DISSERTATIO

Estadística

COLEGIO DE GRADUADOS EN CIENCIAS ECONÓMICAS DE ROSARIO
CONSEJO PROFESIONAL DE CIENCIAS ECONÓMICAS DE LA PROVINCIA DE SANTA FE CÁMARA II
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

TRABAJOS FINALES

RECENSIÓN DE TESINAS Y PRÁCTICAS PROFESIONALES DE LA CARRERA LICENCIATURA EN ESTADÍSTICA

CONSEJO PROFESIONAL DE CIENCIAS ECONOMICAS DE LA PROVINCIA DE SANTA FE
CAMARA II

Colegio de Graduados en Ciencias Económicas de Rosario

UNR Universidad Nacional de Rosario

Año VI 2022, N.° 6, ISSN 2591-5215

ÍNDICE

CONFORMACIONES

Año VI 2022, N.º 6, ISSN 2591-5215

Esta revista se pone a disposición de los profesionales matriculados al Consejo Profesional de Ciencias Económicas de la Provincia de Santa Fe Cámara II (CPCE), asociados del Colegio de Graduados en Ciencias Económicas de Rosario (CGCE), estudiantes y docentes de la Facultad de Ciencias Económicas y Estadística (FCEyE) de la Universidad Nacional de Rosario (UNR) y otras Instituciones vinculadas al quehacer profesional y académico.

El contenido de los trabajos finales no necesariamente refleja la opinión de los Comités responsables de esta publicación digital.

Las Instituciones no son responsables por el contenido de las informaciones y opiniones que viertan en esta revista quienes son identificados como autores de dichos trabajos finales, en todos los casos deberán ser cotejadas por los Profesionales y/o las fuentes.

DISEÑO: AXYOMA GLOBAL
Dir. Nahuel Gasparutti
DG. Victoria Salvarezza

Córdoba 2035
S2000DFB, Rosario, Argentina
+54 (341) 5899241 / +54 (341) 6155910

EDITORIAL


Business Chart - Visual
Business Chart
Space Usage (750 Mb)
375 Mb used
250 Mb available
Who is your audience and what are their needs? This can help you better articulate the benefits of doing business with you and deliver a smarter product or service.
Interactive User
1,505
18,321
Realtime Dashboard
37.91%
31.86%
30.23%
Focus on Your Finances
Whether you're a full time, part time or freelance photographer - or even if you make a few bucks on the side from your photography - you are your own business.
Marketing Chart
31.25%
43.75%
36.50%
41.25%
44.50%
10:24

DE LA FACULTAD A LAS INSTITUCIONES PROFESIONALES

Desde 2015 el Consejo Profesional en Ciencias Económicas de la Provincia de Santa Fe Cámara II, el Colegio de Graduados en Ciencias Económicas de Rosario y la Facultad de Ciencias Económicas y Estadística de la Universidad Nacional de Rosario tomaron la iniciativa de habilitar este espacio que posibilita la publicación de tesinas de grado y trabajos finales de las licenciaturas que integran la Facultad, con el fin de incentivar la investigación y establecer un vínculo entre la vida académica y la profesional.

Convencidos que la formación y desarrollo técnico comienza y sienta sus bases en la facultad pero que debe continuar a lo largo de la vida profesional, para lograr su jerarquización y actualización permanente, es que el Consejo y el Colegio junto a la Facultad desarrollan múltiples actividades para acompañar a estudiantes avanzados y a recientes graduados, en esta importante etapa de sus carreras.

Ante contextos tan cambiantes; desde lo tecnológico hasta los modos de relacionarse, impartir conocimiento y aprender; las instituciones siguen apostando a construir y sostener alternativas de intercambio que agreguen valor y potencien a la comunidad educativa y profesional.

En ese marco, tiene lugar este proyecto de revistas digitales, iniciado con Dissertatio Economía en 2015, hoy en su octava edición; e incorporando en 2017 a Dissertatio Estadística y en 2018 a Dissertatio Administración, que ya cumplen su sexta y quinta edición respectivamente.

INDICADORES GLOBALES DE AUTOCORRELACIÓN ESPACIAL PARA UNIDADES DE DIFERENTE TAMAÑO

Lic. Sebastián Mario Ferraro

Director: Dr. José A. Pagura
Codirector: A.U.S. César Mignoni

En muchos estudios cuantitativos, las variables que se estudian se refieren a unidades que se pueden georreferenciar. Es frecuente observar que unidades cercanas tienen valores de las variables de interés parecidos, lo que se reconoce como correlación espacial positiva. Puede también ocurrir lo contrario: unidades con valores altos tienen vecinos con valores pequeños y se dice que existe correlación espacial negativa. La Estadística Espacial proporciona herramientas para detectar la ocurrencia de estos fenómenos, caracterizarlos mediante métodos descriptivos y construir modelos que expliquen estos comportamientos. Para la detección de la existencia de correlación espacial se emplean herramientas gráficas e indicadores siendo los más usuales el Box-map o el mapa de percentiles, el diagrama de dispersión de Moran y el índice de Moran. Cuando las unidades son de diferente tamaño y la variable en estudio depende de dicho tamaño, el mencionado índice puede conducir a conclusiones erróneas a pesar de algún pre tratamiento que se haya hecho a la variable para quitarle el efecto del tamaño de la unidad. Diferentes autores han presentado propuestas alternativas que tienen en cuenta el tamaño de la unidad; en este trabajo se analizan las propuestas conocidas como índice de Oden (1995) y Empirical Bayes Index (Assunção, 1999) y se los aplica, junto con el índice de Moran a dos problemas, analizando luego los resultados obtenidos y comparando las conclusiones.

INTRODUCCIÓN

En muchos estudios cuantitativos, las variables que se estudian se observan en unidades que se pueden referenciar geográficamente como lo son las manzanas de una ciudad, segmentos o radios censales, pixeles de imágenes de satélite, parcelas de un territorio, etc. Es frecuente observar que unidades cercanas tienen valores de las variables parecidos, por ejemplo características socioeconómicas, tipo de superficie, especies vegetales, como también, puede encontrarse la situación opuesta; esta característica se conoce como correlación espacial positiva o negativa. La Estadística Espacial proporciona herramientas para detectar la ocurrencia de estos fenómenos, caracterizarlos mediante métodos descriptivos y construir modelos que expliquen estos comportamientos. Para la detección de la existencia de correlación espacial se emplean herramientas gráficas e indicadores. Las más usuales son *Box-map* o el mapa de percentiles, el diagrama de dispersión de Moran y el índice de Moran. En cuanto a las representaciones gráficas, no presentan grandes inconvenientes para la elección de la más adecuada, en cambio para el cálculo de los índices de correlación espacial se requieren algunas consideraciones especiales. Cuando las unidades son de diferente tamaño y la variable en estudio depende de dicho tamaño, el mencionado índice puede conducir a conclusiones erróneas a pesar de algún pre tratamiento que se haya hecho a la variable para quitarle el efecto del tamaño de la unidad. Diferentes autores han presentado propuestas alternativas que tienen en cuenta esta situación. En este trabajo se analizan las propuestas conocidas como índice de Oden (1995) y *Empirical Bayes Index* (Assunção, 1999) y se los aplica, junto con el índice de Moran a dos problemas: uno con datos ficticios tomados a partir del problema existente en Rosario referente a delitos con armas de fuego, y otro con datos reales donde se busca caracterizar el comportamiento de la variable “cantidad de hogares con necesidades básicas insatisfechas” en la ciudad de Rosario.

METODOLOGÍA

Los métodos estadísticos tradicionales asumen que las observaciones de una variable se toman bajo condiciones idénticas y de manera independiente. Ellos consideran que los datos son una muestra aleatoria simple, es decir, son independientes e idénticamente distribuidos. Bajo esta suposición se construye la mayoría de la teoría estadística.

Tener en cuenta la dependencia en los datos es un gran inconveniente a la hora de trabajar con los modelos usuales. Sin embargo, en muchos casos los modelos que incluyen dependencia son más realistas que los que no lo hacen. En el contexto de datos espaciales, esta falta de independencia recibe el nombre de dependencia o autocorrelación espacial, la cual se puede describir mediante una relación funcional entre lo que ocurre en una unidad determinada del espacio y en sus unidades vecinas. En otras palabras, existirá autocorrelación espacial cuando el valor observado de una variable en una unidad o área determinada dependa, en cierta

manera, de los valores observados en unidades o áreas vecinas. Las herramientas gráficas permiten una aproximación a esa descripción y los índices de correlación espacial proveen medidas que confirman la existencia o no de correlación espacial así como la magnitud y dirección de la misma. El uso de indicadores requiere ciertas consideraciones y cuidados como se comenta a continuación.

Previo a la presentación de las características de los índices que se consideran, se menciona que el fundamento teórico de los métodos de la Estadística Espacial se encuentra en los procesos estocásticos: a cada unidad de la región le corresponde una variable aleatoria, la de interés para el estudio, con su correspondiente distribución de probabilidad, y las covariancias entre ellas, que reflejan la correlación espacial. Las observaciones obtenidas corresponden a una realización del proceso estocástico. A partir de estas ideas se reconoce la heterogeneidad espacial, fenómeno correspondiente a variancias diferentes de las distribuciones de las variables aleatorias y correlación espacial, que "habla" de la correlación de las variables en localizaciones próximas, y por último, la situación de no dependencia espacial. Estas dos últimas situaciones pueden suceder junto a la existencia de heterogeneidad espacial.

Los siguientes párrafos están dedicados a describir brevemente las características del índice de Moran, el más divulgado, y sus inconvenientes en presencia de unidades espaciales de diferente tamaño, situación usual en estudios socioeconómicos para luego mencionar dos propuestas para evitar dichos inconvenientes, conocidas como Índice de Oden y EBI (*Empirical Bayes Index*).

El índice de Moran (I)

Es el índice más divulgado. Fue desarrollado por Moran en 1950 y en la casi totalidad de los programas geoestadísticos se incluye su cálculo. Su interpretación es sencilla ya que es similar a la del coeficiente de correlación de Pearson. Se conocen sus propiedades estadísticas y pueden hacerse pruebas de hipótesis sobre su significación estadística. Sin embargo, cuando los tamaños de las unidades son diferentes y la variable para la cual se quiere evaluar la correlación espacial es una proporción o razón donde el denominador es una medida del tamaño de la unidad, este índice puede conducir a resultados erróneos.

Considérese una región R dividida en m áreas r_i, $i = 1, ..., m$, por ejemplo la ciudad de Rosario y sus radios censales, y que se desea evaluar la existencia de asociación espacial de "número de hogares con necesidades básicas insatisfechas" designado con n_i . El número de hogares de cada radio censal, x_i, es una variable que refleja el tamaño de los mismos.

No parece correcto abordar el estudio de la correlación espacial empleando n_i debido a las diferencias de tamaño existente en los radios censales. Con la pretensión de corregir el efecto del tamaño del radio, resulta natural plantear el estudio utilizando la razón observada en el área

i definida como $p_i = \frac{n_i}{x_i}$ la que podría mencionarse en forma coloquial como la proporción de hogares con NBI en el radio censal i.

El índice de Moran para la razón p_i, se define como:

$$I = \frac{m}{\sum_{ij} w_{ij}} \frac{\sum_{ij} w_{ij}(p_i - \bar{p})(p_j - \bar{p})}{\sum_i (p_i - \bar{p})^2} \quad \forall\, i \neq j,$$

donde p_i es el valor de la razón en la unidad i a la que se le asocia el conjunto de coordenadas s_i, vector cuyas componentes son las coordenadas espaciales,

$\bar{p} = \frac{\sum_{i=1}^{nm} p_i}{m}$ es la media de las razones p_i, y w_{ij} corresponde al peso entre las unidades i y j que es mayor que 0 si i,j son vecinos e igual a 0 en otro caso.

