UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

TRABAJO FINAL DE GRADO PARA OBTENER EL TÍTULO DE LICENCIADO EN ESTADÍSTICA

Predicción de los flujos faltantes de la matriz de Origen - Destino de la Encuesta de Movilidad del departamento de Montevideo utilizando técnicas de filtrado espacial.

Antonio Rey

Tutores:
María Eugenia Riaño
Fernando Massa

Montevideo
URUGUAY
2021

UNIVERSIDAD DE LA REPÚBLICA

El tribunal docente integrado por los abajo firmantes aprueba el trabajo final de grado:

Predicción de los flujos faltantes de la matriz de Origen - Destino de la Encuesta de Movilidad del departamento de Montevideo utilizando técnicas de filtrado espacial.

Antonio Rey

Tutores:
María Eugenia Riaño
Fernando Massa

Licenciatura en Estadística

Puntaje .....

Tribunal

Profesor Dr. Marco Scavino .....

Profesor Dr.(ver) Leonardo Moreno .....

Profesora María Eugenia Riaño .....

Fecha .....

Agradecimientos

En primer lugar, agradezco a mis tutores Eugenia Riaño y Fernando Massa, por todo el apoyo, dedicación y paciencia brindada durante este tiempo.

En segundo lugar, agradezco al Instituto de Estadística por darme la oportunidad de trabajar como Ayudante de Iniciación a la Investigación. También a la Agencia Nacional de Investigación e Innovación y a la Comisión Sectorial de Investigación Científica por el apoyo financiero a los proyectos en los cuales se enmarca este trabajo.

Por último, agradezco a Orlando Sabogal, por las recomendaciones en cuanto a los datos disponibles para la realización de este trabajo y a Andrés Castrillejo, por toda su colaboración en el armado de los programas de R.

Resumen

Las encuestas de origen- destino, por lo general se realizan con el objetivo de estimar los flujos de transporte entre distintas regiones de una ciudad. Puede suceder, que si las unidades geográficas consideradas como regiones de transporte son pequeñas, el tamaño muestral no sea suficiente y se tengan celdas vacías en la matriz de origen- destino (OD). Si bien se entiende que el tamaño de estos flujos puede ser pequeño (ya que no fue captado por la encuesta), es necesario contar con una estimación para una adecuada planificación del transporte. Contar con una matriz OD completa es fundamental para el análisis de la movilidad urbana y de las necesidades de transporte de la población. Tradicionalmente los flujos de las matrices OD se estiman con modelos gravitacionales, herramienta que también se utiliza para la predicción de las celdas vacías de la matriz OD.

El objetivo del presente trabajo es realizar una imputación para los flujos faltantes de la matriz OD de los viajes por trabajo del departamento de Montevideo, considerando como unidad geográfica de análisis a los Centros Comunales Zonales (CCZ). Para realizar la imputación se comparara el desempeño de los modelos gravitacionales que incorporan la autocorrelación espacial en la imputación de datos faltantes de la matriz OD, con los modelos que no consideran la autocorrelación espacial. La autocorrelación espacial se incorpora utilizando filtros espaciales, obtenidos del conjunto de vectores propios de una apropiada transformación de la matriz de pesos espaciales.

El filtrado espacial parte del supuesto de que la correlación espacial existe debido a una mala especificación del modelo, es decir, existe un conjunto de variables omitidas, que hacen que los errores del modelo presenten autocorrelación espacial. Obteniendo los vectores propios de una transformación de la matriz de pesos espaciales, se obtiene un conjunto de variables "proxy" que al ser incorporadas al modelo, hacen que los residuos del modelo no presenten autocorrelación espacial. La incorporación de los vectores propios se realiza mediante un proceso similar al "forward", incorporando vectores hasta que la autocorrelación espacial medida con el Índice de Moran, sea eliminada de los residuos del modelo.

Para la evaluación de desempeño se cuenta con la matriz origen-destino completa para la ciudad de Bogotá obtenida de la Encuesta de Movilidad de Bogotá realizada en el año 2015. Se simulan distintas cantidades de datos faltantes, teniendo en cuenta el tamaño del flujo original y aumentando la cantidad de celdas vacías de la matriz.

Tomando a la Raíz del Error Cuadrático Medio de la predicción como medida de desempeño, los resultados demuestran una ganancia de hasta el 71 % cuando se trata de flujos pequeños (el escenario más probable), demostrando que la importancia de incorporar de la autocorrelación espacial en la predicción de los flujos faltantes de la matriz.

Los datos a imputar provienen de la Encuesta de Movilidad en el Área Metropolitana de Montevideo. La matriz OD presenta 85 celdas vacías, un 29.4 % del total de sus valores. Luego de realizada la imputación se logró reducir la cantidad de ceros a un 10 %.

Palabras Clave: Estadística espacial, filtrado espacial, imputación, matrices Origen-Destino

Tabla de Contenidos

Índice de figuras

Índice de tablas

Capítulo 1. Introducción

Los modelos de transporte son una herramienta de análisis para la movilidad de una ciudad, constituyen un insumo fundamental para la toma de decisiones en la planificación del transporte en diversas ciudades del mundo. En particular, se suelen usar este tipo de modelos para analizar los impactos que tienen las intervenciones en los sistemas de transporte, principalmente en la infraestructura, pero también a partir de modificaciones en las políticas regulatorias, así como para analizar cambios en la demanda. A partir de estos modelos se puede mejorar las condiciones de movilidad, repercutiendo en la calidad de vida de los ciudadanos.

La matriz de Origen - Destino (OD) es una tabla de contingencia de dos dimensiones, en donde los valores de las celdas representan el volumen de tráfico[1] entre un conjunto de regiones predeterminado. En la matriz OD se observa:

- La magnitud del volumen del tráfico.
- El patrón de Origen - Destino.
- El total de la producción y atracción de viajes.

Cuando se tienen celdas con frecuencia cero en la matriz OD, se deben diferenciar dos situaciones: los denominados "ceros estructurales" y los "ceros muestrales" (Ten Have, 2005). El cero estructural se debe a la imposibilidad de observar una combinación dada de las categorías de una tabla de contingencia. En el caso por ejemplo de la matriz OD de transacciones de bienes y servicios de comercio internacional pueden existir celdas vacías, ya que no todos los países mantienen relaciones comerciales entre sí.

Los ceros muestrales se deben a que en la muestra no se obtuvieron determinadas frecuencias. El cero muestral no implica que el flujo no observado no exista, sino que debido al mecanismo de muestreo utilizado no se observaron realizaciones de algunos flujos en particular.
Los modelos estadísticos a utilizar para realizar inferencias a partir de la matriz OD son diferentes de acuerdo a la cantidad y al origen de las celdas vacías de la matriz. En el caso de ceros estructurales se utilizan modelos de conteo Cero - Inflados (Cameron y Trivedi, 2013) con distribución Poisson o Binomial Negativa. Además, en el caso de que exista un exceso de

[1] Se refiere a tráfico en términos genéricos, pueden ser viajes, flujos migratorios, flujos de capitales, etc.

ceros, los modelos de conteo tipo Poisson o Binomial Negativa no cumplen con los supuestos distribucionales para su correcta modelación, situación que también se resuelve utilizando un modelo de conteo Cero - Inflado.

