FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

DEPARTAMENTO DE ESTADÍSTICA
EYP3417 - Estadística Espacial

Efectos Comunales en la Participación Electoral en Chile

Eduardo Vásquez - Juan Pino

Profesor: Fernando Quintana
Ayudante: Rubén Soza

Diciembre del 2021

Introducción

La participación electoral se define como el porcentaje de personas con derecho a voto, los cuales no necesariamente son todos los adultos de la población, que efectivamente votaron dentro de una elección. Esta participación es fundamental para la democracia, ya que es lo que entrega legitimidad al sistema político y a los integrantes que la componen.

Pese a lo anterior, se ha observado que la tasa de participación electoral ha ido disminuyendo en los últimos 25 años (Ríos, Madrid, and Sacks 2017). En nuestro país, por ejemplo, en las elecciones parlamentarias del año 1990 el porcentaje de votos fue de 86.9%, mientras que en el 2016 fue de un 50.9%, una baja considerable de 36 puntos. No solo eso, sino que además el año 2015 Chile fue el país con el mayor porcentaje de abstinencia dentro de los países con voto voluntario.

Pese a lo anterior, el año 2020 se pudo observar un alza importante en la participación política, donde en el plebiscito para una nueva constitución votó cerca del 50.9% del padrón electoral, lo cual nos llevó a pensar que ocurrió un cambio en cómo la gente se relaciona con las elecciones. Pese a esto, en las elecciones para gobernadores regionales del año 2021 sólo participó un 19.6% del padrón electoral en la segunda vuelta, uno de los porcentajes de participación mas bajos en la historia de nuestro país.

En el presente trabajo estudiaremos las variables que influyen en esta tasa de participación en la ciudad de Santiago, segmentando por comuna, así como estudiar la asociación que existe entre comunas vecinas a través de efectos espaciales. Los datos electorales fueron obtenidos del Servicio Electoral de Chile (SERVEL) para las elecciones presidenciales del año 2017. Por otro lado, los datos comunales fueron obtenidos de la Encuesta Casen 2017, de la Biblioteca del Congreso Nacional de Chile y del Sistema Nacional de Información Municipal.

De manera específica, y dada la naturaleza de los datos, trabajaremos con modelos lineales generalizados mixtos para datos areales, donde introducimos los efectos espaciales por comuna en la segunda etapa de la especificación, a partir de modelos CAR en diferentes versiones. Para esto usaremos el paquete `CARBayes` en R, con el cual podemos definir diferentes prioris para los efectos espaciales.

Datos y Análisis Exploratorio

La construcción de los datos a utilizar en el presente informe está compuesta de tres partes:

- Polígonos de las comunas de Santiago, obtenidas de la Biblioteca del Congreso Nacional.
- Número de inscritos y número de votos para las elecciones presidenciales del año 2017, obtenidas del Servicio Electoral de Chile.
- Datos comunales: ingreso medio (en miles de pesos), porcentaje de gasto en educación y salud, y los porcentajes de pobreza, hogares hacinados y personas sin servicios básicos, obtenidas de la Encuesta Casen 2017, la Biblioteca del Congreso Nacional y el Sistema Nacional de Información Municipal.

Para nuestro trabajo, nuestro interés principal está en la tasa de participación, la cual se define como la razón del número de votantes sobre el número de inscritos. En la Figura 1 se presenta la tasa de participación en la primera vuelta de las elecciones presidenciales del año 2017, para cada una de las comunas de la capital. A partir de la figura es posible ver que las comunas suelen tener vecinos con tasas de participación similares, sobre todo las comunas del sector oriente como Lo Barnechea, Las Condes y Vitacura, las cuales tienen una tasa de participación mayor al 65%.

De la misma manera, podemos ver algunas de nuestras covariables en cada comuna. En la Figura 2 se presenta el ingreso medio por comuna, el gasto en educación, el porcentaje de pobreza y el porcentaje de hacinamiento. En este caso la presencia de autocorrelación no parece evidente, así como correlación con la tasa de participación. Pasamos entonces a estudiar de manera más formal la presencia de autocorrelación y clusters para la tasa de participación en la siguiente sección.