Resulta entonces imprescindible definir cuál es el criterio a aplicar para decidir si dos unidades son o no vecinas. Existen varios criterios de vecindad y en los problemas que se tratan en el presente trabajo se dice que dos unidades son vecinas si tienen algún límite común. El valor del peso w_{ij} para dos unidades vecinas se puede establecer de acuerdo a diferentes criterios que varían en su complejidad, siendo uno de ellos asignar 1 si las unidades son vecinas y 0 si no lo son. El índice de asociación de Moran resume la intensidad y dirección de la dependencia entre los valores de una variable observados en distintas unidades del espacio. Para probar la significación estadística de I y comprobar la hipótesis de no existencia de autocorrelación espacial se puede utilizar un test de hipótesis basado en supuestos de normalidad o, en caso de no verificarse dicho supuesto para la variable en estudio se utiliza un test permutacional. El rechazo de la hipótesis nula implica aceptar la existencia de correlación espacial.

El índice de Oden (I^*_{pop})

Cuando existen tamaños poblacionales distintos, Oden (1995) propone un ajuste al índice de Moran. Este índice que se notará con I^*_{pop} responde a la fórmula:

$$I^*_{pop} = \frac{n^2 \sum_{ij}^{m} M^*_{ij}\left(e_i - d_i\right)\left(e_j - d_j\right) - n(1-2\bar{b}) \sum_{i}^{m} M^*_{ii} e_i - n\bar{b} \sum_{ii}^{m} M^*_{ii} d_i}{\bar{b}\,(1-\bar{b})(x^2 \sum_{ij}^{m} d_i d_j M^*_{ij} - x \sum_{i}^{m} d_i M^*_{ii}},$$

donde:

$n = \sum_{i}^{m} n_i$ total de la variable en estudio en la región.

$x = \sum_{i}^{m} x_i$ es el total en la región de la variable utilizada como denominador.

$\overline{b} = \frac{n}{x}$ proporción de interés en la región.

$e_i = \frac{n_i}{n}$ proporción de la variable en estudio en la unidad i, con respecto al total en la región de la misma.

$d_i = \frac{x_i}{x}$ proporción con respecto al total en la región de la variable utilizada como denominador en la unidad i.

Se llama M_{ij} al elemento i, j de la matriz M de pesos espaciales definida por Oden de la siguiente manera:

$M_{ij} = \{ \quad w_{ij} \qquad\qquad si\ i\ y\ j\ son\ vecinos, \quad i = 1,..,m, \ j = 1,..,m \quad 2$

.

Como puede verse, los elementos de la diagonal principal de la matriz de vecindad tendrán valores iguales a 2, a diferencia de la matriz de vecindad utilizada en el índice de Moran, cuya diagonal está compuesta por elementos iguales a 0. Oden propone esta modificación, con el objeto de diferenciar la situación de dos áreas que no son vecinas a aquella cuando se compara un área consigo misma.

La cantidad M_{ij}^* incluida en el índice de Oden es: $M_{ij}^* = \frac{M_{ij}}{\sqrt{(d_i d_j)}}$.

Al igual que el índice de Moran, la prueba de significación estadística de la existencia de autocorrelación espacial con el índice de Oden se realiza bajo el supuesto de normalidad o mediante un test permutacional.

Oden (1995) muestra mediante estudios por simulación que el test que utiliza a I_{pop}^* es más potente que la prueba asociada a I cuando los tamaños de las unidades espaciales consideradas son muy variables.

Sin embargo debe decirse que el test de significación del índice de Oden plantea una hipótesis nula diferente a la del índice de Moran y su rechazo implica reconocer no solo la existencia de correlación espacial sino también la heterogeneidad espacial.

Assunção y Reis (1999) señalan con más detalle los inconvenientes mencionados destacando que es esperable observar la diferencia mencionada entre las potencias de los tests. Sin embargo estos tests no son comparables ya que prueban hipótesis diferentes.

De esta manera, surge la necesidad de encontrar un índice que tenga en cuenta el tamaño de las distintas áreas consideradas de una región para determinar si existe correlación espacial de una variable aleatoria.

Índice Empírico de Bayes (EBI)

Se trata de una propuesta realizada por Assunção y Reis (1999) y cuyos fundamentos se exponen brevemente en los párrafos siguientes.

Un fenómeno en el espacio se trata como un proceso estocástico, es decir como una colección de variables aleatorias para las que se indican sus ubicaciones en la región de estudio. En los estudios tratados en este trabajo, la variable aleatoria es una razón y se consideran los θ_1, θ_2, ..., θ_i, ..., θ_m parámetros - razones o proporciones- en las m áreas en estudio. Se realiza el supuesto que el número de eventos observados n_i sigue una distribución Poisson con media condicional $E(n_i|\theta_i) = Var(n_i|\theta_i) = x_i\,\theta$ siendo x_i el "tamaño" poblacional del área i. De esta forma, la media condicional de la razón estimada p_i es $E(p_i|\theta_i) = \theta_i$ y su variancia condicional es igual a $Var(p_i|\theta_i) = \frac{\theta_i}{x_i}$, por lo tanto, las razones estimadas poseen distintas medias y variancias condicionales.

Assunção y Reis (1999), considerando el enfoque bayesiano, tratan a los parámetros θi como variables aleatorias y realizan el supuesto de que las razones θi tienen una distribución a priori con esperanza $E(\theta i) = \beta$ y variancia $Var(\theta i) = \alpha$ y encuentran que la esperanza marginal de p_i es $E(p_i) = \beta$ y su variancia marginal es $Var(p_i) = \alpha + \beta x_i$. Puede verse que las razones poseen la misma esperanza marginal (β) y las variancias marginales difieren entre ellas dependiendo de los tamaños de las unidades (x_i). Las variancias marginales de las razones p_i se incrementan a medida que los tamaños de las áreas disminuyen.

Para estimar los parámetros α y β desconocidos, Marshall (1991) propone utilizar el método de los Momentos, los cuales conducen a los siguientes resultados:

$$\hat{\alpha} = a = s^2 - \frac{b}{(\frac{x}{m})},\ \hat{\beta} = b = \frac{n}{x},\ donde\ s^2 = \sum_i^m \frac{x_i(p_i - b)^2}{x}$$

De esta manera, la esperanza y variancia marginal son estimadas por b y $v_i = a + \frac{b}{x_i}$, respectivamente. Por convención, si $v_i < 0$, se define $v_i = \frac{b}{x_i}$.

En lugar de utilizar las proporciones p_i (como se emplean en el índice de Moran), se propone utilizar las razones estandarizadas, aplicando las estimaciones presentadas anteriormente: $y_i = \frac{p_i - b}{\sqrt{v_i}}$

El Índice Empírico de Bayes (EBI) se define de la siguiente manera:

$$EBI = \frac{m}{\sum_{ij}^{m} w_{ij}} \frac{\sum_{ij}^{m} w_{ij} y_i y_j}{\sum_i^m (y_i - \bar{y})^2}$$

Al igual que el Índice de Moran, EBI será positivo si las razones están directamente correlacionadas espacialmente. La prueba de independencia espacial se realiza a partir de la distribución del EBI obtenida mediante permutaciones.

Es decir, se permuta independientemente el vector $(y_1, y_2, ..., y_m)$ y se asignan aleatoriamente a las áreas una determinada cantidad de veces (en general se utilizan 999 permutaciones al igual que en el índice de Moran). Para cada una de las permutaciones se calcula el EBI. El valor de la probabilidad asociada al test de hipótesis del EBI está dado por el cociente entre la cantidad de veces que el EBI permutado excede el EBI observado (numerador) y la cantidad de permutaciones utilizadas (denominador).

Assunção y Reis (1999) estudian el efecto de "tamaños" de unidades heterogéneos sobre la potencia del test, es decir evalúan el impacto de la variación de los "tamaños" de las áreas cuando existe una correlación espacial entre las razones.

Las representaciones gráficas y los cálculos para el índice de Moran y EBI se obtuvieron recurriendo a los paquetes *sp* y *spdep* de R. Para el cálculo del índice de Oden se debió desarrollar un programa R específico. Los programas utilizados se encuentran disponibles en el siguiente repositorio de código en GitHub: *https://github.com/Seferra18/Tesina*

APLICACIÓN

La ciudad de Rosario se encuentra dividida en radios censales, unidades para las que usualmente se observan diferentes variables. Estas unidades presentan tamaños diferentes en término de superficie, número de viviendas o de otras cantidades similares. Muchas de las variables de interés en los estudios socioeconómicos presentan valores relacionados con dichas cantidades. Ejemplo de ello son el número de hogares con NBI y la cantidad de heridos por delitos con armas de fuego en la ciudad de Rosario durante un determinado período. Este trabajo se centra en la aplicación de los tres índices de correlación espacial que se mencionan, realizando finalmente los comentarios emergentes de los resultados que se obtienen.

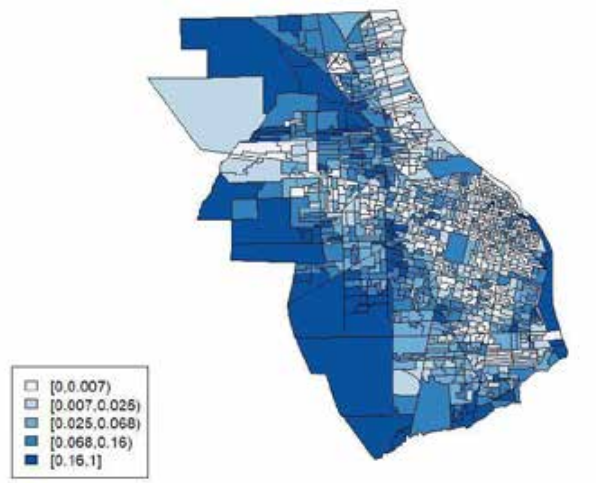
Estudio del comportamiento espacial de la cantidad de hogares con NBI en la ciudad de Rosario

Se analiza la proporción de hogares con NBI con respecto al total de hogares en el Censo de 2010, por radio censal. En primer lugar se utiliza una herramienta gráfica como lo es el *Box-map*, presentado en la Figura 1. Las categorías definidas para este gráfico son:

- [0; 0,007): corresponde a los radios con proporciones de NBI inferiores al primer cuartil que es 0,007 (expresado en porcentaje: 0,7% de hogares con NBI).
- [0,007; 0,025): agrupa los radios con proporciones de hogares con NBI comprendidos entre el primer cuartil y la mediana (2,5%).
- [0,025; 0,068): corresponde al 25% de radios censales con proporciones entre la mediana y el tercer cuartil (6,8%).
- [0,068; 0,16): agrupa los radios con proporciones de hogares con NBI iguales o mayores que el tercer cuartil y menores que el valor máximo habiendo excluido los "outliers".

- [0,16; 1): corresponde a “outliers” superiores.

Figura 1: Box Map de la proporción de hogares con NBI en los radios censales de la ciudad de Rosario. Año 2010



Los radios censales con menor proporción de hogares con NBI se observan en tonalidades más claras y se encuentran principalmente agrupados en la zona céntrica de la ciudad y en un sector de la zona norte. En la zona noroeste también se pueden distinguir agrupamientos de radios censales con baja proporción de hogares con NBI. Por otro lado, los radios con mayor proporción de hogares con NBI, representados con tonalidades más oscuras se concentran mayormente en la zona sur y oeste de la ciudad. Existen tres puntos que asumen un valor de la proporción igual a 1 correspondientes a radios censales con pocos hogares.

En segundo lugar, se calculan los tres índices de interés para la totalidad de los radios censales, excluyendo los radios con datos atípicos. Los valores calculados se presentan en la Tabla 1, junto a las probabilidades asociadas a las hipótesis nulas correspondientes. Se observan resultados estadísticamente significativos para las tres propuestas, es decir conducen a aceptar la existencia de correlación espacial. Vale destacar que, al excluir los 3 valores atípicos, se obtienen cambios importantes en el índice de Moran y diferencias menores en el EBI, lo cual puede deberse a la robustez mencionada por los autores del EBI.