Si los ceros son muestrales, el total de celdas vacías de la matriz OD no debería reflejar un problema de exceso de ceros. Si así fuera, el problema radicaría en una regionalización no acorde con el tamaño de muestra. Los flujos de las matrices OD no son uniformes, y la mayoría de los viajes se encuentran concentrados en las regiones más atractivas de una ciudad. Aún utilizando una estructura de regiones acorde con el tamaño de muestra, pueden haber celdas con frecuencia cero. En general los modelos de conteo con distribución Poisson o Binomial Negativa son adecuados para esta situación. (Cameron y Trivedi, 2013)

Los métodos encontrados en la literatura para estimar los valores faltantes de las matrices OD (Wilson, 1967; Jou, Cho, Lin, y Wang, 2006) son los mismos que se utilizan para estimar las celdas de la matriz cuando sólo se cuenta con la información de las marginales de la tabla. La información de las marginales de la tabla es menos costosa que la información a nivel de celdas, por lo que en muchas aplicaciones se parte de esta información. Aún cuando se tiene información a nivel de celdas, como en las encuestas de Origen - Destino, se utilizan estos métodos, con el argumento de que las estimaciones a nivel de las marginales son más precisas que a nivel de las celdas de la tabla. Para esto, se utilizan métodos determinísticos como el de ajuste biproporcional (Furness, 1970), o de máxima entropía (Wilson, 1967) o modelos probabilísticos como los modelos gravitacionales. Los métodos antedichos comparten el supuesto de independencia de los flujos de la matriz OD: se asume que los viajes entre el par i,j son independientes de los viajes entre el par k, l. En el caso de matrices OD de viajes por trabajo, este supuesto no se cumple en la mayoría de las aplicaciones. Las fuentes de empleo tienden a presentar una correlación espacial positiva fuerte, con zonas de alta concentración de empresas, lo que genera un efecto "derrame" en las zonas aledañas, en cuanto a la localización de las fuentes de empleo. Los modelos propuestos para la estimación de las matrices OD en el caso de viajes por trabajo deberían incorporar de alguna manera la correlación espacial existente entre las regiones de la ciudad.

La inclusión de efectos espaciales en la estimación de modelos de Origen - Destino es de reciente aplicación. En el caso particular de viajes por trabajo destacan los artículos de Daniel Griffith. En su primer artículo trabaja con datos de 24 áreas urbanas de Canadá (Griffith y Jones, 1980), ajustando un modelo gravitacional con restricciones e incorporando autocorrelación espacial. La discusión en este caso se enfoca en la importancia de considerar la autocorrelación en la estimación del modelo para la movilidad interurbana de viajes por trabajo. En su segunda publicación, el autor analiza el caso de los viajes por trabajo para 469 distritos de Alemania, con información recolectada en el año 2002. Estima un modelo Poisson con restricciones y filtrado espacial, encontrando un mejor ajuste en el modelo que incluye variables espaciales

(Griffith, 2009).

Los trabajos antedichos y la literatura en general sobre modelos de interacción espacial se concentran en estudiar las distintas alternativas para la modelización de los efectos espaciales y en analizar su efecto sobre la estimación de los parámetros del modelo. En el presente trabajo se propone comparar el desempeño de los modelos tradicionales utilizados para el análisis de matrices OD con los modelos que consideran la correlación espacial, en la predicción de los flujos faltantes de la matriz. Si bien los modelos de conteo adminten ceros en su variable de respuesta, la estimación de los flujos faltantes es un objetivo en sí mismo en el sentido de que es necesario contar con una aproximación a estos valores para la planificación del transporte.

Objetivos

El objetivo general de este trabajo es realizar una predicción para los flujos faltantes de la matriz OD de los viajes por trabajo del departamento de Montevideo, considerando como unidad geográfica de análisis a los Centros Comunales Zonales (CCZ).

Como objetivos específicos se plantea:

- Comparar el desempeño de los métodos tradicionales frente a los modelos con efectos espaciales, de acuerdo al tamaño de los flujos faltantes de la matriz.
- Comparar el desempeño de los métodos tradicionales frente a los modelos con efectos espaciales, de acuerdo a la cantidad de celdas vacías de la matriz.

El documento se estructura en seis capítulos. En el presente capítulo se realiza una introducción al problema. El segundo capítulo abarca el marco teórico en el cual se sustenta el presente trabajo. En el capítulo tercero se especifica la metodología aplicada durante el proceso de investigación. El cuarto capítulo contiene una descripción de los datos utilizados, tanto para la comparación del desempeño de los métodos como para la imputación. En el quinto capítulo se presentan los resultados y por último, en el capitulo seis se presentan las conclusiones y las líneas futuras de investigación.

Capítulo 2. Marco Teórico

2.1. Modelos de interacción espacial

De acuerdo a O'Kelly (2009), los modelos de interacción espacial, en términos generales, intentan explicar los flujos y movimientos entre las regiones de una ciudad, en base a su separación espacial, su complementariedad y elementos de la estructura espacial que intervienen en la capacidad de atracción o propulsión de viajes.

Dadas n regiones, se tienen n^2 combinaciones de flujos posibles entre dichas regiones. Sea y_{ij} el flujo entre la región de origen i y la de destino j, con $i, j = 1, ..., n$, formando así la matriz OD:

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nn} \end{pmatrix} \tag{2.1}$$

Los modelos de interacción espacial se conocen también como modelos gravitacionales, terminología de amplio uso en la problemática del transporte. El término gravitacional se debe a su analogía con la teoría newtoniana, en donde una medida asociada a las regiones funciona como una fuerza de atracción de interacciones (análogo a la masa de un objeto) y la distancia, que al igual que en la física, tiene un efecto negativo en las interacciones. El modelo, en su forma más genérica, se define de la siguiente manera:

$$y_{ij} = A_i B_j O_i D_j f(d_{ij}) \tag{2.2}$$

donde A_i y B_j son factores de balance, O_i y D_j representan características del origen y del destino respectivamente, y $f(d_{ij})$ es una medida de resistencia o disuasión denominada como función de impedancia, en la cual se refleja la forma en que la separación espacial o la distancia impiden los movimientos a través del espacio.

Existen diversas especificaciones para esta función (Fischer y Wang, 2013). Una de las más utilizadas es la función en potencias en donde $f(d_{ij}) = d_{ij}^{-\theta}$, siendo d_{ij} la medida de distancia (o costo) entre el origen i y el destino j y θ es el parámetro a estimar (*distance decay parameter*)

que representa el grado en el cual los flujos disminuyen en función de la distancia que los separa.

Para poder especificar la matriz de diseño del modelo, es necesario ordenar los valores de la matriz (2.1) en un vector de dimensión $N \times 1$ (con $N = n \times n$). Para ello LeSage y Pace (2008) proponen dos alternativas, llamadas *origin-centric* y *destination-centric*.