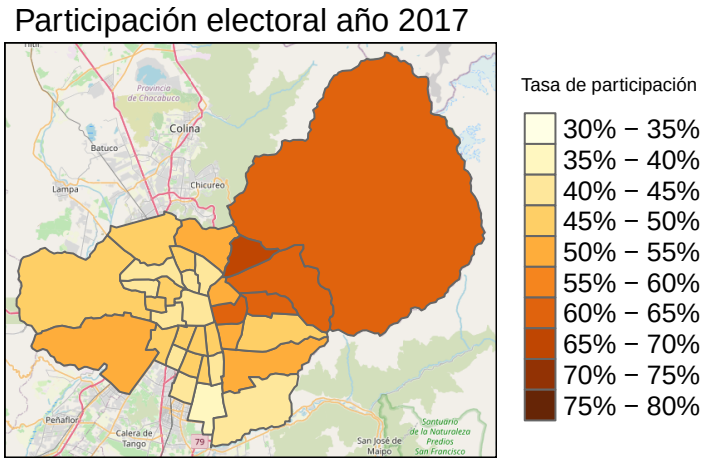

Figura 1: Tasa de participación electoral para las comunas de Santiago en la segunda vuelta de las elecciones presidenciales del año 2017.

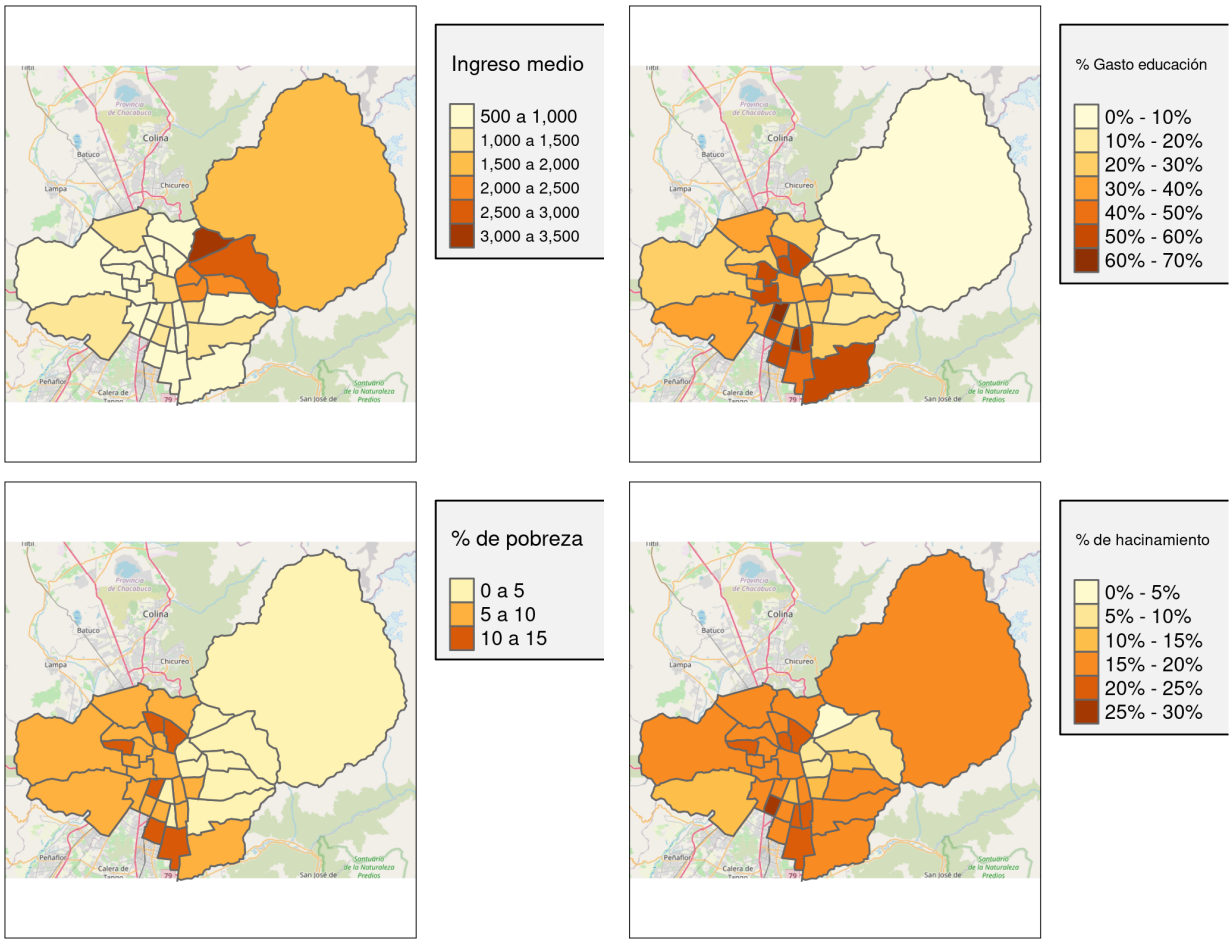

Figura 2: Covariables dentro de cada comuna: ingreso medio (en miles de pesos), gasto en educación, porcentaje de pobreza y porcentaje de hacinamiento

Análisis de Autocorrelación Espacial

Como pudimos observar en el análisis exploratorio de los datos, es posible notar que existe autocorrelación espacial, esto es, las comunas suelen tener características similares a sus vecinos, tanto para la tasa de participación como para algunas de las covariables. Para el estudio formal de autocorrelación veremos el coeficiente I de Moran y el coeficiente C de Geary como medidas globales, e introduciremos una medida local de autocorrelación denominada estadísticos de Getis-Ord, la cual es utilizada para identificar clusters. Decidimos incluir una medida local ya que, como se puede apreciar de la Figura 3, cada una de las comunas tiene al menos 3 vecinos, por lo que tenemos información suficiente para el estudio por separado.

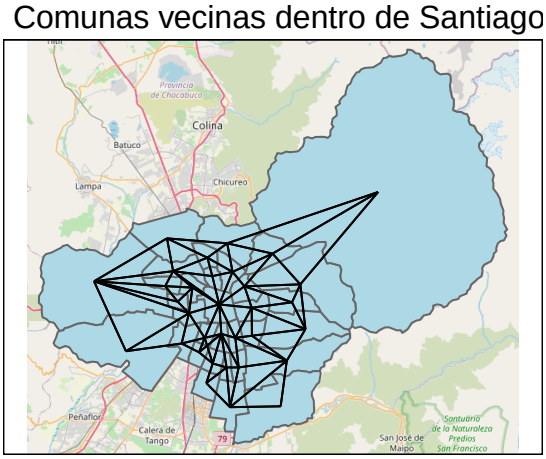