Tabla 1. Índices de autocorrelación espacial calculados para la proporción de hogares con NBI

Índice	Estadístico	P-Valor
Conjunto completo de datos		
Moran (I)	0,39364	0,001
Oden (I^*_{pop})	0,08953	< 0,001
EBI	0,43339	0,001
Conjunto excluyendo radios con datos anómalos		
Moran (I)	0,47990	< 0,001
Oden (I^*_{pop})	0,08838	< 0,001
EBI	0,48152	0,001

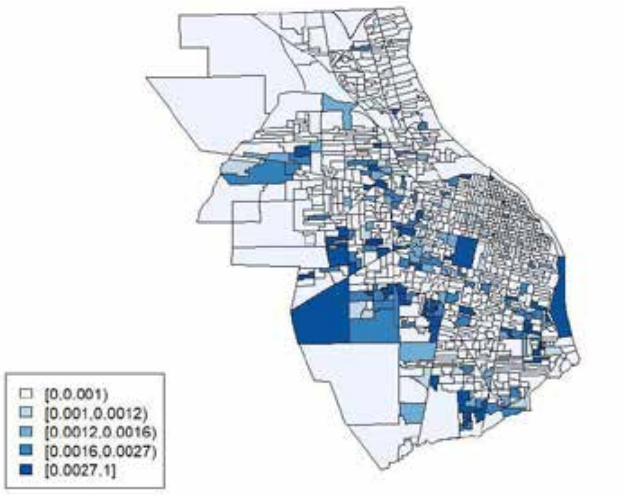
Estudio del comportamiento espacial de heridos por armas de fuego en la ciudad de Rosario

El otro problema a considerar es el comportamiento espacial de heridos por armas de fuego en la ciudad de Rosario. Los datos son ficticios, pero imitan el fenómeno de acuerdo a la Tesina presentada por Castro (2017).

La variable que se utiliza es la razón de número de heridos con respecto al total de habitantes en cada radio censal. Debido a los valores de las razones observadas, se presenta un mapa de percentiles, determinados de 5 en 5 comenzando por el percentil P_{75}.

El mapa de la Figura 2 muestra la agrupación de radios con menor razón de heridos por delitos con armas de fuego en las áreas representadas con una tonalidad más clara, por otro lado, los radios con mayor razón de heridos por delitos con armas de fuego asumen un color más oscuro. Si bien no se observa un patrón claro, puede identificarse mayoritariamente concentraciones de radios censales con razones altas en las zonas oeste y sur de la ciudad de Rosario.

Figura 2. Mapa de percentiles de la razón de heridos por arma de fuego en los radios censales de la ciudad de Rosario



La Tabla 2 contiene los valores de los índices junto con la probabilidad asociada a sus pruebas de hipótesis para los dos conjuntos de datos: completo y excluyendo outliers. Los resultados de los tres índices conducen a aceptar la existencia de correlación espacial y se encuentran las mismas particularidades mencionadas en el caso anterior.

Tabla 2. Índices de autocorrelación espacial para la razón de heridos por arma de fuego

Índice	Estadístico	P-Valor
Conjunto completo de datos		
Moran (I)	0,18779	0,001
Oden (I^*_{pop})	0,00155	<0,001
EBI	0,21715	0,001
Conjunto excluyendo tres radios con información anómala		
Moran (I)	0,21887	0,001
Oden (I^*_{pop})	0,00156	<0,001
EBI	0,22653	0,001

CONCLUSIONES

Se aplicaron tres índices para estudiar correlación espacial: índice de Moran, índice de Oden y *Empirical Bayes Index* (EBI) a problemas en los que las unidades espaciales son los radios censales de la ciudad de Rosario. El primero de ellos es el más divulgado, sin embargo algunos autores indican que no debería utilizarse en situaciones en las que los tamaños de las unidades presentan diferencias. Esto es muy frecuente en estudios socioeconómicos y en particular en los problemas de referencia: estudio del comportamiento espacial de la proporción de hogares

con NBI y razón de delitos con armas de fuego por número de habitantes. Las herramientas gráficas de empleo usual muestran asociación espacial y los indicadores confirman la existencia de este fenómeno mediante los resultados de las pruebas de hipótesis realizadas. Un comentario importante para realizar es que el EBI mostró resultados compatibles con la propiedad de robustez, lo que no se observó en el índice de Moran.

BIBLIOGRAFÍA

Anselin, L.; Syabri, I.; Kho, Y. (2006). "GeoDa: An Introduction to Spatial Data Analysis". Geographical Analysis 38 (1), 5-22.

Assunção, R. M.; Reis, E. A. (1999) "A new proposal to adjust Moran´s I for population density". Statist. Med. 18, 2147-2162.

Borra, V. (2015). "Estadística Espacial. Muestreo y modelización para la aplicación en estudios socioeconómicos". Tesis de Maestría en Estadística Aplicada, FCECON, UNR.

Castro, M. (2017). "Heridos por armas de fuego en la ciudad de Rosario en el 2012. Su comportamiento espacial". Tesina de Licenciatura en Estadística, FCECON, UNR.

Marshall, R. J. (1991). "Mapping disease and mortality rates using empirical Bayes estimators". Applied Statistics, 40, 283–294.

Moran, P. A. P. (1950). "Notes on continuous stochastic phenomena". Biometrika, 37, 17–23.

Oden, N. (1995). "Adjusting Moran's I for population density". Statistics in Medicine, 14, 17-26.

R Core Team (2020). "R: A language and environment for statistical computing".

Tobler, W. (1970). "A Computer Movie Simulation Urban Growth in the Detroit Region". Economic Geography 46(2),234-240.

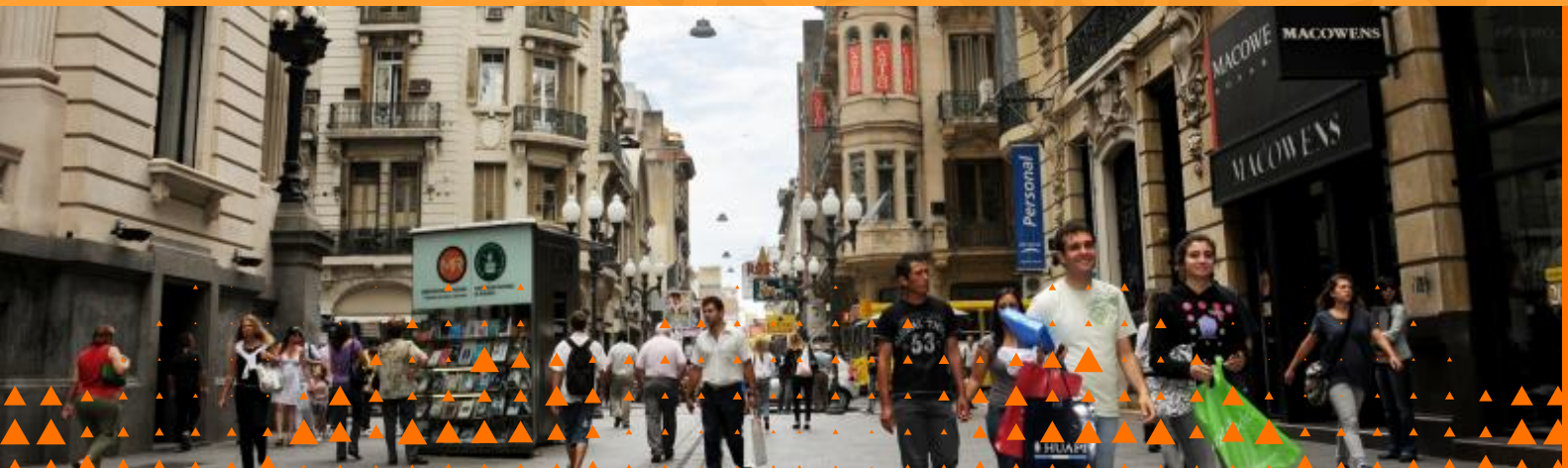
Walter, S. D. (1992). "The analysis of regional patterns in health data. I. Distributional considerations". American Journal of Epidemiology, 136, 730-741.

COMPARACIÓN DE LOS MÉTODOS DE ESTIMACIÓN DE LOS PARÁMETROS DEL MODELO DE SEMIVARIOGRAMA A TRAVÉS DE LA MODELIZACIÓN DE LA VARIABILIDAD ESPACIAL DEL PORCENTAJE DE PERSONAS DESOCUPADAS EN EL AÑO 2010 EN LA CIUDAD DE ROSARIO

Lic. Daiana Emili

Directora: Mg. Virginia Laura Borra

Un dato espacial refiere a una variable que se encuentra asociada a una localización del espacio físico. Para comprender cómo se distribuye la variable en el espacio y en qué grado las unidades pueden verse afectadas por sus vecinos, se emplea la autocorrelación espacial. Para determinar dicha correlación espacial se utilizan los modelos de semivariograma.
Existen diversos modelos teóricos de semivariograma y los mismos tienen tres parámetros comunes que deben estimarse: rango, efecto pepita y meseta. El objetivo es elegir entre todos los semivariogramas posibles aquel que mejor se ajuste a las observaciones, dado que dicho modelo será utilizado en la etapa de predicción. Por lo tanto, la estimación de los parámetros del modelo cumple un rol fundamental en el análisis. Los métodos estadísticos más frecuentes para la estimación de los parámetros son mínimos cuadrados y máxima verosimilitud.
En este trabajo se analizó el comportamiento espacial de la variable "porcentaje de personas desocupadas en el año 2010" en Rosario. Dado que los datos presentaron correlación espacial, se propusieron tres modelos que parecían adecuarse al semivariograma (exponencial, gaussiano y esférico). Se realizó la estimación de parámetros utilizando diferentes métodos de estimación y se compararon los resultados.
Se observó que los modelos estimados por mínimos cuadrados presentaron un ajuste al semivariograma experimental diferente al de los modelos estimados por máxima verosimilitud. Además, el método de máxima verosimilitud, mostró gran sensibilidad respecto a los valores iniciales que se utilizan en los procesos iterativos para estimar los parámetros de los modelos de semivariograma. Lo cual implicó que se deban realizar muchos ajustes considerando diversos valores iniciales hasta obtener un modelo adecuado para ajustar el semivariograma experimental.



INTRODUCCIÓN

Se define un dato espacial como todo aquel que tiene asociada una referencia geográfica, de tal modo que se puede localizar exactamente en un mapa. En estas ocasiones suele presentarse correlación espacial, lo cual indica que las unidades tomadas en una región específica no son independientes.

La geoestadística es una rama de la estadística que estudia datos espaciales. La misma tiene como objetivo la descripción de la correlación espacial, la estimación de parámetros y la obtención de predicciones de la variable de interés en puntos del espacio que no han sido muestreados. Una de las etapas en el desarrollo de un análisis geoestadístico es la determinación de la dependencia espacial o correlación entre las mediciones de la variable en estudio. Es decir, es necesario expresar mediante algún modelo espacial, la forma de la relación existente entre los valores de la variable y la distancia que separa las correspondientes unidades. Para llevar a cabo este análisis estructural, con base en la información muestral, se puede utilizar la función de semivariancia que permite representar la dependencia espacial entre unidades de muestreo vecinas, utilizando modelos espaciales.

La modelización de la dependencia espacial se obtiene a través de un modelo espacial teórico ajustable a la función semivariancia llamado modelo de semivariograma y las estimaciones de los parámetros del modelo ajustado se derivan a partir de los datos de la muestra.

Entre los modelos teóricos más conocidos se destacan los modelos esférico, exponencial y gaussiano. Todos estos modelos, tienen tres parámetros comunes: rango, meseta y efecto pepita.

En un estudio sobre la dependencia espacial de una variable, el objetivo será elegir entre diferentes semivariogramas posibles aquel que mejor se ajuste a las observaciones realizadas. Es sumamente importante que el modelo teórico ajustado sea el adecuado para el semivariograma experimental (Mc. Bratney & Webster, 1986). Por lo tanto, la estimación de los parámetros del modelo cumple un rol fundamental en el análisis, ya que la calidad de los resultados depende de la adecuación a la realidad del modelo que se proponga. Dicho modelo será utilizado posteriormente en la interpolación espacial de valores en lugares no muestreados.