El vector *origin-centric* tendrá en los n primeros elementos los flujos entre el origen 1 y los n destinos, en los n siguientes elementos tendrá los flujos entre el origen 2 y los n destinos, y así sucesivamente, como se muestra en (2.3).

$$\begin{array}{ccc} o^{(o)} & & d^{(o)} \quad l^{(o)} \\ 1 & \longrightarrow & 1 \quad\quad 1 \\ \vdots & \vdots & \vdots \quad\quad \vdots \\ 1 & \longrightarrow & n \quad\quad n \\ 2 & \longrightarrow & 1 \quad\quad n+1 \\ \vdots & \vdots & \vdots \quad\quad \vdots \\ 2 & \longrightarrow & n \quad\quad 2n \\ \vdots & \vdots & \vdots \quad\quad \vdots \\ n & \longrightarrow & 1 \quad\quad N-n+1 \\ \vdots & \vdots & \vdots \quad\quad \vdots \\ n & \longrightarrow & n \quad\quad N \end{array} \tag{2.3}$$

Denominamos $\mathbf{y} \in \mathbb{R}^N$ al vector *origin-centric*. De forma análoga se puede construir el vector *destination-centric* en el cual los primeros n elementos corresponderán a los flujos de los orígenes 1 a n al destino 1, los siguientes hacia destino 2, y así hasta completar los n destinos.

Por último, la matriz de distancias se define como una matriz simétrica $\mathbf{D} \in \mathbb{R}^{n \times n}$ (distancia en forma genérica, pudiendo estar medida en tiempo, en costos, etc.) entre las regiones i y j.

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{pmatrix} \tag{2.4}$$

Las distancias se incorporan al modelo definiendo $\mathbf{d} = \text{vec}(\mathbf{D})$ al vector de $\mathbb{R}^{\mathbf{N}}$ que resulta de apilar las filas de la matriz $\mathbf{D}$, manteniendo el orden establecido en (2.3).

2.2. Modelos de Conteo[1]

Dada la forma multiplicativa del modelo gravitacional (2.2), la primer aproximación utilizada para su estimación fue mediante una transformación logarítmica. De forma de simplificar el análisis se toman $O_i = 1$ y $D_j = 1$ y aplicando logaritmos a la ecuación (2.2), se obtiene

$$log(y_{ij}) = log(A_i) + log(B_j) - \theta log(d_{ij}) + \varepsilon_{ij} \text{ con } \varepsilon_{ij} \sim N(0, \sigma^2) \tag{2.5}$$

estimando el parámetro θ aplicando Mínimos Cuadrados Ordinarios (MCO).

Esta especificación no es correcta en muchas aplicaciones y presenta varios problemas. Flowerdew y Aitkin (1982) mencionan tres en particular: el primero es el sesgo creado por la transformación logarítmica. El segundo es la caída del supuesto de homoscedasticidad implícito en la estimación MCO: no es correcto suponer dada la gran variación en las magnitudes de los flujos, que los errores tengan igual varianza para cada par de orígenes y destinos. Por último, el uso de la transformación logarítmica no admite ceros. Usualmente se sustituye el cero por un valor pequeño, y en el caso de que existan muchos ceros el modelo termina siendo muy sensible al valor elegido para sustituir al cero.

Es así que surgen las especificaciones utilizando modelos de conteo. Los modelos de conteo son un tipo específico de regresión con datos discretos. Estos modelos son los usados cuando los datos provienen de un conteo de eventos, es decir, cuando la variable de interés mide la cantidad de veces que ocurre un suceso en en un intervalo de tiempo o en una región del espacio. De esta manera, los valores de la variable de respuesta deberán ser siempre enteros no negativos. El primer modelo a aplicarse bajo este contexto es el modelo Poisson que se describe a continuación.

2.2.1. Regresión Poisson

Dados los problemas que presenta la transformación logarítmica del modelo gravitacional, se plantea el uso de modelos de conteo, el primero de ellos utilizando la distribución Poisson, cuya función de cuantía es:

$$Pr(Y_{ij} = y_{ij}) = \frac{exp(-\mu_{ij})\mu_{ij}^{y_{ij}}}{y_{ij}!} \text{ con } y_{ij} = 0, 1, 2, \ldots \tag{2.6}$$

con

$$E[y_{ij}|d_{ij}] = \mu_{ij} = exp(-\theta d_{ij}) \tag{2.7}$$

A esta especificación se le llama "modelo de regresión Poisson con función de media exponencial". También se lo conoce como modelo log-lineal, ya que el logaritmo de la media condicional

[1] Esta sección se encuentra basada en el libro de Cameron y Trivedi (2013, Capítulo 3).

es lineal en los parámetros: $ln(E[y_{ij}|d_{ij}]) = -\theta d_{ij}$.

Si las observaciones son independientes, el estimador usado para el modelo de regresión Poisson es el estimador máximo verosímil. El logaritmo de la función de verosimilitud es:

$$ln(L(\theta)) = \sum_{j=1}^{n}\sum_{i=1}^{n} -y_{ij}d_{ij}\theta - exp(-\theta d_{ij}) - ln(y_{ij}!) \tag{2.8}$$

Y por lo tanto el estimador máximo verosímil $\hat{\theta}$ es la solución de:

$$\sum_{j=1}^{n}\sum_{i=1}^{n} -y_{ij}d_{ij} + exp(-\theta d_{ij})d_{ij} = 0 \tag{2.9}$$

Si y_{ij} se distribuye efectivamente Poisson con media $exp(-\theta d_{ij})$, entonces por las propiedades del estimador máximo verosímil se tiene que $\hat{\theta}$ se distribuye Normal:

$$\hat{\theta} \overset{a}{\sim} N(\theta, V_{MV}(\hat{\theta})) \tag{2.10}$$

La media es θ y su varianza está dada por:

$$V_{MV}(\hat{\theta}) = \left(\sum_{j=1}^{n}\sum_{i=1}^{n} \mu_{ij}d_{ij}d'_{ij}\right)^{-1} \tag{2.11}$$

El modelo Poisson es más adecuado para modelizar los flujos de viajes que una regresión lineal, pero asume equidispersión, ya que la media y la varianza condicional se suponen iguales. En la mayoría de las aplicaciones reales a flujos de Origen - Destino la varianza condicional es mayor que la media, lo que implica que la variable de respuesta presenta sobredispersión. Con el modelo Poisson pueden obtenerse estimaciones consistentes pero ineficientes, pudiendo llevar a conclusiones erróneas sobre la significación de los parámetros del modelo.

2.2.2. Binomial Negativa

Con el fin de corregir el problema de la sobredispersión de los datos surge la regresión con la especificación Binomial Negativa. La media condicional es la misma que para el modelo Poisson, pero la varianza se especifica como una función de la media condicional μ_{ij} y un parámetro de dispersión α tal que $V(y_{ij}|d_{ij}) = \mu_{ij} + \alpha\mu_{ij}^2$. En este caso la función de cuantía es:

$$Pr(y_{ij}) = \frac{\Gamma(y_{ij} + \alpha^{-1})}{y_{ij}!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_{ij}}\right)^{\alpha^{-1}}\left(\frac{\mu_{ij}}{\alpha^{-1} + \mu_{ij}}\right)^{y_{ij}} \tag{2.12}$$

Cuanto mayor es el valor de α mayor es la sobredispersión en los datos. Si α es aproximadamente cero, el modelo con la Binomial Negativa se reduce a un modelo Poisson.