Figura 3: Mapa de vecinos para la provincia de Santiago, utilizando una matriz de adyacencia Queen

Coeficiente I de Moran y Coeficiente C de Geary

En la Figura 3 se presenta de manera conjunta la densidad aproximada del coeficiente I de Moran (arriba) y el coeficiente C de Geary (abajo), obtenidas a partir de permutaciones en los datos, junto con el valor observado. Vemos que en ambos casos el valor observado es muy lejano, por lo que tenemos evidencia de que existe autocorrelación espacial positiva para la tasa de participación en el año 2017, dado que el valor obtenido es mayor a 0 en el caso de Moran y menor a 1 en el caso de Geary.

Coeficiente de Getis-Ord

El coeficiente de Getis-Ord (Getis and Ord 1992) fue propuesto para estudiar la autocorrelación espacial local. Existen dos versiones, una que toma en cuenta el valor en la localización, denominada G_i^*, y otra que no, denominada G_i.

La definición específica está dada por:

$$G_i = \frac{\sum_{j \neq i} w_{ij} x_j}{\sum_{j \neq i} x_j} \quad \text{y} \quad G_i^* = \frac{\sum_j w_{ij} x_j}{\sum_j x_j}$$

esto es, es una razón de la suma ponderada de los valores vecinos (e incluyendo a sí mismo en el caso de G_i^*) sobre la suma de todos los valores.