Los métodos estadísticos más frecuentes de estimación de los parámetros del modelo de semivariograma que definen la estructura de dependencia espacial presentados en la literatura son: Mínimos Cuadrados Ordinarios, Mínimos Cuadrados Ponderados, Máxima Verosimilitud y Máxima Verosimilitud Restringida.

Con el objetivo de comparar dichos métodos, se presenta un caso práctico en el cual se analiza el comportamiento espacial de la variable “porcentaje de personas desocupadas en el año 2010” en la ciudad de Rosario por radio censal según datos del Censo Nacional de Población, Hogares y Viviendas del año 2010.

OBJETIVO

El objetivo de este trabajo consiste en comparar los resultados que proporcionan los métodos de estimación de parámetros del modelo de semivariograma a través de una aplicación práctica donde se modela la variabilidad espacial del porcentaje de personas desocupadas en el año 2010 en la ciudad de Rosario.

METODOLOGÍA

Para dar respuesta al objetivo mencionado se presenta una breve reseña de los métodos de estimación de los parámetros del modelo de semivariograma y se mencionan dos paquetes del software R que permiten llevar a cabo la aplicación de dichos métodos.

Si a cada sitio $s_i \in D \subset R^d$ le corresponde una variable aleatoria $Z(s_i)$, donde D es el dominio espacial del proceso, s_i representa una ubicación en el espacio euclidiano d-dimensional, y $Z(s_i)$ es la variable aleatoria en la ubicación s_i, entonces al conjunto de variables aleatorias $\{Z(s_i) / s_i \in D\}$ espacialmente distribuidas se lo llama proceso espacial (Ruiz, 2008).

Sea $Z = \{Z(s_i) / s_i \in D\} = \{Z(s_i)\}_{s_i \in D}$ un proceso espacial de segundo orden, intrínseco.

Entonces se define la función variograma del proceso espacial Z como:

$$\gamma_Z: D \to R^+_{\{o\}} / 2\gamma_Z(h) = Var\left[Z(s_i + h) - Z(s_i)\right],$$

para cualquier $h \in D$, siendo h el vector de separación. Notar que γ_Z no depende de s_i en D.

Cuando $Z = \{Z(s_i)\}_{s_i \in D}$ es un proceso espacial de segundo orden e isotrópico, la función de covariancia del proceso espacial Z, $C_Z(s_i, s_i + h)$, no depende de la dirección en la que ésta se calcule, $\forall\, s_i, s_i + h \in D$. A partir de ahora se decide denotar con t al módulo de h, $t = |h|$, y se utiliza la nomenclatura $Z(s_i + t)$ para hacer referencia a un punto separado de s_i por una distancia t en cualquier dirección, aunque en términos estrictos se está sumando al vector s_i otro vector en cualquier dirección, que tiene módulo t.

Cuando el proceso es estacionario de segundo orden e isotrópico, el valor esperado de Z es constante para todas las unidades y la expresión del variograma se simplifica a:

$$2\gamma_Z(t) = 2Var\left[Z(s_i + t)\right] - 2Cov\left[Z(s_i + t), Z(s_i)\right].$$

Por lo tanto, la función de semivariograma $\gamma_Z(t)$, en este caso, se puede escribir como:

$$\gamma_Z(t) = \sigma_Z^2 - C_Z(t).$$

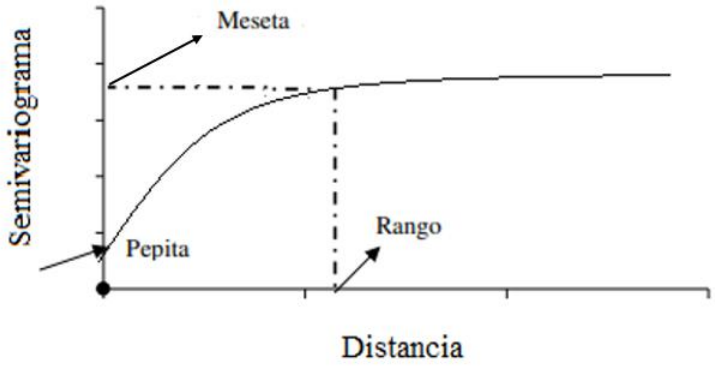
Modelos teóricos de semivariograma

Existen diversos modelos teóricos de semivariograma dependiendo de las características y condiciones que éstos deben cumplir (Samper & Carrera, 1990). En general dichos modelos pueden dividirse en no acotados (lineal, logarítmico, potencial) y acotados (esférico, exponencial, gaussiano) (Warrick, Myers, & Nielsen, 1986). Los del segundo grupo garantizan que la covariancia de los incrementos es finita, por lo cual son ampliamente usados cuando hay evidencia de que presentan buen ajuste. Todos estos modelos tienen tres parámetros comunes: el rango, el efecto pepita y la meseta.

Los semivariogramas más usuales para predicción alcanzan un valor límite denominado meseta (σ_Z^2) y es equivalente a la variancia de la población. El efecto pepita (τ^2) representa una discontinuidad puntual del semivariograma en el origen. Puede ser debido a errores de medición en la variable o a la escala de la misma. Cuando el efecto pepita es diferente de cero, se define la meseta como la suma entre el efecto pepita y la meseta parcial ($\sigma^2_{Z\,parcial}$). El rango ($\frac{1}{\phi}$) corresponde a la distancia t donde se alcanza la meseta. Cabe mencionar que algunos semivariogramas alcanzan la meseta sólo de manera asintótica.

En la Figura 1 se presenta como ejemplo un modelo teórico de semivariograma en forma gráfica donde se indica el significado geométrico de cada uno de sus parámetros.

Figura 1. Parámetros del semivariograma



Como primer paso se construye el semivariograma experimental que consiste en la estimación de las semivariancias por el método de los momentos. Posteriormente, se debe postular un modelo matemático que describa de la mejor forma posible al semivariograma experimental, el cual es conocido como semivariograma teórico o semivariograma paramétrico. Entre los modelos de semivarioagrama más conocidos se pueden mencionar: esférico, exponencial, gaussiano, wave y lineal. Ellos presentan diversas fórmulas y características.

Los criterios para seleccionar un modelo u otro dependen de los objetivos del trabajo. Si se cuenta con información a priori del comportamiento de la variable, puede ser interesante

realizar un ajuste manual de los modelos al semivariograma experimental. De esta forma el investigador puede fijar el efecto pepita, la meseta o el rango dependiendo del tipo de información que tenga de su variable y ajustar los parámetros de los que no tiene información.
Si el objetivo del trabajo es comparar los parámetros de los semivariogramas, la utilización de modelos diferentes resulta poco útil. Hay que tener en cuenta que, por ejemplo, los rangos del modelo esférico y el exponencial no son directamente comparables. El modelo esférico es el único que tiene una meseta verdadera, mientras que en los modelos exponencial y gaussiano la meseta es asíntota. Estos tres modelos se conocen como modelos de transición porque en ellos se puede estimar la meseta. El modelo lineal (al igual que otros modelos aquí no considerados) ni siquiera tiene meseta, y no es un modelo de transición. En este caso, es más conveniente elegir, cuando sea posible, un único modelo con motivo de comparar semivariogramas. El modelo esférico es el más usado seguido por el exponencial y en tercer lugar el gaussiano. Este último refleja muy bien la continuidad espacial, la interpolación de puntos basada en este modelo es muy exigente, produciendo frecuentemente representaciones gráficas alejadas de la realidad (Gallardo, 2006).

Estimación de parámetros

Para poder obtener predicciones de la variable aleatoria bajo estudio en sitios para los cuales no se cuenta con una muestra, es necesario poder estimar los parámetros del modelo de semivariograma escogido.

A continuación, se presentan los métodos estadísticos más frecuentes de estimación de los parámetros $\varphi = (\frac{1}{\phi}, \sigma_z^2, \tau^2)^T$ que definen la estructura de dependencia espacial, los cuales son: Mínimos Cuadrados Ordinarios, Mínimos Cuadrados Ponderados, Máxima Verosimilitud y Máxima Verosimilitud Restringida.

<u>Estimación por mínimos cuadrados</u>

El semivariograma es una función de la distancia t y para calcularlo se toman intervalos de distancia. Todos los pares de unidades se clasifican en estos intervalos de acuerdo a la distancia de a pares, es decir, se discretiza la distancia $t_1, t_2, \ldots, t_3$ formando K intervalos.

El método de mínimos cuadrados ordinarios (OLS) consiste en minimizar la función:

$$R^2 = \sum_{j=1}^{k} (\hat{\gamma}(t_j) - \gamma(t_j, \varphi))^2,$$

donde $\hat{\gamma}(t_j)$ es el valor estimado del semivariograma experimental correspondiente a la distancia t_j, utilizando el estimador de Matheron y $\gamma(t_j, \varphi)$ es el valor estimado del

semivariograma correspondiente a la distancia t_j, dado por el modelo espacial teórico seleccionado para ser ajustado, evaluado en $\varphi = \hat{\varphi}$.

Por su parte, el método de mínimos cuadrados ponderados (*Weighted Least Squares*, WLS) consiste en minimizar la función:

$$R^2 = \sum_{j=1}^{k} w_j\left(\hat{\gamma}\left(t_j\right) - \gamma\left(t_j, \varphi\right)\right)^2 \qquad [1]$$

donde los pesos w_j, atribuidos a los k intervalos calculados para graficar el semivariograma experimental son seleccionados de acuerdo con algún criterio estadístico.

En la literatura, tal como refieren Cressie (1993), Diggle y Reibeiro Jr. (2007) y en programas tales como Idrisi 32 (Eastman, 2001) (sistema abierto de análisis geográfico), aparecen disponibles tres criterios de ajuste por el método de mínimos cuadrados ponderados (WLS1 y WLS2, WLS3).

Estimación por máxima verosimilitud

Sea $Z = \{Z(s_i)\}_{s_i \in D}$ un proceso espacial intrínsicamente estacionario y gaussiano. Se asume que para cualquier conjunto de n puntos $s_1, s_2, ..., s_n \in D$, $Z \sim N\left(\mu_Z I, \sum_{\varphi}\right)$, donde $\mu_Z: D \rightarrow R$ / $\mu_Z\left(s_i\right) = E\left[Z\left(s_i\right)\right]$ $\forall\, s_i \in D$, I es un vector de unos de dimensión $n * 1$ y $\sum_{\varphi}$ es la matriz de variancias y covariancias de las observaciones. $\sum_{\varphi}$ es una matriz simétrica, definida positiva. Un elemento genérico de $\sum_{\varphi}$ es $C_Z\left(s_i, s_j\right) = Cov\left(Z\left(s_i\right), Z\left(s_j\right)\right)$ $\forall\, s_i, s_j \in D$ y se puede expresar a través de los parámetros de los modelos de semivariogramas válidos, $\varphi = \left(\frac{1}{\phi}, \sigma_z^2, \tau^2\right)^T$.

La estimación mediante máxima verosimilitud (*Maximum Likelihood*, ML) consiste en obtener, de forma simultánea, los valores de μ_Z y φ que minimizan la función:

$$L\left(\mu_Z, \varphi\right) = \left(Z - \mu_Z I\right)^T \sum_{\varphi}^{-1} \left(Z - \mu_Z I\right) + \left(\left|\sum_{\varphi}\right|\right) + n * ln(2\pi).$$

Generalmente, φ sólo incluye tres parámetros (el efecto pepita, la meseta parcial y el rango).

Otro método utilizado para estimar los parámetros de la función de covariancia $\sum_{\varphi}$ es el de Máxima Verosimilitud Restringida (*Restricted Maximum Likelihood*, REML) (Cressie, 1993). Este método tiene mejores propiedades de sesgo que la estimación por máxima verosimilitud. Consiste en eliminar la media de la función de verosimilitud de modo que quede definida sólo en términos de la matriz de variancias y covariancias.