En esta especificación del modelo se estiman dos parámetros y, al igual que en el modelo Poisson, se usa el estimador máximo verosímil.
Usando la igualdad $\Gamma(y+1) = y\Gamma(y)$ con lo cual $\Gamma(y+\alpha^{-1}) = \Gamma(\alpha^{-1})\prod_{k=0}^{y-1}(k+\alpha^{-1})$. El logaritmo de la función de verosimilitud es:

$$
\begin{aligned}
ln(L(\theta,\alpha)) = \sum_{j=1}^{n}\sum_{i=1}^{n}\Bigg\{ & \left(\sum_{k=0}^{y_{ij}-1} ln(k+\alpha^{-1})\right) - ln(y_{ij}!) \\
& -(y_{ij}+\alpha^{-1})ln(1+\alpha exp(-\theta d_{ij})) \\
& +y_{ij}ln(\alpha) - y_{ij}\theta d_{ij}\Bigg\}
\end{aligned}
$$

El estimador máximo verosímil de $\hat{\theta}$ y $\hat{\alpha}$ es la solución de:

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{y_{ij}-\mu_{ij}}{1+\alpha\mu_{ij}}d_{ij} = 0 \tag{2.13}$$

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\left\{\frac{1}{\alpha^2}\left(ln(1+\alpha\mu_{ij}) - \sum_{k=0}^{y_i-1}\frac{1}{k+\alpha^{-1}}\right) + \frac{y_{ij}-\mu_{ij}}{\alpha(1+\alpha\mu_{ij})}\right\} = 0 \tag{2.14}$$

Si la especificación Binomial Negativa es correcta, por las propiedades del estimador máximo verosímil se tiene que asintóticamente los parámetros tienen una distribución Normal:

$$\begin{bmatrix}\hat{\theta}\\ \hat{\alpha}\end{bmatrix} \overset{a}{\sim} N\left(\begin{bmatrix}\theta\\ \alpha\end{bmatrix}, \begin{bmatrix}V(\hat{\theta}) & Cov(\hat{\theta},\hat{\alpha})\\ Cov(\hat{\theta},\hat{\alpha}) & V(\hat{\alpha})\end{bmatrix}\right) \tag{2.15}$$

en donde

$$V(\hat{\theta}) = \left(\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\mu_{ij}}{1+\alpha\mu_{ij}}d_{ij}d'_{ij}\right)^{-1} \tag{2.16}$$

$$V(\hat{\alpha}) = \left(\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{1}{\alpha^4}\left(ln(1+\alpha\mu_{ij}) - \sum_{k=0}^{y_{ij}-1}\frac{1}{k+\alpha^{-1}}\right)^2 + \frac{\mu_{ij}}{\alpha^2(1+\alpha\mu_{ij})}\right)^{-1} \tag{2.17}$$

$$Cov(\hat{\theta},\hat{\alpha}) = 0 \tag{2.18}$$

Una ventaja del modelo especificado con la Binomial Negativa, además de incorporar un parámetro que permite modelar adecuadamente la existencia de sobredispersión que presenta la especificación Poisson, es que es robusto ante una mala especificación de la distribución. Esto se debe a que si α es conocido la distribución Binomial Negativa pertenece a la familia exponencial. Si además, la media condicional está correctamente especificada, el método es consistente en la estimación del parámetro θ del modelo gravitacional.

2.3. Econometría Espacial

Según Gaetan y Guyon (2010), existen tres grandes áreas de estudio dentro de la Estadística Espacial: el Análisis de Patrones de Puntos (*Spatial Point Pattern*), la Geoestadística y la Econometría Espacial (o análisis de datos de área). En el Análisis de Patrones de Puntos el interés se centra en el lugar en donde ocurrirán los eventos. Se utiliza por ejemplo en epidemiología, y se concentra en la modelización de procesos estocásticos en tiempo y espacio. En el caso de la Geoestadística, los datos pueden medirse en un principio en cualquier punto del espacio (datos continuos). El interés no es el patrón de puntos observados en sí mismo, sino en la predicción sobre un espacio continuo de una variable de interés medida en los sitios observados. Cuando los datos espaciales son observados en polígonos, se trata de datos discretos, y corresponde a lo que se denomina como Econometría Espacial.

Los modelos espaciales en el caso de la Econometría Espacial deben admitir la existencia de una dependencia entre polígonos "vecinos" (J. LeSage y Pace, 2009). Existe, por tanto, una correlación entre los valores de la variable de interés que es atribuible a la proximidad de las unidades geográficas (polígonos), y es lo que se denomina como "autocorrelación espacial". En la mayoría de los casos los polígonos corresponden a unidades administrativas, como ser un municipio, un departamento, país, etc. Los datos observados son frecuentemente agregados dentro de los límites del polígono, como por ejemplo totales o promedios dentro de las áreas administrativas. Los polígonos pueden ser definidos según el criterio del investigador o pueden estar predefinidos con cualquier otro tipo de criterio arbitrario (Bivand, Pebesma, y Gómez-Rubio, 2013), es decir, no relacionado con el uso para un análisis específico de los datos. En la gran mayoría de los casos los polígonos coinciden con unidades administrativas y si coincide con una agregación inadecuada puede ser la causa de que exista autocorrelación espacial.

En el caso de la Geoestadística, las distancias sobre una superficie continua son la base para determinar la estructura de la autocorrelación espacial. En el caso de la Econometría Espacial no existe una noción de distancia euclidea, si no que se deben definir estructuras de vecindad - cercanía entre los polígonos. Además del tipo de datos con el que se trabaja esta diferencia entre Geoestadística y Econometría Espacial es fundamental, ya que en el segundo caso es el investigador quien define el área de influencia. Al poder obtener diferentes resultados con diferentes áreas de influencia, ésta se convierte en un parámetro más del modelo, volviéndose un objeto central de análisis.

Este trabajo se enmarca dentro de la Econometría Espacial. Los flujos de viajes de una matriz OD no son entidades espaciales en sí mismas, pero deberá incorporarse a su análisis la autocorrelación espacial existente entre las regiones que definen la matriz OD. En las siguientes secciones se presentan los conceptos y definiciones básicos de Econometría Espacial, y los diferentes enfoques para el tratamiento de la autocorrelación.

2.3.1. La matriz de pesos $\mathbf{W}$[2]

La matriz de pesos espaciales es el concepto con el cual se define el área de influencia de las observaciones. Para poder incorporar la autocorrelación espacial, se crea una matriz $W = ((w_{ij}))_{i,j=1,\ldots,n}$ en donde las cantidades w_{ij} son fijas y están basadas en la estructura geográfica subyacente de las regiones. En la matriz de pesos W se refleja la estructura de "vecindad - cercanía" de las regiones de análisis. Se pueden aplicar distintos criterios para definir las regiones vecinas, y además se pueden definir pesos que diferencien, según su valor, la intensidad del vínculo entre cada par de vecinos.

Una primera forma de construir la matriz de pesos W es con una matriz de conectividad binaria en la que sus elementos son:

$$w_{ij} = \begin{cases} 1 & \text{si } i \text{ y } j \text{ son vecinos} \\ 0 & \text{si } i \text{ y } j \text{ no son vecinos} \end{cases} \tag{2.19}$$

Además, se tiene que para cualquier región i, $w_{ii} = 0$. Los vecinos pueden definirse por contigüidad o adyacencia:

$$w_{ij} = \begin{cases} 1 & \text{si } i \text{ y } j \text{ son polígonos contiguos} \\ 0 & \text{si } i \text{ y } j \text{ no son polígonos contiguos} \end{cases} \tag{2.20}$$

La elección de esta definición resulta en una matriz W simétrica ya que $w_{ij} = w_{ji}$.