Un valor alto de G más alto que la media sugiere un cluster de valores altos, denominado normalmente como High-High o "hot spot," mientras que un valor más bajo que la media sugiere un cluster de valores bajos, denominado Low-Low o "cold spot."

De manera similar a el coeficiente de Moran y Geary, se sugiere hacer inferencia a partir de permutaciones de los datos, obteniendo el coeficiente para cada permutación y finalmente comparar con el valor obtenido. Para

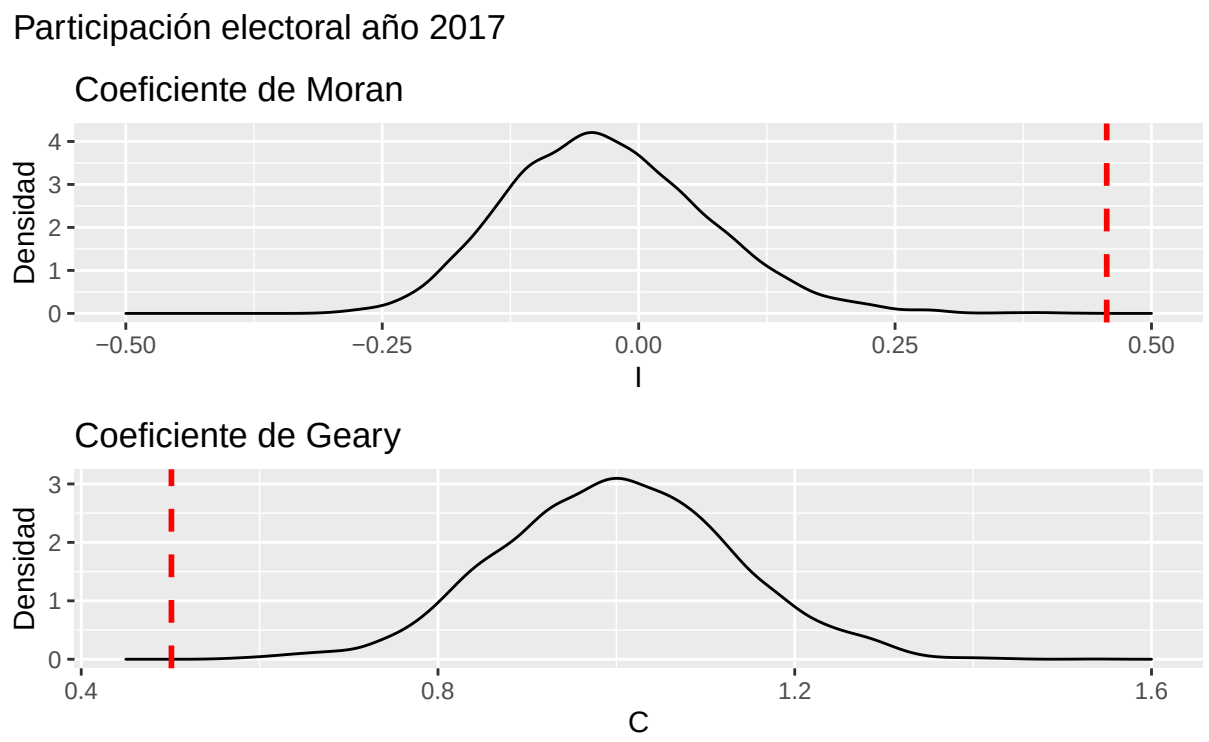

Figura 4: Densidad aproximada del coeficiente I de Moran y C de Geary, elecciones año 2017, junto al valor observado (en celeste)

el cálculo del estadístico se utilizó el paquete `rgeoda` (Li and Anselin 2021), el cual nos entrega las regiones con un coeficiente significativo, indicando además si son hot spots o cold spots.

