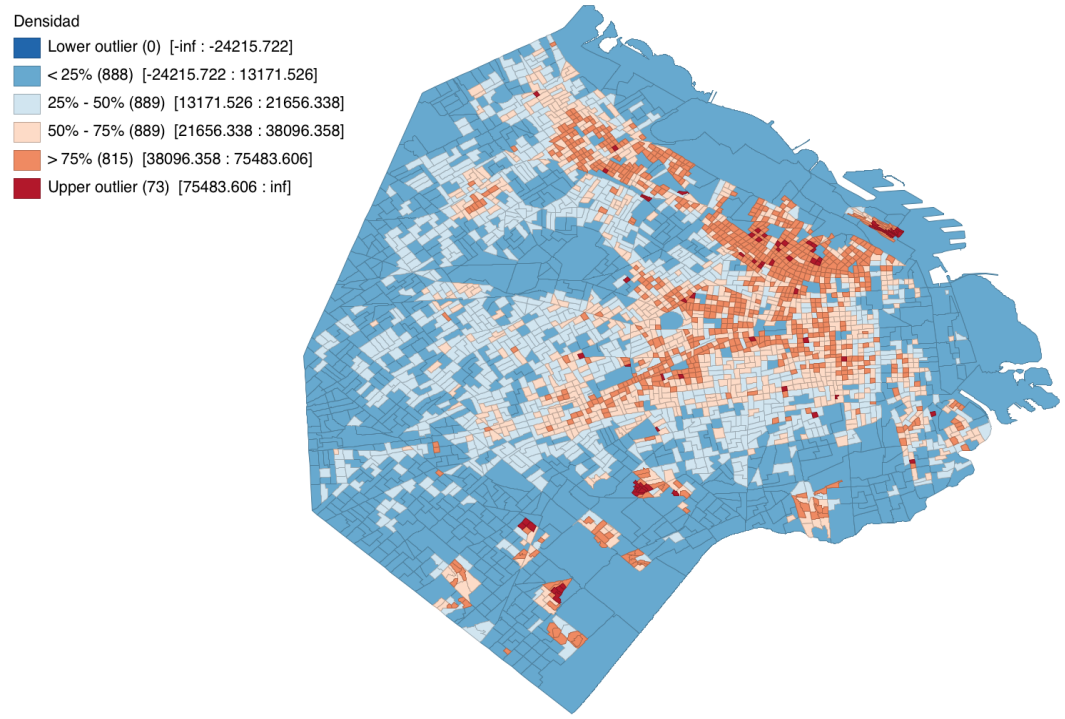
The blue areas represent census tracts with low numbers of cases, similarly surrounded by census tracts with the same characteristic.

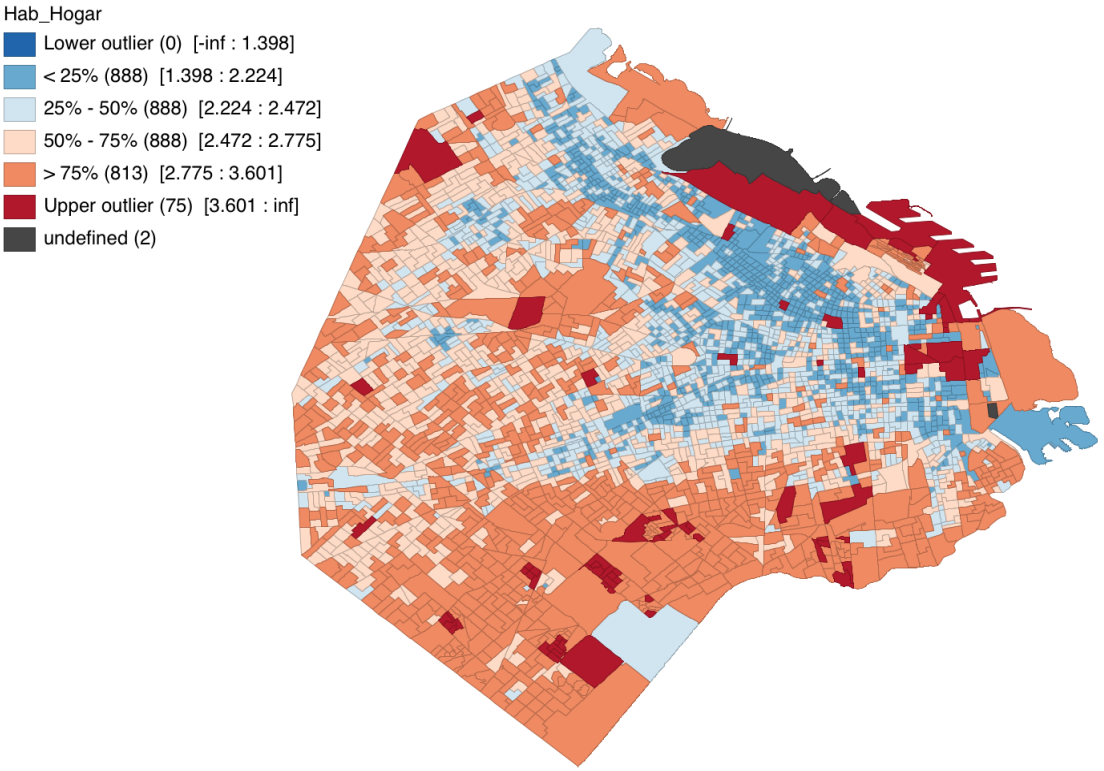
In this map, four areas in red correspond to neighborhoods with high case numbers and are part of slums in the city (Villa 31, Villa 1-11-14, Villa 21-24, and Carbonilla).



In this map, the red areas represent census tracts with many households with NBI, surrounded by others with a similar characteristic. The blue areas represent census tracts with low numbers of households with NBI, showing a relation between high levels of NBI households and vulnerable neighborhoods, as well as a north-south inequality in the city.

Finally, two box maps display the population density in the city and the average number of inhabitants per household. While these variables are not synonymous with vulnerability by themselves, high values in both contribute to the spread of the disease and are characteristics observed in deprived areas.


Densidad
Lower outlier (0) [-inf : -24215.722]
< 25% (888) [-24215.722 : 13171.526]
25% - 50% (889) [13171.526 : 21656.338]
50% - 75% (889) [21656.338 : 38096.358]
> 75% (815) [38096.358 : 75483.606]
Upper outlier (73) [75483.606 : inf]


Hab_Hogar
Lower outlier (0) [-inf : 1.398]
< 25% (888) [1.398 : 2.224]
25% - 50% (888) [2.224 : 2.472]
50% - 75% (888) [2.472 : 2.775]
> 75% (813) [2.775 : 3.601]
Upper outlier (75) [3.601 : inf]
undefined (2)

Conclusions

Based on the analysis of the sampled data of COVID-19 infected cases in the city, it can be concluded that **there is a significant relationship between living in an unfavorable socioeconomic and housing context and the probability of contracting the coronavirus.**

Through the number of households with unmet basic needs, the most vulnerable areas of the city were identified. This variable seems to have a significant influence on the number of cases observed per census tract and shows high values in neighborhoods considered vulnerable. The same applies to the variables of average inhabitants per household and population density, which may not directly indicate vulnerability but have shown high values in deprived areas and also significantly influence the number of cases.