Si $Z \sim N(\mu_z I, \sum_{\varphi})$, entonces, $F = AZ \sim N\left(0, A\sum_{\varphi} A^T\right)$ y el estimador de máxima verosimilitud restringido de φ consiste en maximizar el logaritmo de la función de verosimilitud restringida, o lo que es lo mismo, minimizar la función:

$$L(\varphi) = (n - 1)\ln ln\,(2\pi) + \ln ln \left|A\sum_{\varphi} A^T\right| + F^T\left(A\sum_{\varphi} A^T\right)^{-1} F.$$

En consecuencia, los estimadores de máxima verosimilitud restringida, denominados también de máxima verosimilitud de los residuos, son estimadores de los parámetros desconocidos de $\sum_{\varphi}$, que se obtienen maximizando la función de verosimilitud de una nueva variable definida a partir de la original.

Medidas de bondad de ajuste

Existen diversas medidas de bondad de ajuste que permiten la comparación y elección del mejor modelo de semivariograma.

Cuando se utiliza mínimos cuadrados ordinarios, el modelo que mejor se ajusta al semivariograma experimental es aquel que presenta menor valor de la estadística R^2 definida en la Ecuación 1. Cuando se utilizan métodos de máxima verosimilitud para estimar los parámetros del semivariograma, se puede recurrir a estadísticas como el valor de menos dos veces la log-verosimilitud asociada $(-2ln)$, el criterio de información de Akaike (AIC) (Akaike, 1974) o el criterio de información bayesiano (BIC) (Schwarz, 1978) con el fin de elegir el mejor modelo de semivariograma (Faraway, 2005).

El AIC es definido como:

$$AIC = -2\ln ln\,(L(\varphi)) + 2u$$

donde, $\ln ln\,(L(\varphi))$ es el logaritmo de la función de verosimilitud evaluado en $\varphi = \hat{\varphi}$ y u es el número de parámetros del modelo ajustado que se calcula como la suma entre la cantidad de covariables bajo estudio y 3, haciendo referencia a los 3 parámetros del semivariograma que se deben estimar.

En consecuencia, entre dos modelos con el mismo valor de $\ln ln\,(L(\varphi))$, el AIC clasifica mejor al que tiene menor cantidad de parámetros. La decisión de elegir entre los modelos utilizados en el ajuste, recae sobre aquel modelo que presenta el menor valor de AIC.

APLICACIÓN

Con el objetivo de comparar los métodos de estimación de los parámetros del semivariograma, se analiza el comportamiento espacial de la variable “porcentaje de personas desocupadas en el año 2010” en la ciudad de Rosario. Se obtuvo las coordenadas “X” e “Y” para el centroide de cada uno de los 1069 radios censales y se asignó a dicho punto el valor de la variable. Posteriormente, se realizó un análisis descriptivo acerca de la distribución de la variable bajo estudio. Para esto, se realizaron los gráficos de los datos frente a las coordenadas, el histograma de frecuencia y se observó cómo se distribuye el porcentaje de personas desocupadas en las posiciones espaciales distinguiendo los cuartiles de la distribución.

Se detectó que el porcentaje de personas desocupadas por radio censal en la ciudad de Rosario en el año 2010 varía entre 0.00% y 11.28%, donde el 50% de los radios censales bajo estudio presentó al menos un 3.84% de personas desocupadas (IQR = 2.17%).

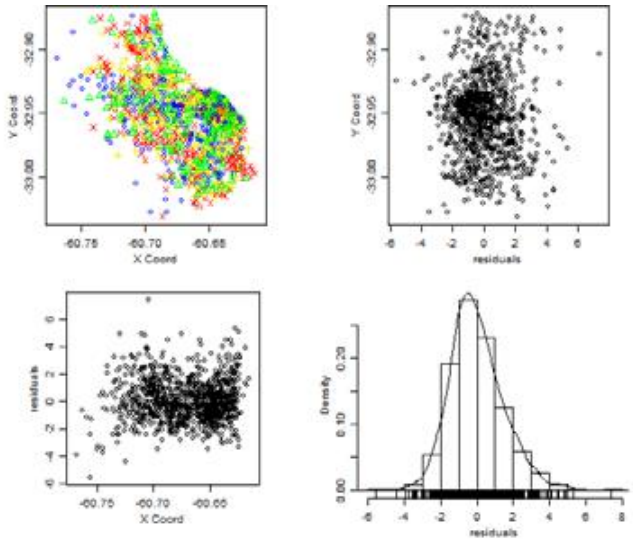
Los gráficos de dispersión de los datos frente a las coordenadas pueden ayudar a determinar si existe una tendencia espacial, es decir detectan si los valores de una variable están relacionados con sus propias coordenadas geográficas. Dichos gráficos mostraron una mayor concentración de puntos en valores más altos de las coordenadas “X” de los radios censales y valores bajos de los porcentajes de personas desocupadas por radio censal. Lo cual podría indicar que existe una leve tendencia dado que a medida que aumentan los valores de las coordenadas “X” de los radios censales, disminuye el porcentaje de personas desocupadas en los radios censales. El mismo escenario se observó al graficar la variable de interés versus las coordenadas “Y” de los radios censales.

El siguiente paso consistió en eliminar la tendencia encontrada ajustando un modelo lineal que utiliza las coordenadas “X” e “Y” de los radios censales como covariables y el porcentaje de personas desocupadas en cada radio censal como variable respuesta. Ambas variables resultaron estadísticamente significativas. Luego se calcularon los residuos del modelo como la diferencia entre los valores observados y los valores ajustados por dicho modelo.

Mediante el test de Shapiro wilk, se comprobó que los residuos no seguían una distribución normal. Sin embargo, la asimetría de la distribución de los residuos fue 0.54 y la curtosis 4.04, lo cual indica que los residuos siguen una distribución leptocúrtica, lo cual significa que hay una mayor concentración de los datos en torno a la media de la distribución.

Posteriormente se graficaron los residuos versus las coordenadas “X” e “Y” de los radios censales y se detectó que la tendencia que parecía existir en los datos desapareció, ya que en la Figura 2 se observa que los puntos están dispersos alrededor del cero tanto para las coordenadas del eje “X” como para el eje “Y”.

Figura 2. Resumen descriptivo de los residuos del modelo ajustado al “porcentaje de personas desocupadas en el año 2010”

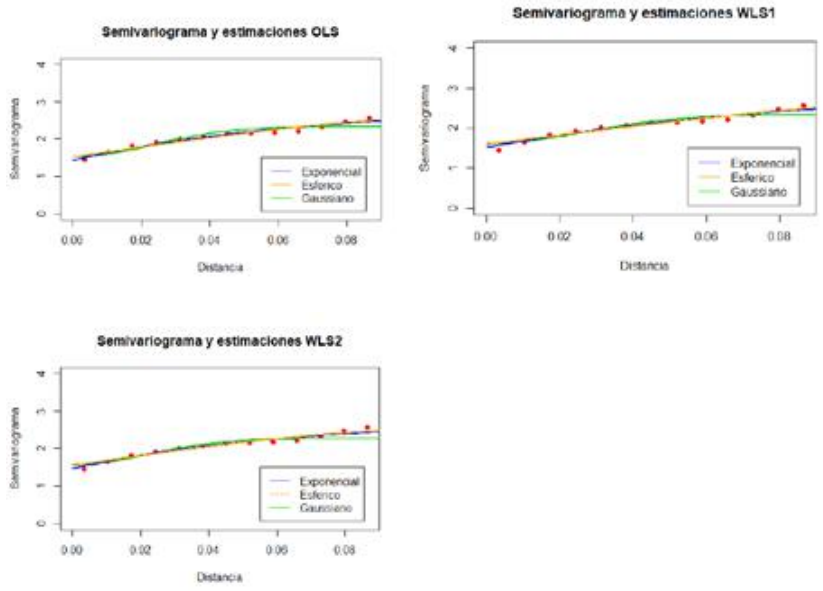


Para continuar, se ajustó el semivariograma experimental a los residuos del modelo y se concluyó que los datos presentaban correlación espacial. Mediante una inspección visual del semivariograma se aproximaron los valores de los parámetros del modelo, donde una primera estimación visual sería pensar una meseta igual a 2.50, una meseta parcial de 1.10, un efecto pepita de 1.40 y un rango de 0.09. Luego se propusieron tres modelos teóricos que parecían adecuarse al semivariograma experimental: el modelo exponencial, gaussiano y esférico.

Para cada uno de los modelos se realizó la estimación de los parámetros con los diversos métodos expuestos en la metodología y luego se compararon los modelos utilizando la estadística R^2 en las estimaciones por mínimos cuadrados y el AIC en las estimaciones por máxima verosimilitud.

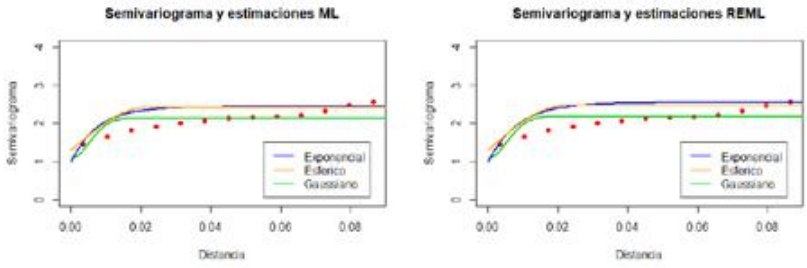
En la Figura 3 se puede observar el gráfico del semivariograma experimental y los semivariogramas estimados por los métodos OLS, WLS1 y WLS2. Las estimaciones por mínimos cuadrados brindan un buen ajuste para el semivariograma experimental.

Figura 3. Semivariograma experimental y semivariogramas estimados por los métodos OLS, WLS1 y WLS2



Al utilizar ML o REML para obtener la estimación de los parámetros del semivariograma, se observó que dichos métodos son muy sensibles a los valores iniciales otorgados para iniciar el proceso iterativo. En consecuencia, se tuvieron que realizar varias pruebas con diferentes valores iniciales hasta lograr obtener aquellas estimaciones que se ajusten al semivariograma experimental.

Figura 4. Semivariograma experimental y semivariogramas estimados por los métodos ML y REML



A pesar de haber realizado varias pruebas con diferentes valores iniciales para los procesos iterativos, la Figura 4 muestra que aún existe una leve discrepancia entre el semivariograma experimental y los modelos ajustados por ML y REML.

Finalmente, para cada método de estimación de parámetros se seleccionaron los modelos que mejor ajustaban el conjunto de datos, los cuales fueron: el modelo exponencial en los métodos OLS y WLS1 y el modelo esférico en los métodos WLS2, ML y REML.

Se observó que los modelos estimados por mínimos cuadrados presentaron un ajuste al semivariograma experimental diferente al de los modelos estimados por máxima verosimilitud. En la Tabla 1, se observa que los métodos por mínimos cuadrados estiman que la correlación espacial desaparece en promedio luego de una distancia de 0.09, mientras que los métodos por máxima verosimilitud estiman que la autocorrelación es nula a partir de 0.02.

Tabla 1. Estimaciones de los parámetros de los modelos elegidos

Método de estimación	Modelo	Meseta Parcial	Rango	Efecto Pepita
Estimación visual		1.10	0.09	1.40
OLS	Exponencial	1.45	0.07	1.43
WLS1	Exponencial	1.51	0.09	1.52
WLS2	Esférico	0.96	0.11	1.55
ML	Esférico	1.12	0.02	1.30
REML	Esférico	1.17	0.02	1.30

CONSIDERACIONES FINALES

Se realizó la comparación de los resultados de los métodos de estimación de los parámetros del semivariograma en el comportamiento espacial de la variable “porcentaje de personas desocupadas por radio censal en el año 2010” en la ciudad de Rosario. Se observó que los resultados hallados por ambos métodos de estimación difieren entre ellos. Los métodos de máxima verosimilitud mostraron una gran sensibilidad respecto a los valores iniciales utilizados en los procesos iterativos que estiman los parámetros de los modelos de semivariograma. Esto no ocurre al utilizar el método de mínimos cuadrados, ya que tomando como valores iniciales las estimaciones de los parámetros que surgen a partir de una inspección visual del semivariograma experimental, dichos métodos proporcionan un buen ajuste al semivariograma experimental.