En algunas aplicaciones el área de influencia puede exceder al área definida por el criterio de adyacencia. Existen diversos criterios para definir los vecinos en este caso. El primero sería utilizando los k vecinos más cercanos, de la siguiente forma:

$$w_{ij} = \begin{cases} 1 & \text{si el centroide de } j \text{ es uno de los } k \text{ más cercanos al centroide de } i \\ 0 & \text{en otro caso} \end{cases}$$

La matriz W resultante no tiene porque ser simétrica, es decir, no necesariamente se cumple que $w_{ij} = w_{ji}$. Sí se obtiene la misma cantidad de vecinos para todas las regiones.

Un segundo criterio consiste en fijar una valor δ y usar alguna función de distancia d_{ij} tal que se consideran vecinos de i todos los j cuyos centroides estén a una distancia d_{ij} menor que δ:

$$w_{ij} = \begin{cases} 1 & \text{si } d_{ij} < \delta \\ 0 & \text{en otro caso} \end{cases}$$

[2]Esta sección se encuentra basada en el libro de Bivand, Pebesma y Gómez Rubio (2013, Capítulo 9).

Nuevamente, al definir los vecinos de esta manera, W es simétrica.

Las matrices de conectividad binarias son recomendadas cuando se conoce poco sobre la estructura espacial subyacente, sin embargo, si se cuenta con información, se pueden construir matrices W con pesos no binarios. De manera similar al anterior ejemplo, se puede definir:

$$w_{ij} = \begin{cases} d_{ij}^{-1} & \text{si } d_{ij} < \delta \\ 0 & \text{en otro caso} \end{cases}$$

Y análogamente a la definición de la matriz de conectividad binaria por contigüidad:

$$w_{ij} = \begin{cases} l_{ij}/l_i & \text{si } i \text{ y } j \text{ son contiguos} \\ 0 & \text{si } i \text{ y } j \text{ no son contiguos} \end{cases}$$

Donde l_{ij} es el largo del borde compartido entre las regiones j e i y l_i es el perímetro de la región i. El valor w_{ij} es la proporción de borde de i que es compartido con j. Notar que en esta estructura, a diferencia de su análoga binaria, la matriz W no es simétrica.

Para una interpretación más sencillas de los parámetros del modelo a estimar, es conveniente ajustar los pesos de acuerdo a la cantidad de vecinos de cada región y de esta manera tener una matriz W_{std} estandarizada por filas:

$$w_{std,ij} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}}$$

En el caso de los flujos espaciales, la matriz refleja la interacción entre las regiones geográficas de la matriz OD, que deberá ser llevada a la estructura de flujos. Para ello se utiliza el producto de Kronecker:

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{pmatrix}$$

Y se definen tres tipos de interacción entre los flujos:

Sólo en origen

$$\mathbf{W}_o = \mathbf{W} \otimes \mathbf{I}_n$$

Sólo en destino

$$\mathbf{W}_d = \mathbf{I}_n \otimes \mathbf{W}$$

Cada una de estas especificaciones se encuentra asociada a un modelo sobre el comportamiento de los "viajantes" (Chun, 2008). El primero de ellos es el denominado *competing destinations*

model, el cual se basa en que la elección de un destino ocurre en un proceso en dos etapas, evaluándose primero la atracción de un grupo de destinos. Luego, en una segunda etapa, se evalúan los destinos individuales dentro del grupo seleccionado previamente (Fotheringham, 1983). Así, un destino individual se encuentra afectado por la proximidad de otros destinos. En el proceso de elección, el modelo se enfoca en la disposición o estructura espacial de los destinos, como lo determina la definición de la matriz $\mathbf{W}_d$.

Por otro lado, el modelo denominado *intervening opportunities model* se basa en la idea de que el número de viajes entre dos regiones se encuentra determinado por las oportunidades intermedias (o interpuestas), como el número de trabajos disponibles que existen entre el origen y el destino. Asumiendo que los "viajantes" se trasladan a la menor distancia posible, este modelo provee un argumento para la búsqueda espacial de destinos en forma secuencial en el contexto de los viajes por trabajo (Jayet, 1990). De esta manera, la disposición espacial de las regiones cercanas al origen tienen una gran influencia sobre el número potencial de oportunidades intermedias. La estructura de este modelo se ve reflejada en la matriz $\mathbf{W}_o$.

Por último, se consideran ambos modelos, teniendo en cuenta los efectos de las oportunidades intermedias, y la elección jerárquica de destinos que implica el *competing destinations model*. Ambos fenómenos se ven reflejados en la estructura de la matriz $\mathbf{W}_{od}$, que puede definirse de diferentes maneras:

$$\mathbf{W}_{od} = \mathbf{W} \otimes \mathbf{W}$$

$$\mathbf{W}_{od} = 1/2(\mathbf{W}_o + \mathbf{W}_d)$$

La primera definición refleja la dependencia basada en la interacción entre vecinos de origen y destino. En la segunda definición basa la relación de dependencia en un impacto acumulativo de los efectos de las interacciones de Origen - Destino (J. P. LeSage y Fischer, 2010).

2.3.2. Índice de Moran[3]

El índice de Moran es una de las medidas de autocorrelación más utilizadas en econometría espacial. Es usado para probar la existencia de correlación espacial, planteando la prueba de hipótesis:

H_0 : No existe correlación espacial entre las observaciones

H_1 : Existe correlación espacial entre las observaciones

Para la construcción del índice se tienen en cuenta tres factores:

[3]Esta sección se encuentra basada en el libro de Waller y Gotway (2004, Capítulo 7)

- Una medida de similaridad entre regiones, $sim_{ij} = (y_i - \bar{y})(y_j - \bar{y})$
- La varianza muestral de los valores observados, $s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Un factor de ajuste para los pesos espaciales usados, $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$

Para obtener el índice se suman todas las medidas de similaridades ponderadas por su respectivo peso espacial, se divide por la varianza muestral y se le aplica el factor de ajuste. De esta manera el índice I de Moran toma la forma:

$$I = \frac{1}{s^2}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} sim_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \tag{2.21}$$

Si las regiones vecinas tienden a tener valores similares, el índice I sera positivo. Cuando las regiones vecinas tiendan a tener valores disímiles, el índice I sera negativo. Bajo la hipótesis nula, cuando no exista correlación entre los valores de las regiones vecinas, el valor esperado del índice I es:

$$E(I) = -\frac{1}{n-1} \tag{2.22}$$

Para obtener la distribución de I bajo la hipótesis nula se toma el supuesto de *randomization*, es decir, que los valores observados están aleatoriamente asignados a las regiones. Luego se reasignan los valores observados entre todas las localidades, obteniendo así una distribución con la cual poder comparar el I observado.

2.3.3. Modelos Simultáneos Autorregresivos (SAR)[4]

Los modelos SAR son modelos endógenos, en el sentido de que la dependencia espacial se define a partir de los valores de la propia variable en los polígonos más cercanos.