En la Figura 5 se presenta el mapa con las comunas indicadas como hot spot de rojo y las indicadas como cold spot de azul. Las comunas sin color significa que no tenemos suficiente evidencia para concluir la presencia de clusters.

A partir de la figura vemos que, como habíamos notado en la sección de análisis exploratorio, existe un cluster de comunas con alta tasa de participación correspondiente a las comunas del sector oriente de la capital: Lo Barnechea, Vitacura, Las Condes, Providencia, La Reina y Peñalolén, agregando también la comuna de Huechuraba. También vemos que hay dos clusters de tasa baja de participación, una compuesta por las comunas de Independencia, Renca, Quinta Normal y Lo Prado, mientras que la otra está compuesta de San Miguel, San Ramón, Pedro Aguirre Cerda, El Bosque, La Pintana y La Florida.

Modelamiento

Para modelar nuestros datos, usaremos el paquete `CARBayes` (Lee 2013), el cual nos permite ajustar modelos lineales mixtos generalizados, a partir del punto de vista bayesiano, en el cual los efectos aleatorios corresponden a las áreas de estudio, que en este caso son nuestras comunas. El paquete nos entrega diferentes prioris para estos efectos, con las cuales trabajaremos a continuación.

Específicamente, la librería ajusta, de manera general, el siguiente modelo:

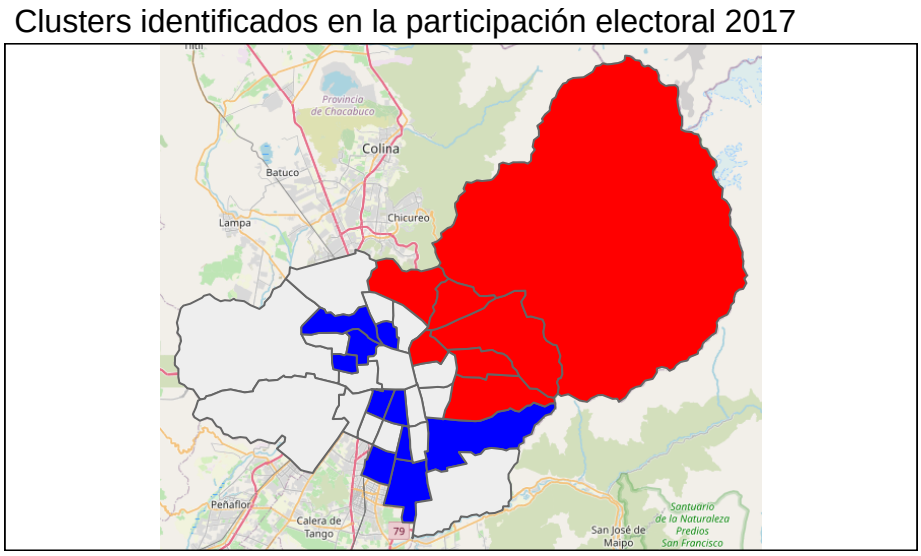

Figura 5: Clusters identificados por los estadísticos de Getis y Ord, en rojo para valores altos y en azul para valores bajos

$$
\begin{aligned}
Y_k|\mu_k &\sim f(y_k|\mu_k, \nu^2), \quad k = 1, ..., K \\
g(\mu_k) &= x_k^T\beta + O_k + \psi_k \\
\beta &\sim \mathrm{N}(\mu_\beta, \Sigma_\beta) \\
\nu^2 &\sim \text{Inv-Gamma}(a, b)
\end{aligned}
$$

donde K es el número de áreas (comunas), $\mathbf{Y} = (Y_1, ..., Y_k)$ el vector de respuestas (cantidad de votos), O_k las exposiciones (específicamente, el log de éstos). Además, $\mathbf{x}_k$ corresponde a las covariables para la k-ésima área, $E(Y_k) = \mu_k$ y $\beta = (\beta_1, ..., \beta_p)$ el vector de parámetros de la regresión.