La sensibilidad de los métodos de máxima verosimilitud, implicó que se deban realizar muchos ajustes considerando diversos valores iniciales hasta obtener un modelo adecuado para ajustar el semivariograma experimental. Aun así, las estimaciones por mínimos cuadrados mostraron un mejor ajuste.

BIBLIOGRAFÍA

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, vol. 19, no. 6., 716-723.

Cressie, N. (1993). Statistics for Spatial Data. New York: John Wiley & Sons.

Diggle, P., & Reibeiro Jr., P. (2007). Model based geostatistics. New York: Springer.

Eastman, J. (2001). IDRISI 32 Release 2 - Guide To GIS and Image Processing Volume 2. Worcester, MA: Clark Labs, Clark University.

Faraway, J. (2005). Linear Models with R. London: Chapman and Hall.

Gallardo, A. (2006). Geoestadística. Ecosistemas, 15(3).

Matheron, G. (1962). Traité de Géostatistique Appliquée. Tomo 1. Paris: Ediciones Technip.

Mc. Bratney, A., & Webster, R. (1986). Choosing Functions for Semi-Variograms of Soil Properties and Fitting Them to Sampling Estimates. European Journal of Soil Science, 37, 617-639.

Ruiz, F. M. (2008). Modelización de la función de covarianza en procesos espacio-temporales: análisis y aplicaciones. Valencia: Universidad de Valencia.

Samper, F., & Carrera, J. (1990). Geoestadística: aplicaciones a la hidrogeología subterránea. Barcelona: Centro Internacional de Métodos Numéricos en Ingeniería.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics 6, 461-464.

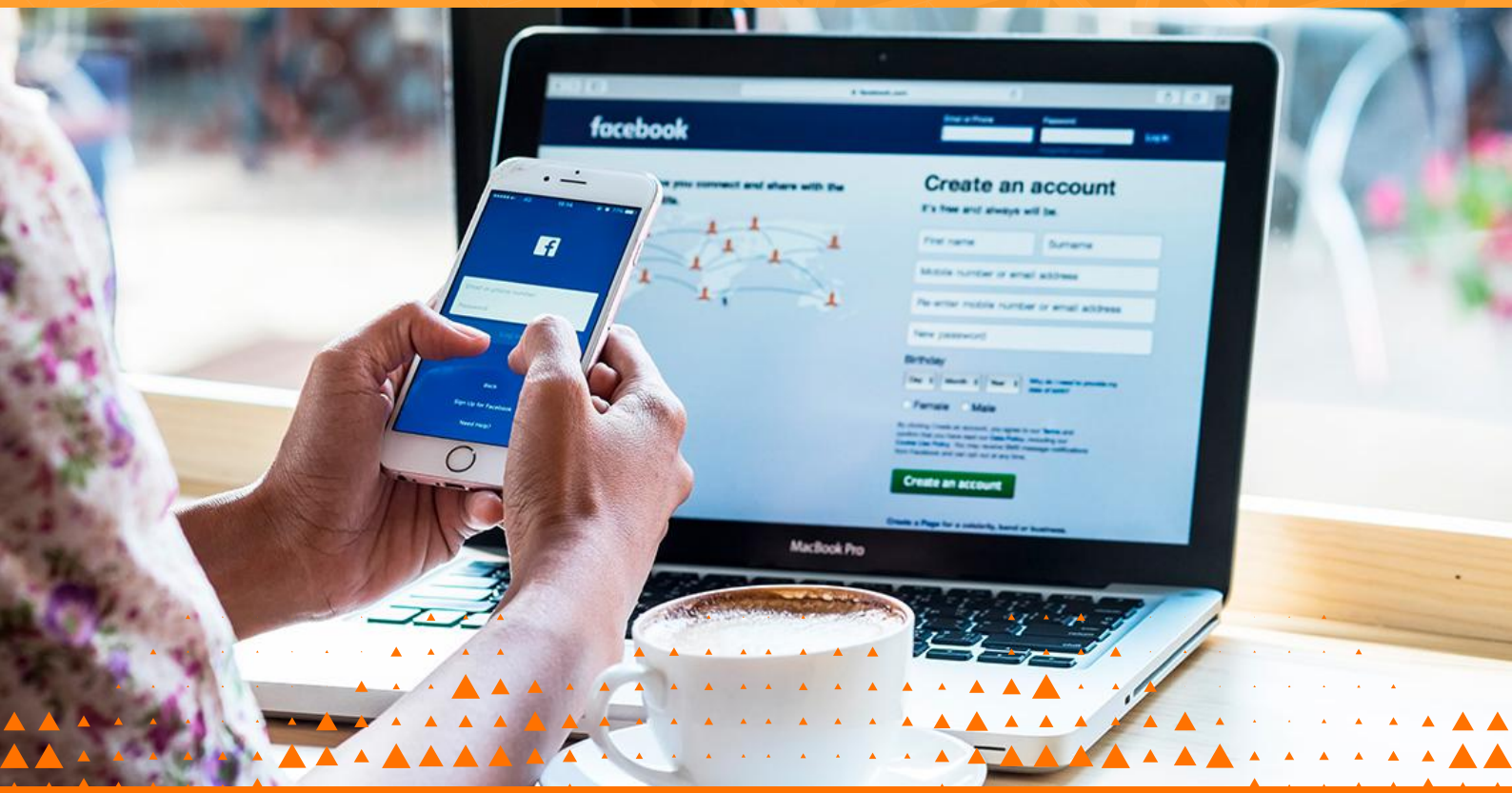
Warrick, A., Myers, D., & Nielsen, D. (1986). Geostatistical Methods Applied to Soil Science. Methods of Soil Analysis. Part 1. Physical and Mineralogical Methods. Agronomy Monograph 9, 53 - 81.

ANÁLISIS DE LOS TEXTOS PUBLICADOS EN FACEBOOK POR LOS CONCEJALES DE ROSARIO, ARGENTINA, DURANTE EL AÑO 2019

Lic. Luisina Rubio

Responsable de la Facultad de Ciencias Económicas y Estadística: Mg. Marcos Prunello
Responsable de la entidad: Lic. Julián Crucella

El presente trabajo es el producto de una Práctica Profesional desarrollada en la empresa Inmediata que tuvo por objetivo analizar la comunicación en redes sociales de los concejales de Rosario en el año electoral 2019. Se tomaron todos los mensajes publicados en sus páginas públicas de Facebook entre febrero y noviembre y se caracterizó el contenido de las publicaciones de cada bloque empleando técnicas para el análisis de texto, tales como nubes de palabras y grafos de asociaciones. Además, se ajustó un modelo de tópicos con Asignación Latente de Dirichlet, siendo posible identificar 3 temáticas en los mensajes de los concejales: propuestas políticas, jerga de campaña y cuestiones de género y derechos humanos, así como también señalar los temas de preponderancia para cada bloque político.



INTRODUCCIÓN

En el último tiempo las redes sociales han alterado las interacciones humanas permitiendo a las personas tanto compartir sus opiniones y experiencias como reaccionar sobre lo publicado por otros. La popularidad alcanzada por las redes sociales se ha trasladado rápidamente al terreno de la política, donde la mayoría de los partidos, gobernantes y/o funcionarios, cuentan con ellas para conectarse con la ciudadanía. En este trabajo se emplearon técnicas de análisis provenientes del *social media listening* (escucha de redes sociales) con el objetivo de explorar los temas abordados en *Facebook* por los concejales de la ciudad de Rosario durante el año electoral de 2019.

El Concejo Municipal de Rosario está compuesto por 28 ediles. En el año 2019, antes de la renovación electoral, los concejales se encontraban divididos en tres bloques y dos interbloques, conformados como se muestra en la Tabla 1 (se utilizará de ahora en adelante el término "bloques" para referirse tanto a los bloques como a los interbloques). Se analizaron todas las publicaciones realizadas en *Facebook* entre el 1º de febrero y el 30 de noviembre de 2019 por los 26 miembros del Concejo Municipal que contaban con página pública en dicha red social.

Los mensajes publicados fueron sometidos a procesos de limpieza característicos del análisis textual para posteriormente analizar las frecuencias de términos y las asociaciones entre palabras y ajustar un modelo de tópicos, permitiendo realizar comparaciones entre los distintos bloques políticos.

Tabla 1. Composición del Concejo Municipal de Rosario en el año 2019 según bloques e interbloques, previo a la renovación de sus miembros

	Nombre	Color de identificación	Concejales
Bloques	Cambiemos (C)		Agapito Blanco, Agustina Bouza, Alejandro Roselló, Anita Martínez, Charly Cardozo, Gabriel Chumpitaz, Germana Figueroa Casas, Renata Ghilotti y Roy López Molina
	Ciudad Futura (CF)		Juan Monteverde, Caren Tepp, Pedro Salinas y Jessica Pellegrini (los últimos dos no contaban con página pública de *Facebook*).
	Frente Social y Popular (FSyP)		Celeste Lepratti
Interbloques	Frente Progresista Cívico y Social (FP)		Aldo Poy, Lisandro Zeno, Horacio Ghirardi, Enrique Estévez, Verónica Irizar, María Eugenia Schmuck y Pablo Javkin
	Frente Nacional y Popular (FNyP)		Norma López, Roberto Sukerman, Marina Magnani, Andrés Giménez, Eduardo Toniolli, Osvaldo Miatello y Fernanda Gigliani

DESCRIPCIÓN Y TRATAMIENTO DE LOS DATOS

Los datos analizados en este trabajo fueron recolectados mediante la plataforma *Social Insider* que permite capturar las publicaciones realizadas en cuentas públicas de *Twitter*, *Instagram* y *Facebook* junto con métricas asociadas (por ejemplo, número de reacciones como "Me Gusta", número de veces que fue compartida, número de respuestas o comentarios, número de seguidores de la cuenta, etc.). Durante la primera quincena del mes de enero de 2020 se obtuvieron los datos correspondientes a todas las publicaciones de *Facebook* de cada concejal realizadas entre febrero y noviembre del año 2019, resultando en un total de 4781. De éstas, 4399 fueron retenidas para el análisis por incluir mensajes textuales.

El análisis de datos textuales se caracteriza por la necesidad de realizar un proceso particular de limpieza de los mismos, siendo necesario reconocer la terminología propia de la disciplina. En primer lugar, se define como *token* al conjunto de caracteres que constituyen la unidad de análisis, pudiendo tratarse de una única palabra, un conjunto de n palabras (n-grama), una oración, un párrafo, una página, etc. En este trabajo, cada palabra de los mensajes publicados fue considerada como un *token* (de esta forma, las expresiones "*token*", "palabra" o "término" refieren al mismo concepto). El proceso de dividir un texto en *tokens* se conoce como *tokenización.* En segundo lugar, se define como documento a una secuencia de *tokens* vinculados entre sí. En este contexto, el conjunto de *tokens* que compone cada mensaje conforma un documento. Finalmente, se le dice *corpus* a la colección de documentos a analizar, tratándose en este caso del conjunto de 4399 mensajes analizados.

Con el objetivo de buscar *tokens* no óptimos o de poca utilidad para reemplazarlos o removerlos, se realizó un proceso de depuración de los documentos que consistió en:

- Convertir todo el texto a minúscula para evitar que una misma palabra escrita con distintos estilos de capitalización sea distinguida como palabras diferentes.
- Unir nombres propios para convertirlos en una sola palabra con el fin de evitar redundancias (por ejemplo, "mauricio macri" se reemplazó por "mauricio_macri").
- Unificar distintas formas de nombrar a la misma persona (por ejemplo, "cristina fernandez", "cristina fernández o "cristina kirchner").
- Remover palabras vacías como preposiciones o conectores, conocidas como *stopwords*.
- Eliminar espacios vacíos, saltos de línea y puntuaciones.
- Eliminar enlaces y etiquetas.
- Acortar palabras que hayan sido distorsionadas mediante la repetición de caracteres (por ejemplo, "buenooo" se convirtió en "bueno")
- Eliminar fechas, horas, números y caracteres especiales (por ejemplo, emojis).