La presencia de autocorrelación espacial implica que el valor observado en cada región depende de los valores observados en otras regiones. La especificación del modelo SAR es la siguiente:

$$y_i = \rho\sum_{j=1}^{n} w_{ij} y_j + X_{i1}\beta_1 + \cdots + X_{ik}\beta_k + \varepsilon_i \tag{2.23}$$

con

$$\varepsilon_i \sim N(0, \sigma^2) \qquad\qquad i = 1, ..., n \tag{2.24}$$

[4]Esta sección se encuentra basada en el libro de LeSage y Pace (2009, Capítulo 1)

donde $X_1, \ldots, X_k$ es un conjunto de variables explicativas y $\beta_1, \ldots, \beta_k$ los parámetros asociados (como en una regresión lineal), w_{ij} son los valores de la matriz de pesos espaciales, y ρ es el parámetro a estimar que describe qué tan fuerte es la dependencia espacial entre las observaciones. Como se puede observar, los valores y_i se generan con dependencia de los valores y_j para todos los i en simultáneo. Esto quiere decir que los valores de y_i dependen de y_j y viceversa. Al no tener constante, se asume que los valores de y_j se encuentran centrados en la media. El término $\sum_{j=1}^{n} w_{ij} y_j$ se llama rezago espacial ya que es una combinación lineal de valores de y construida a partir de las observaciones en regiones vecinas de i y la matriz de pesos W (J. LeSage y Pace, 2009).

En forma matricial

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.25}$$

donde $\mathbf{X}$ es la matriz de diseño del modelo, de dimensión $n \times k$, $\boldsymbol{\beta}$ es el vector de parámetros del modelo de dimensión $k \times 1$ y $\boldsymbol{\varepsilon}$ es el vector de residuos, de dimensión $n \times 1$. El rezago espacial queda definido por $\mathbf{W}\mathbf{y}$. Es conveniente trabajar con una matriz $\mathbf{W}$ estandarizada por fila (de forma que las filas sumen uno). De esta forma se obtienen directamente los promedios al calcular el rezago, y la estimación del ρ es directamente interpretable.

Este modelo tiene implícito el siguiente proceso generador de datos:

$$\begin{aligned} (\mathbf{I} - \rho\mathbf{W})\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} &+ (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim N(0, \sigma^2\mathbf{I}) \end{aligned} \tag{2.26}$$

donde $\mathbf{I}$ es la matriz identidad de dimensión $n \times n$. Se asume que $\boldsymbol{\varepsilon}$ sigue una distribución Normal multivariada, con media cero y matriz de varianzas y covarianzas $\sigma^2\mathbf{I}$. El proceso generador de datos expresa la naturaleza de "simultaneidad" del proceso autorregresivo espacial.

2.3.4. Filtrado Espacial[5]

El filtrado espacial es una técnica basada en los vectores propios de una transformación de la matriz de pesos espaciales. Los vectores tienen patrones espaciales distintivos asociados a distintos niveles de autocorrelación. La incorporación de subconjuntos de vectores propios en el modelo como variables regresoras "filtran" los residuos, haciendo que éstos no presenten correlación espacial.

Desde una perspectiva de una especificación errónea del modelo, se asume que el modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$ tiene errores autocorrelacionados espacialmente $\boldsymbol{\varepsilon}^*$, que pueden ser descompuestos

[5]Esta sección se encuentra basada en el artículo de Tiefelsdorf y Griffith (2007)

en un componente de ruido blanco, ε, y en un conjunto de variables exógenas no especificadas, $\mathbf{E}$. La estructura del modelo sería:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{2.27}$$

donde $\mathbf{E}\gamma$ es el término no especificado (omitido) en el modelo original, cuya inclusión elimina la autocorrelación espacial que presentan los residuos ε^*. El objetivo del filtrado es usar variables espaciales que aproximen este término, y así separar el componente $\mathbf{E}$ del ruido blanco en ε^*.

Relación entre el modelo SAR y el término omitido

Si el proceso espacial subyacente es estacionario, se puede expandir el término $(\mathbf{I} - \rho\mathbf{W})^{-1}$ de la ecuación (2.26):

$$(\mathbf{I} - \rho\mathbf{W})^{-1} = \sum_{k=0}^{\infty} \rho^k \mathbf{W}^k \tag{2.28}$$

Aplicando esta igualdad al modelo SAR se obtiene:

$$\begin{aligned}
\mathbf{y} &= \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\
\mathbf{y} - \rho\mathbf{W}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\
\mathbf{y} &= (\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
\mathbf{y} &= \sum_{k=0}^{\infty} \rho^k \mathbf{W}^k (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \underbrace{\sum_{k=1}^{\infty} \rho^k \mathbf{W}^k (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})}_{\text{término omitido}} + \boldsymbol{\varepsilon}
\end{aligned}$$

El término omitido $\sum_{k=1}^{\infty} \rho^k \mathbf{W}^k (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$ incluye a las variables exógenas $\mathbf{X}$. Por lo tanto no hay incorrelación entre las variables exógenas y el término omitido y bajo esta condición las estimaciones MCO de $\hat{\boldsymbol{\beta}}$ para el modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon^*$ son sesgadas, como así también las estimaciones de los desvíos estándar. El objetivo del filtrado espacial es utilizar variables proxy espaciales que puedan remplazar el término omitido en el modelo (2.27).

Obtención de los filtros espaciales

Considerando la matriz de proyección:

$$M_{(1)} = I - 1(1^T 1)^{-1} 1^T \tag{2.29}$$

el conjunto de vectores propios $\{\boldsymbol{e}_1, ...\boldsymbol{e}_n\}$ obtenido a partir de la matriz:

$$T_1 = M_{(1)} 1/2(\mathbf{W} + \mathbf{W}^T) M_{(1)} \tag{2.30}$$

puede estar correlacionado con las variables $\mathbf{X}$, ya que es ortogonal sólo al vector unidad. Este conjunto de vectores propios establece la base para las variables proxy espaciales a utilizar. Los vectores propios $\boldsymbol{e}_i$ y $\boldsymbol{e}_j$ son ortogonales entre sí, ya que la transformación simétrica $1/2(\mathbf{W} + \mathbf{W}^T)$ garantiza una forma cuadrática en la expresión $M_{(1)}$. Sea $\mathbf{E}$ una matriz compuesta por un subconjunto de $\{\boldsymbol{e}_1, ..., \boldsymbol{e}_n\}$. El término omitido será aproximado por una combinación lineal de este subconjunto tal que:

$$\mathbf{E}\boldsymbol{\gamma} \approx \sum_{k=1}^{\infty} \rho^k \mathbf{W}^k \left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) \tag{2.31}$$

$\mathbf{E}\boldsymbol{\gamma}$ se encuentra correlacionado con las variables exógenas $\mathbf{X}$, y su incorporación al modelo corrige el sesgo en la estimación MCO de los parámetros $\hat{\boldsymbol{\beta}}$ en el modelo (2.27). El modelo $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{E}\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\varepsilon}}$ es una descomposición del modelo (2.25) en un componente de tendencia, una señal estocástica y en un ruido blanco. Para este modelo, la tendencia y la señal estocástica no se encuentran incorrelacionadas, pudiendo existir algún problema de multicolinealidad. De todas formas, aunque pueda existir algún factor de inflación de la varianza, se logra una disminución en el sesgo en la estimación de $\hat{\boldsymbol{\beta}}$ y sus desvíos, ya que se elimina la autocorrelación espacial existente en los residuos.