La variable respuesta puede ser modelada con cuatro distribuciones: Binomial, Normal, Poisson y Poisson cero-inflada. Como nuestros datos corresponden a conteos con una cantidad máxima posible (la cantidad de electores inscritos), usaremos la distribución binomial, esto es, $Y_k \sim \text{Binomial}(n_k, \theta_k)$, con $\log(\theta_k/(1-\theta_k)) = \mathbf{x}_k^T\beta + O_k + \psi_k$.

Modelo sin efectos especiales

La librería `CARBayes` también nos permite ajustar un modelo lineal generalizado, sin incluir los efectos espaciales, esto es, $\psi_k = 0$ para todo k. Ajustamos este modelo para así tener una base de referencia al momento de ajustar los otros modelos que sí incluyen los efectos.

Modelo de Besag-York-Mollie

En este caso, los efectos espaciales están modelados como:

$$
\begin{aligned}
\psi_k &= \phi_k + \theta_k \\
\phi_k|\phi_{-k}, \mathrm{W}, \tau^2 &\sim \mathrm{N}\left(\frac{\sum_{i=1}^K w_{ki}\phi_k}{\sum_{i=1}^K w_{ki}}, \frac{\tau^2}{\sum_{i=1}^K w_{ki}}\right) \\
\theta_k &\sim \mathrm{N}(0, \sigma^2) \\
\tau^2, \sigma^2 &\sim \text{Inv-Gamma(a, b)}
\end{aligned}
$$

donde W es la matriz de adyacencia que obtuvimos anteriormente, y $\theta = (\theta_1, ..., \theta_K)$ son efectos aleatorios independientes y los efectos espaciales son modelados por $\phi = (\phi_1, ..., \phi_K)$. Es importante notar que este modelo no es identificable en θ_k y ϕ_k, pero sí en ψ_k.

Modelo de Leroux

Este modelo es similar al anterior, pero se agrega un parámetro ρ de la siguiente manera:

$$\begin{aligned}
\psi_k &= \phi_k \\
\phi_k|\phi_{-k}, \mathrm{W}, \tau^2 &\sim \mathrm{N}\left(\frac{\rho\sum_{i=1}^{K} w_{ki}\phi_i}{\rho\sum_{i=1}^{K} w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{i=1}^{K} w_{ki} + 1 - \rho}\right) \\
\tau^2 &\sim \text{Inv-Gamma}(a, b) \\
\rho &\sim \text{Uniforme}(0, 1)
\end{aligned}$$

Notar que cuando $\rho = 1$ se obtiene el modelo ICAR visto en clases, mientras que con $\rho = 0$ hay independencia espacial.

Modelo localised

Los dos modelos anteriores realizan un suavizamiento espacial global. El problema con estos modelos es que pueden simplificar mucho el problema si es que la región no pareciera comportarse de manera uniforme. Por ejemplo, de la Figura 1, parece ser que existe una correlación alta entre las comunas del sector oriente de la capital, pero no pareciera ser correcto asumir los mismo niveles de correlación en otros sectores de la capital.

El modelo localised es utilizado para capturar autocorrelación espacial localizada. La forma en que hace esto es aumentando el conjunto de efectos espaciales ϕ con interceptos constantes, particionando las K áreas en un máximo de G clusters, cada uno con su propio intercepto λ_g.

En nuestro caso, tomaremos $G = 4$ dado lo observado en la Figura 5, ya que identificamos tres clusters en ese caso, y dejaremos abierta la posibilidad de un cuarto cluster.

No incluiremos los detalles de este modelo en el informe, ya que lo alargaríamos en demasía, pero los detalles se encuentran en (Lee and Sarran 2015).

Resultados

Utilizando la librería ajustamos los cuatro modelos propuestos. Cada uno de los modelos considera de manera común los parámetros de la regresión, y luego se van diferenciando en los otros que van incluyendo.

Sobre los resultados obtenidos, tenemos que solo para el modelo sin efectos espaciales tenemos evidencia de convergencia a la distribución estacionaria, con valores de Geweke entre -1.96 y 1.96, y valores de $\hat{R}$ bajo el 1.01 (al menos de manera puntual). En este caso, fue necesario un periodo de quema de 20000 iteraciones para obtener la convergencia.