El análisis se realizó empleando el software estadístico R, utilizando paquetes específicos para la limpieza de datos de texto (*tidytext*, Silge y Robinson, 2016; *tm,* Feinerer y Hornik, 2020),

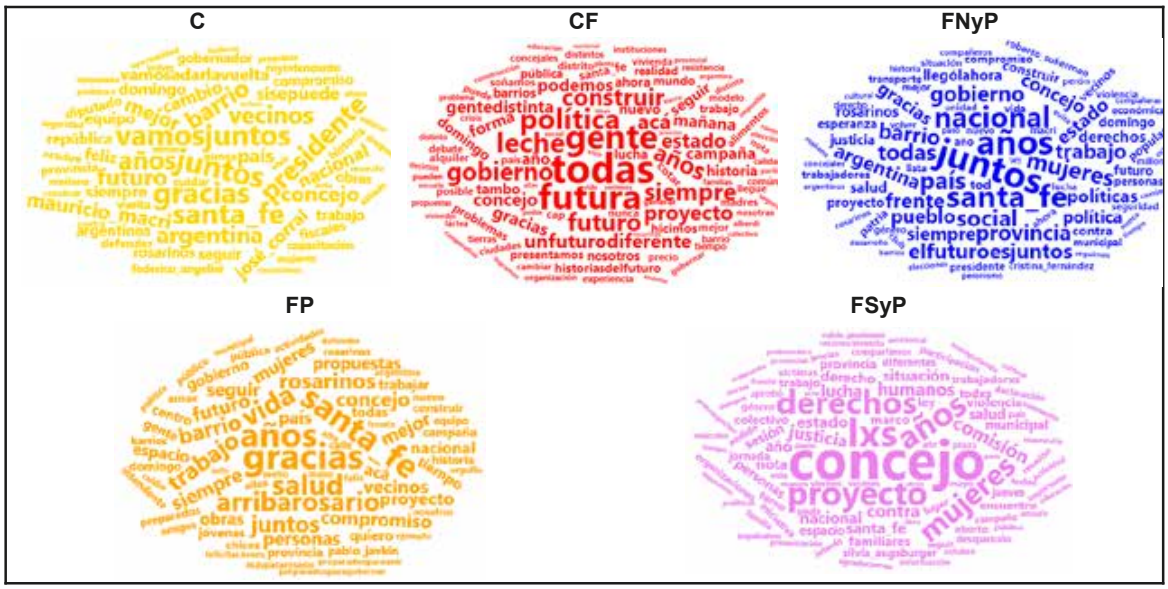
para el ajuste de modelos de tópicos (*textmineR*, Jones, 2019) y otros para tareas generales de manipulación de datos y visualización (*tidyverse*, Wickham et al., 2019). Tanto el código de R empleado como los datos para replicarlo se encuentran en https://github.com/luisinarubio/PracticaProfesional.

PALABRAS DE USO FRECUENTE

Una de las formas de analizar el contenido de un documento consiste en determinar el número de veces que se presenta cada *token* que lo compone. Estas frecuencias de palabras permiten, además de identificar los términos más utilizados, realizar comparaciones entre grupos de documentos definidos según algún criterio. En este trabajo se compararon frecuencias de términos entre los bloques de concejales, produciendo como resultado nubes de palabras, en las que cada término aparece con un tamaño de fuente proporcional a su frecuencia (Gráfico 1).

Entre las palabras más utilizadas por Cambiemos se observan nombres propios como “Mauricio Macri”, “José Corral” y “Federico Angelini”, términos relacionados al ámbito nacional y provincial y *slogans* de campaña. En el caso de CF se destaca la ausencia de nombres propios y el uso de términos como “leche”, “alimentos” o “tambo” que se asocian con los proyectos del bloque. En el FNyP se observan nombres propios relacionados con el escenario nacional relativos al partido (como “Cristina Fernández”) y a sus oponentes (como “Macri”) y otras palabras relacionadas al escenario local. Las palabras más frecuentes del FP están relacionadas con la campaña local y se destaca el término “gracias” por las publicaciones de agradecimiento luego de ganar las elecciones. El FSyP no tuvo campaña para intendente, por lo cual las palabras que se observan están relacionadas a su participación en el Concejo Municipal, destacándose también el uso del lenguaje inclusivo.

Gráfico 1. Nubes de palabras de publicaciones de cada bloque.



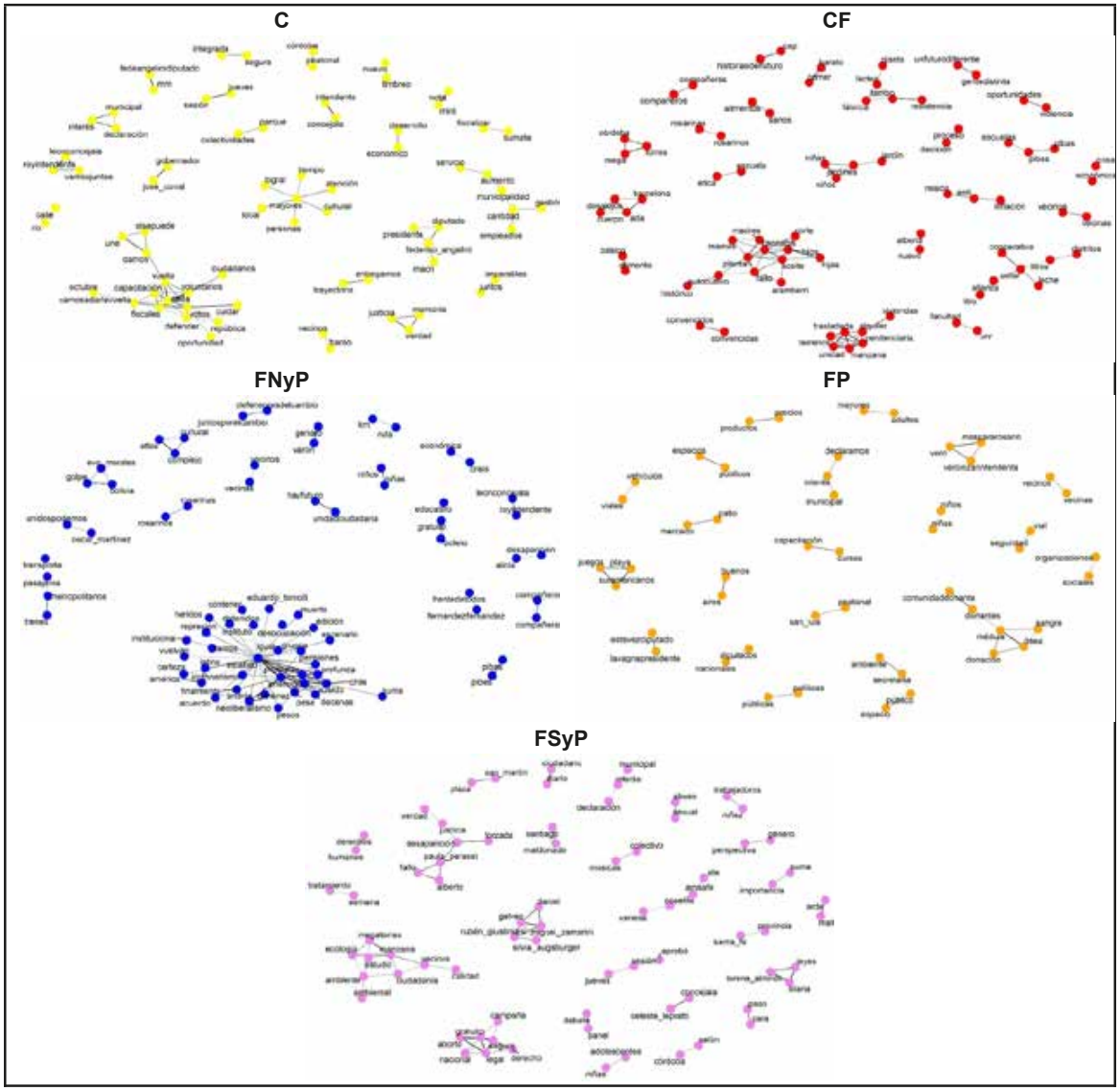
CORRELACIONES ENTRE PALABRAS

Para analizar el nivel de asociación existente entre cada par posible de palabras se calculó el coeficiente de Cramer para tablas de contingencia 2x2. Los resultados fueron explorados gráficamente mediante grafos, figuras que muestran nodos conectados con líneas, donde cada nodo es una palabra y las líneas que los unen representan las asociaciones entre las mismas. A mayor tonalidad en su trazo, mayor es la asociación; palabras no conectadas no están asociadas (Fradejas Rueda, 2020).

La utilización de grafos permite observar *clusters* entre las palabras más asociadas entre sí (Gráfico 2). En Cambiemos el *cluster* más grande contiene palabras y *hashtags* como “voluntarios”, “fiscales”, “VamosADarlaVuelta”, “SiSePuede”, etc., relacionados con las elecciones nacionales de octubre. Se observa otro *cluster* que contiene los *hashtags* “LeonConcejala” y “RoyIntendente” demostrando su campaña conjunta. En el caso de CF dos grandes *clusters* hacen referencia a proyectos del partido (creación de un parque de Vivienda Pública de Alquiler en Manzana 125) y a la revocación de un fallo a favor de la organización Madres Que Se Plantan. Se destaca también el uso de la misma palabra en ambos géneros (por ejemplo, niños y niñas). En el FNyP, la mayoría de las palabras en el grupo más grande tienen una connotación negativa, estando algunas relacionadas a los hechos acontecidos en Chile en 2019 y con críticas al gobierno vigente en Argentina. También aparecen grupos de palabras expresadas en ambos géneros. En el FP no se observan grandes grupos de palabras asociadas entre sí. Se pueden ver dos grupos diferenciados correspondientes a Pablo Javkin y a Verónica Irizar (candidatos a intendente en las elecciones primarias) y otro grupo relacionado a la campaña de donación de médula ósea incentivada por un concejal del bloque. Para el FSyP, uno de los principales grupos de palabras se relaciona con el proyecto de Manzana 125

nombrado anteriormente. Se pueden ver otros grupos relacionados con violencia de género y aborto, demostrando una perspectiva feminista.

Gráfico 2. Grafo de asociaciones de palabras para cada bloque



IDENTIFICACIÓN DE TÓPICOS

Asignación Latente de Dirichlet

Los modelos de tópicos buscan descubrir la estructura semántica subyacente de una colección de documentos. Permiten descubrir patrones de usos de palabras, conectan documentos que exhiben patrones similares y proporcionan un marco probabilístico para describir las frecuencias de los términos en documentos de un corpus (Blei *et al*, 2003; Blei, 2011).

Dentro de esta familia de modelos se encuentra la Asignación Latente de Dirichlet (ALD), que propone modelar documentos como si surgieran de diferentes tópicos, donde un tópico es definido como una distribución de probabilidad sobre un determinado conjunto de términos o

palabras (vocabulario). Se asume que existen K tópicos asociados con la colección de documentos, cada uno con un conjunto característico de palabras, y que cada documento se genera seleccionando primero una distribución de estos tópicos para su contenido y luego palabras provenientes de los mismos, de forma tal que cada documento exhibe estos tópicos en diferentes proporciones. Este supuesto es razonable ya que los documentos en un corpus son generalmente heterogéneos y combinan un subconjunto de ideas o temas.

Dado que estos tópicos no se conocen de antemano, el objetivo es encontrarlos a partir de los datos. Asumiendo que los tópicos son especificados antes de la generación de los documentos, el modelo ALD postula un conjunto de variables latentes: la distribución de tópicos en cada documento, la de palabras en cada tópico y la asignación de un tópico a cada palabra de un documento. A partir de las variables observadas (documentos y palabras), se estiman parámetros de interés que permiten describir la estructura de tópicos subyacentes.