Relación entre los filtros espaciales y el Índice de Moran

Los vectores propios $\boldsymbol{e}_i$ exhiben patrones espaciales distintivos de acuerdo a su valor propio asociado λ_i, tanto por su signo como por su magnitud. El Índice de Moran asociado a cada vector propio $\boldsymbol{e}_i$ es igual a su valor propio asociado

$$\lambda_i = \left[\boldsymbol{e}_i^T \left(\mathbf{W} + \mathbf{W}^T\right) \boldsymbol{e}_i\right] / \left(2\boldsymbol{e}_i^T \boldsymbol{e}_i\right) \tag{2.32}$$

si $\mathbf{W}$ se encuentra escalada de forma que $\left[\mathbf{1}^T \left(\mathbf{W} + \mathbf{W}^T\right) \mathbf{1}\right] /2 = n$. De esta forma se tiene que el signo de λ_i determina el signo de la correlación del vector $\boldsymbol{e}_i$. El primer vector propio $\boldsymbol{e}_1$ se encuentra asociado al valor propio de mayor valor absoluto, λ_1, el segundo vector $\boldsymbol{e}_2$ se encuentra asociado al valor propio que le sigue al primero en orden de magnitud, λ_2, y así sucesivamente. Los primeros vectores propios exhiben patrones espaciales globales, y a medida que aumenta el orden se asocian a escalas más locales.

Determinación del subconjunto $\mathbf{E}$

La selección de los vectores propios se realiza con un procedimiento similar al *forward stepwise* (Chambers, Hastie, y Pregibon, 1990), midiendo en cada iteración el Índice de Moran de los residuos, hasta que no presenten correlación espacial.

En una primera iteración se estima el modelo $\mathbf{X}\boldsymbol{\beta} + \varepsilon^*$ y se calcula el Índice de Moran a los residuos del modelo, para determinar la existencia de correlación espacial. En caso de que se

rechace la hipótesis de no correlación se procede al siguiente paso. En la segunda iteración se estima el modelo (2.27) con la totalidad de vectores $\{\boldsymbol{e}_1, ..., \boldsymbol{e}_n\}$ y se determina cuál es el vector que tiene una mayor significación en el modelo (2.27). Denotando como $\boldsymbol{e}_1^*$ al vector que cumpla esta condición, se estima el modelo:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}_1^*\gamma_1 + \boldsymbol{\varepsilon}^{(1)}$$

y se calcula el Índice de Moran a los residuos $\boldsymbol{\varepsilon}^{(1)}$. Si no se rechaza la hipótesis de no autocorrelación se detiene el algoritmo y el conjunto $\mathbf{E}$ queda definido por $\{\boldsymbol{e}_1^*\}$. Si se rechaza se estima el modelo (2.27) con el conjunto de $n-1$ vectores que no fueron seleccionados en el paso anterior. Procediendo de la misma forma, se elije el vector $\boldsymbol{e}_2^*$ que es más significativo en la estimación del modelo (2.27) considerando los $n-1$ vectores y se estima el modelo:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}_1^*\gamma_1 + +\boldsymbol{e}_2^*\gamma_2 + \boldsymbol{\varepsilon}^{(2)}$$

Si los residuos $\boldsymbol{\varepsilon}^{(2)}$ no presentan correlación espacial, el algoritmo se detiene y el conjunto $\mathbf{E} = \{\boldsymbol{e}_1^*, \boldsymbol{e}_2^*\}$. Si los residuos $\boldsymbol{\varepsilon}^{(2)}$ aún presentan autocorrelación espacial de acuerdo al Índice de Moran se procede a seleccionar un tercer vector $\boldsymbol{\varepsilon}^{(3)}$ de los $n-2$ no seleccionados en los pasos anteriores, y así sucesivamente hasta que los residuos del modelo no presenten autocorrelación espacial. El conjunto $\mathbf{E}$ queda determinado por el conjunto de vectores seleccionado que hacen que los residuos del modelo se comporten como un ruido blanco.

Filtrado espacial aplicado a los flujos de Origen - Destino

En el caso de los flujos OD se tienen tres posibles matrices de pesos espaciales, W_o, W_d y W_{od}. Se obtienen sus respectivas matrices transformadas, de acuerdo a (2.30), a las que denominaremos T_o, T_d y T_{od}. Se calculan los vectores propios y se aplica el mismo algoritmo de selección, incorporando un subconjunto de vectores al modelo de forma que los residuos no presenten correlación espacial.

En el contexto de un modelo gravitacional asumiendo una distribución Poisson o Binomial Negativa, los filtros se incorporan agregándolos a la especificación del modelo,

$$\mu_{ij} = exp(\lambda + \theta d_{ij} + \sum_{q=1}^{Q} \gamma_q e_q^*) \tag{2.33}$$

siendo Q la cantidad de vectores propios seleccionados de las matrices transformadas T_o, T_d o T_{od} y γ_q los parámetros asociados que indican la importancia relativa que tiene cada patrón espacial asociado a los vectores $\boldsymbol{e}_q$ en explicar la autocorrelación espacial de la estructura de flujos.

Los modelos a estimar en este trabajo son como el especificado en la ecuación (2.33). Será necesario definir la distribución que mejor ajusta a los datos (Poisson o Binomial Negativa), la estructura de la matriz W_{od} e implementar la selección de los filtros espaciales. En el siguiente capítulo se describe el detalle de la metodología a implementar para la selección del modelo a utilizar en la imputación de los valores faltantes de la matriz OD.

Capítulo 3. Metodología

El objetivo de este trabajo es realizar una predicción para los flujos faltantes de la matriz OD de los viajes por trabajo para el departamento de Montevideo, considerando como unidad geográfica de análisis a los Centros Comunales Zonales (CCZ).

La matriz OD considerando a los CCZ como unidad geográfica de análisis presenta aproximadamente un 30 $\%$ de celdas vacías. Para tener una buena medida del desempeño de ambos modelos, la evaluación de los mismos debería realizarse sobre una matriz sin datos faltantes, en donde puedan simularse distintos patrones de celdas vacías, y comparar los resultados de las predicciones con los verdaderos valores de la matriz. Para elegir el método con mejor desempeño en la predicción se trabaja con una matriz completa, correspondiente a la encuesta de movilidad de Bogotá 2015 – Caracterización viajes – Origen / Destino (Alcaldía Mayor de Bogotá, 2015), en donde se entrevistaron 28.213 hogares y se tiene información sobre 147.251 viajes.

La metodología se desarrolla en las siguientes etapas:

- Definición de la matriz W.
- Definición del tipo de interacción entre flujos ($\mathbf{W}_o$, $\mathbf{W}_d$ o $\mathbf{W}_{od}$).
- Elección de la distribucíón para el modelo gravitacional de conteo.
- Cálculo del Índice de Moran para los residuos del modelo estimado en el paso anterior.
- Obtención del conjunto $\{\boldsymbol{e}_1, ..., \boldsymbol{e}_n\}$.
- Simulación de distintos escenarios de celdas vacías en la matriz OD, teniendo en cuenta diferentes cantidades y patrones de flujos faltantes.
- Predicción de los flujos con el modelo gravitacional sin considerar la autocorrelación espacial.
- Predicción de los flujos del modelo gravitacional incorporando los vectores del filtrado espacial.