Por otro lado, en ninguno de los tres modelos que incluyen efectos espaciales tenemos evidencia de convergencia. En todos los casos probamos hasta con 1 millón de iteraciones para el periodo de quema, sin obtener resultados positivos. Intentamos también usar JAGS y Stan, sin obtener mejores resultados. Un ejemplo de los valores obtenidos se presenta en la Tabla 1, que entrega los valores de $\hat{R}$ para el vector β en el modelo de Besag-York-Mollie.

En lo que queda de esta sección asumiremos convergencia de nuestras cadenas, por lo que los resultados y conclusiones que siguen no son correctos (o al menos no sabemos si realmente lo son).

Para poder elegir entre los cuatro modelos, podemos utilizar el Criterio de Información de la Devianza y el Criterio de Información de Watanabe-Akaike, entregados por el paquete utilizado. En la Tabla 2 se presentan los cuatro DIC y WAIC obtenidos.

Tabla 1: Valores del factor de reducción de escala potencial para los parámetros de regresión en el modelo de Besag-York-Mollie

	Estimación puntual	Límite superior IC
Intercepto	2.161257	4.531872
Ingreso medio	1.650889	3.129096
Pobreza	1.626397	2.666483
Sin Servicios Básicos	1.741212	3.110491
Hacinamiento	3.281798	7.435879
Gasto en salud	1.353185	1.986123
Gasto en educación	1.133752	1.362560

Tabla 2: Valores de DIC y WAIC para cada uno de los modelos ajustados

	Sin efectos	BYM	Leroux	Localised
DIC	18757.24	470.5418	468.5580	470.5296
WAIC	24290.64	460.3300	458.2042	463.4375

A partir de la tabla es posible ver que el modelo sin efectos espaciales tiene un DIC y WAIC mucho mayor que los demás. Entre los que incluyen efectos espaciales, vemos que el modelo propuesto por Leroux tiene tanto un DIC como un WAIC menor que los demás, por lo que concluimos que éste es el "mejor" modelo.

En la Tabla 3 se presenta la mediana y un intervalo de credibilidad de 95% para cada uno de los parámetros ajustados en el modelo.

A partir de la Tabla 3 podemos ver que los intervalos de credibilidad del intercepto, porcentaje de viviendas sin servicios básicos, porcentaje de hogares hacinados, gasto municipal en salud y gasto municipal en educación contienen al 0, por lo que tenemos evidencia que aquellas variables no son significativas para el estudio de la tasa de participación electoral.

Por otro lado, las covariables que sí son significativas son el ingreso medio y el porcentaje de pobreza, donde el logit del parámetro aumenta en 0.03 por cada cien mil pesos que aumenta el ingreso medio, y disminuye a medida que aumenta la tasa de pobreza.

En cuanto al valor de ρ, que nos entrega el nivel de suavizamiento espacial, vemos que el intervalo de credibilidad es bastante ancho entre el 0.0077 y 0.7499. Por su parte, el parámetro τ^2, que entrega de cierta manera el nivel de variabilidad global, tiene una estimación de 0.0307.

Tabla 3: Inferencia a posteriori de los parámetros en el modelo de Leroux

Parámetro	Mediana	Límite inferior IC	Límite superior IC
Intercepto	-0.1585	-0.3661	0.1137
Ingreso medio	0.0003	0.0002	0.0004
% Pobreza	-0.0230	-0.0318	-0.0024
% sin Servicios Básicos	-0.0020	-0.0147	0.0070
% Hacinados	-0.0039	-0.0119	0.0117
Gasto salud	0.2892	-0.2127	0.6203
Gasto educación	-0.2719	-0.6375	0.0640
tau	0.0307	0.0157	0.0844
rho	0.1500	0.0077	0.7499

Conclusión

Dado los resultados anteriores, la conclusión es incierta. En caso que sí se pueda concluir convergencia a la distribución estacionaria, tendríamos que el modelo a utilizar sería el propuesto por Leroux, al obtener valores de DIC y WAIC menores, como vimos en la sección anterior.