Dado un corpus con M documentos, la ALD asume el siguiente proceso generativo para un documento $d = (w_{d1}, ..., w_{dN_d})$, que contiene N_d palabras w_{dj}, $j \in \{1, ..., V\}$ (V = cantidad total de términos en el vocabulario):

1. Se asume que φ_k, el vector de probabilidades de términos en el tópico k, tiene distribución Dirichlet: $\varphi_k \sim Dir(\beta)$.
2. Se asume que θ_d, el vector de probabilidades de tópicos en el documento d, tiene distribución Dirichlet: $\theta_d \sim Dir(\alpha)$.
3. Se determina N_d, la cantidad de términos en el documento d.
4. Para cada término j del documento d, w_{dj}:
    a. Se elige un tópico según: $z_{dj} \sim Multinomial(\theta_d)$. Z es el vector de dimensión $\sum_{d=1}^{M} N_d$ que reúne a las variables z_{dj}.
    b. Se elige del vocabulario un término w_{dj} según: $w_{dj} \sim Multinomial(\varphi_{z_{dj}})$. W es el vector de dimensión $\sum_{d=1}^{M} N_d$ que reúne a las variables w_{dj}.

El proceso generativo para la ALD se corresponde con la siguiente función de probabilidad conjunta de las variables ocultas y observadas:

$$p(\varphi_{1:K}, \theta_{1:M}, Z, W) = \prod_{k=1}^{K} p(\varphi_k) \prod_{d=1}^{M} p(\theta_d) \left(\prod_{d=1}^{M} \prod_{j=1}^{N_d} p(\theta_d) p(w_{dj} | \varphi_k, z_{dj}) \right)$$

Para estudiar la estructura oculta de tópicos en el corpus se plantea la correspondiente distribución a posteriori de las variables ocultas dados los M documentos observados:

$$p(W) = \frac{p(\varphi_{1:K}, \theta_{1:M}, Z, W)}{p(W)}$$

Para caracterizar esta distribución de interés se empleó el "muestreador" de Gibbs, un caso particular del algoritmo de Metropoli-Hastings ampliamente utilizado en estadística bayesiana para obtener muestras de la distribución a posteriori con fines inferenciales (Blei, 2011, pp.7-8).

Interpretación de los tópicos encontrados

En la selección del número K se priorizó la claridad en la interpretación de los tópicos originados, como suele ser sugerido en la literatura, optando por $K = 3$. A partir de las estimaciones de los vectores φ_k se observó cuáles fueron los términos más importantes en cada tópico (Gráfico 3) y a partir de esta información se eligió un nombre para cada uno de ellos con el fin de resumir el tema al que hacen referencia dichas palabras importantes. Estos son: "Género y Derechos Humanos" (Tópico 1), "Jerga de campaña" (Tópico 2) y "Propuestas políticas" (Tópico 3).

Además se determinó el tópico principal en cada mensaje como aquel que presentara mayor probabilidad de ocurrencia, empleando las estimaciones de los vectores θ_k. De esta forma fue posible estudiar la distribución del tópico principal dentro de cada bloque político. Se encontró que en los mensajes del FSyP predominó "Género y Derechos Humanos", mientras que el resto de los bloques presentó mayormente mensajes asociados a "Jerga de campaña" (Gráfico 4).

Por otro lado, se comparó la distribución del tópico principal en los mensajes de cada bloque durante el período preelectoral (febrero a junio) y el período postelectoral (julio a noviembre), observando que el porcentaje de publicaciones asociadas a "Jerga de campaña" disminuyó al pasar al segundo período para todos los casos con excepción de Cambiemos (Gráfico 5).

Gráfico 3. Palabras más importantes en cada tópico

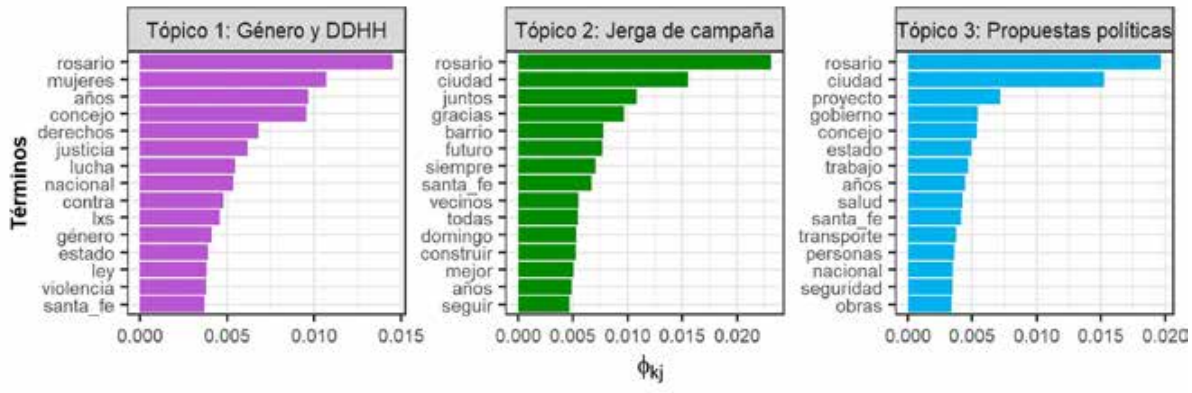


Gráfico 4. Distribución del tópico principal en los mensajes de cada bloque

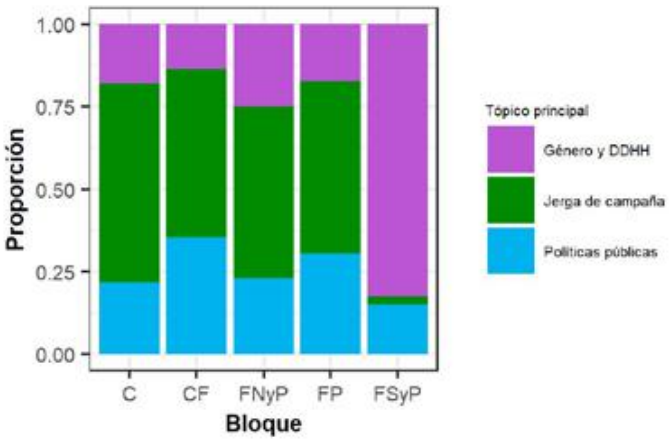
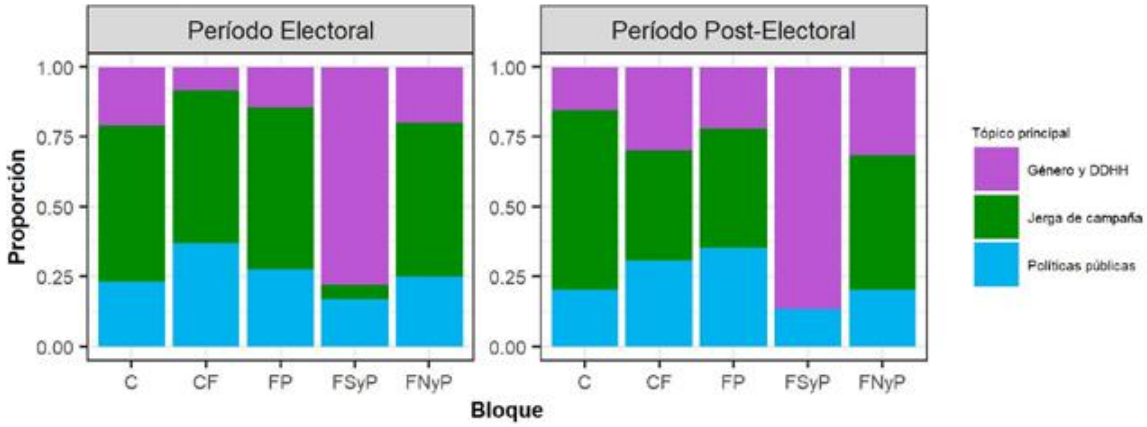


Gráfico 5. Distribución del tópico principal en los mensajes según bloque y período



COMENTARIOS FINALES

En este trabajo se realizó un estudio exhaustivo de las páginas de *Facebook* de los concejales de la ciudad de Rosario entre los meses de febrero y noviembre del año electoral de 2019. Los mensajes publicados fueron sometidos a un análisis textual con el objetivo de caracterizar el contenido de la comunicación de cada bloque político y de realizar comparaciones entre ellos. Se implementaron herramientas gráficas desarrolladas para datos textuales como nubes de palabras y grafos, se analizaron asociaciones entre palabras y se estudió el modelado de tópicos en base al método de ALD.

Con las herramientas para el análisis de frecuencias y de asociaciones entre términos fue posible distinguir, exploratoriamente, diferencias en la comunicación de los bloques. En Cambiemos se evidenció una mayor participación en la campaña nacional y provincial y el uso

frecuente de nombres propios correspondientes a candidatos o referentes del partido. La comunicación del FP se centró, en su mayoría, en la campaña de intendencia de Pablo Javkin y en sus propuestas. El FNyP mostró participación en las campañas a nivel local, provincial y nacional. El único bloque en el cual no se destaca el uso de nombres propios es CF y su comunicación se centró en propuestas y proyectos del partido. El FSyP por su parte, basó su comunicación en temáticas correspondientes al Concejo Municipal. El FSyP, el FNyP y CF mostraron uso de lenguaje inclusivo.

Mediante el ajuste del modelo de tópicos con ALD se logró identificar tres ejes temáticos en las comunicaciones; uno referido a género y derechos humanos, otro a propuestas de políticas públicas y el último a jerga de campaña. Los bloques se diferenciaron en cuanto a la distribución de sus publicaciones según tópicos. En base al modelo, en el FSyP la mayoría de los mensajes fueron referidos a género y derechos humanos, mientras que el resto de los bloques realizó publicaciones mayoritariamente de jerga de campaña. Queda propuesto como futura línea de investigación estudiar con mayor detalle el efecto de cambios que podrían aplicarse sobre la parametrización del modelo, el alcance de las medidas de bondad de ajuste disponibles y las ventajas o desventajas de analizar textos cortos como lo son las publicaciones de Facebook.

En el informe final de la práctica profesional que sirvió de marco para este trabajo (Rubio *et al.*, 2021) se incluyen resultados adicionales surgidos del análisis de numerosas variables referidas a las publicaciones de los concejales. A través del uso de métodos no paramétricos, fue posible concluir que todos los bloques disminuyeron el nivel de actividad y de interacciones durante el período post-electoral (julio a noviembre), con excepción de Cambiemos que se comporta de manera inversa. Además, se observó que los concejales de CF tuvieron un mejor desempeño en su comunicación a lo largo del año, presentando la mayor cantidad de seguidores en promedio y de *engagement* (reacciones y comentarios) por publicación. Por otro lado, estudiando la pauta publicitaria de las publicaciones se pudo observar que el *engagement* obtenido fue mayor en las comunicaciones publicitadas para todos los bloques.

REFERENCIAS

Blei, D., Ng, A., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993-1022.

Blei, D. (2011). Introduction to Probabilistic Topic Models. Princeton University.

Feinerer, I. & Hornik, K. (2020). tm: Text Mining Package. R package version 0.7-8. https://CRAN.R-project.org/package=tm.

Fradejas Rueda, J. M. (2020). Cuentapalabras: Estilometría y análisis de textos con R para filólogos. Recuperado de http://www.aic.uva.es/cuentapalabras/

Jones, T. (2019). textmineR: Functions for Text Mining and Topic Modeling. R package version 3.0.4. URL https://CRAN.R-project.org/package=textmineR

Silge, J., & Robinson, D. (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." _JOSS_, *1*(3). doi: 10.21105/joss.00037 (URL: https://doi.org/10.21105/joss.00037).

Rubio, L., Prunello, M., Crucella, J. (2021). Análisis de redes sociales de los integrantes del Concejo Municipal de Rosario en el año 2019. Recuperado de https://www.fcecon.unr.edu.ar/web-nueva/practicas-profesionales-aprobadas

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686.


CONSEJO PROFESIONAL
DE CIENCIAS ECONOMICAS
DE LA PROVINCIA DE SANTA FE
CAMARA II
CPCE
Colegio de Graduados
en Ciencias Económicas
de Rosario
UNR
Universidad
Nacional
de Rosario
340
280
120
50
70
120
june
july
aug.
sep.
nov.
dec.
jan.