De este modelo obtuvimos que hay dos covariables para las cuales tenemos evidencia que afectan la tasa de participación electoral: el ingreso medio de la población y el porcentaje de pobreza, donde el primero afecta de manera positiva, mientras que el segundo de manera negativa. Creemos que puede ser sensible lo que se pueda concluir a partir de ésto, considerando que nosotros no entendemos bien los mecanismos sociales que subyacen a la participación política y democrática. Como se menciona en (Ríos, Madrid, and Sacks 2017), algunas de las razones por las que existen tasas de participación bajas son: debilitamiento del sistema de representación, rol de los partidos políticos, declive en la percepción de la eficacia política y transformaciones en el mundo juvenil. Sería interesante entonces tratar de generar encuestas que se enfoquen en estas variables dentro de la población, las cuales no se encuentran disponibles actualmente.

En cuanto al análisis espacial, pudimos ver en la sección de Análisis de Autocorrelación Espacial que existen similitudes entre comunas que son vecinas entre sí, por lo que incluir los efectos espaciales logra enriquecer el modelo con esta información que entrega.

Una de nuestras intenciones iniciales era trabajar con modelos espacio-temporales, pero las bases de datos del SERVEL no se encuentran bien optimizadas para obtener datos de antes del 2013, por lo que lo único que pudimos conseguir fueron los resultados del 2013 y del 2017. Además, los espacios de tiempo entre elecciones son grandes, por lo que el efecto temporal puede ser tenue en comparación con el efecto de covariables y el efecto espacial. Un paso natural sería incluir los datos que se obtengan este domingo en las elecciones presidenciales entre Gabriel Boric y José Antonio Kast.

Por último, existe un nuevo tipo de modelo denominado DAGAR (Datta, Banerjee, and Hodges 2017) el cual puede ser interesante de aplicar como otro paso siguiente al presente informe. Este modelo lo que hace es que, en vez de modelar la matriz de varianzas-covarianzas directamente, modela el factor de Cholesky de esta, dándole un carácter dirigido.

Referencias

Anselin, Luc. 2016. "Spatial Autocorrelation." GeoDa Software. 2016. https://www.youtube.com/playlist?list=PLzREt6r1NennT20oeK46QOloT2k9QM_xB.

Banerjee, Sudipto. 2014. *Hierarchical Modeling and Analysis for Spatial Data*. 2nd ed.. Monographs on Statistics and Applied Probability (Series) ; 135.

Datta, Abhirup, Sudipto Banerjee, and James S Hodges. 2017. "Spatial Disease Mapping Using Directed Acyclic Graph Auto-Regressive (DAGAR) Models."

Getis, Arthur, and J. K. Ord. 1992. "The Analysis of Spatial Association by Use of Distance Statistics." *Geographical Analysis* 24 (3): 189–206. https://doi.org/https://doi.org/10.1111/j.1538-4632.1992.tb00261.x.

Lee, Duncan. 2013. "CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors." *Journal of Statistical Software* 55 (13): 1–24. https://www.jstatsoft.org/htaccess.php?volume=55&type=i&issue=13.

Lee, Duncan, and Christophe Sarran. 2015. "Controlling for Unmeasured Confounding and Spatial Misalignment in Long-Term Air Pollution and Health Studies." *Environmetrics (London, Ont.)* 26 (7): 477–87.

Li, Xun, and Luc Anselin. 2021. *Rgeoda: R Library for Spatial Data Analysis*. https://CRAN.R-project.org/package=rgeoda.

Ríos, M., S. Madrid, and S. Sacks. 2017. "Diagnóstico Sobre La Participación Electoral En Chile." Santiago, Chile: Programa de las Naciones Unidas para el Desarrollo (PNUD). https://www.cl.undp.org/content/chile/es/home/library/democratic_governance/diagnostico-sobre-la-participacion-electoral-en-chile.html.

Tennekes, Martijn. 2018. "tmap: Thematic Maps in R." *Journal of Statistical Software* 84 (6): 1–39. https://doi.org/10.18637/jss.v084.i06.