- Comparación de resultados de acuerdo a los patrones de datos faltantes simulados utilizando como criterio la Raíz cuadrada de la media de los errores al cuadrado ($RMSE$).
- Predicción de los flujos faltantes de la matriz OD de Montevideo con el método que presenta mejor desempeño.

En los siguientes capítulos se presenta la descripción de los conjuntos de datos con los que se trabaja (encuestas de movilidad de Bogotá y de Montevideo), el detalle de las decisiones tomadas en cada una de las etapas de la metodología, y finalmente los resultados obtenidos en cuanto a la imputación de las celdas vacías de la matriz OD, junto con una descripción de los filtros espaciales que capturan la correlación presente entre los flujos de la matriz.

Capítulo 4. Datos

En este trabajo se utilizarán dos conjuntos de datos. Si bien el objetivo es estimar los flujos faltantes de la matriz OD para la ciudad de Montevideo considerando como unidades geográficas a los CCZ para el caso de los viajes por trabajo, como fue explicado en el capítulo anterior, para elegir el método de imputación se trabaja con una matriz OD completa, que permita evaluar mejor el desempeño de los modelos en cuanto a los valores imputados. Se trabaja en este caso con la Encuesta de movilidad de Bogotá 2015 – Caracterización viajes – Origen / Destino.

4.1. Viajes entre localidades de Bogotá

Las unidades geográficas consideradas serán las localidades de Bogotá. La ciudad esta dividida administrativamente en 20 localidades, sin embargo la matriz OD se construirá con los viajes realizados por motivos de trabajo entre las 15 más céntricas (ver Figura 4.1). Estarán excluidas las 3 localidades que se encuentran más al sur (Sumapaz, Ciudad Bolívar y Usme) y las dos localidades que se encuentran mas al norte (Suba y Usaquén). Sumapaz es una localidad rural dentro de Bogotá y las otras cuatro, si bien tienen parte del casco urbano dentro de sus límites, se tratan de zonas en su mayoría rurales.

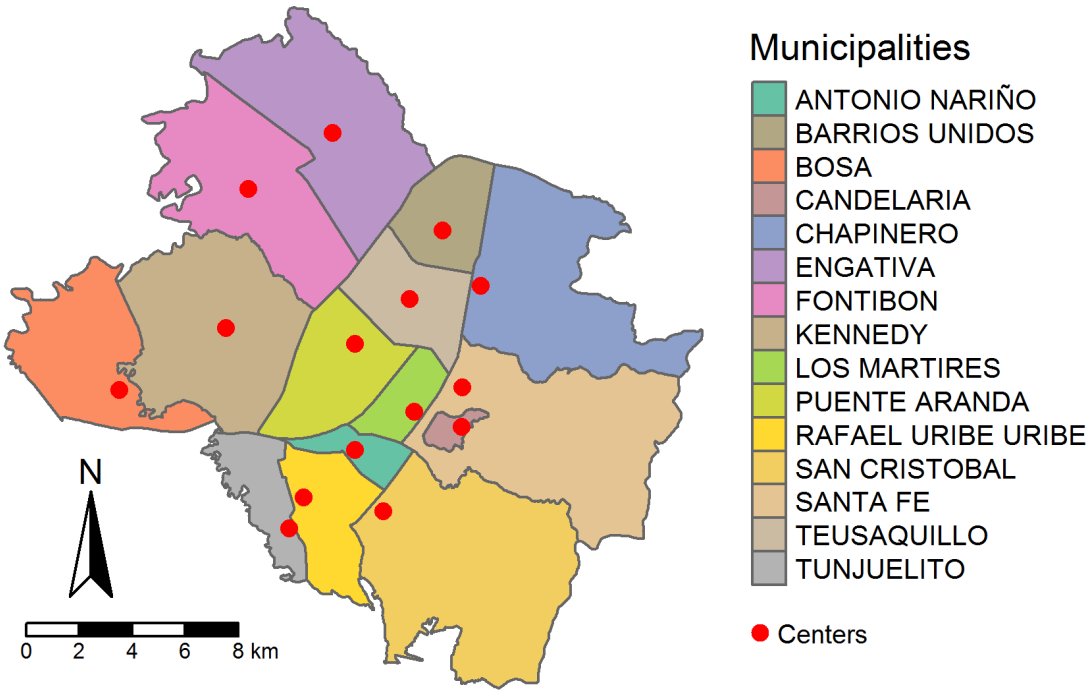

Figura 4.1: Localidades de la matriz de Origen - Destino

Los datos provienen de la encuesta de movilidad de Bogotá 2015 – Caracterización viajes – Origen / Destino, en donde se entrevistaron 28.213 hogares de los cuales se obtuvo información sobre 147.251 viajes. Los datos usados son los que corresponden a viajes por motivos laborales de las 15 localidades mencionadas en el párrafo anterior. Estos suman en total 13.426 viajes en la muestra para las zonas consideradas, habiendo sido relevada la información sobre los viajes realizados de lunes a sábado.

La matriz OD para localidades de Bogotá es:

	Antonio Nariño	Barrios Unidos	Bosa	Candelaria	Chapinero	Engativá	Fontibón	Keneddy	Los Mártires	Puente Aranda	Rafael Uribe Uribe	San Cristóbal	Santa Fe	Teusaquillo	Tunjuelito
Antonio Nariño	76	23	4	6	75	28	22	40	40	42	27	11	60	34	20
Barrios Unidos	4	127	1	32	124	38	20	23	6	23	3	1	49	58	0
Bosa	15	52	149	15	86	87	86	167	37	73	14	9	51	26	13
Candelaria	4	21	1	80	32	17	18	12	19	17	3	14	108	19	4
Chapinero	7	70	1	25	377	50	52	26	20	28	4	8	83	66	3
Engativá	16	121	14	22	177	462	152	91	36	62	9	3	124	139	3
Fontibón	14	47	4	20	94	80	370	76	36	52	6	11	86	92	2
Keneddy	24	94	76	44	188	125	216	615	105	181	24	22	181	136	26
Los Mártires	20	45	18	21	88	29	57	32	250	50	11	12	112	70	6
Puente Aranda	9	31	8	31	85	33	47	67	39	178	11	7	81	63	9
Rafael Uribe Uribe	31	16	6	11	93	46	35	52	31	48	104	34	96	100	27
San Cristóbal	61	38	11	50	150	55	74	69	43	80	34	156	166	69	7
Santa Fe	35	26	1	88	102	36	56	28	54	48	16	19	247	93	6
Teusaquillo	1	68	5	51	225	50	61	42	59	86	9	12	131	213	3
Tunjuelito	27	33	14	25	76	47	31	69	45	65	28	11	46	78	83

Se observa, como es frecuente en este tipo de matrices, que en la diagonal están los mayores flujos. Kennedy es la localidad de la diagonal con mayor número de viajes, esto se debe a que a pesar de no ser el centro de Bogotá, es la localidad de mayor población y una de las de mayor superficie. Las localidades que atraen más viajes son Chapinero y Santa Fe. Chapinero está en una zona céntrica de la ciudad de gran actividad comercial y Santa Fe es el centro tradicional (histórico) de Bogotá junto con la localidad de La Candelaria. Esta última es una localidad especial, está completamente situada en el interior de Santa Fe y su superficie apenas supera los $2\ km^2$, por eso pese a ser céntrica no es de las localidades que más viajes recibe. Como tercer localidad más receptora de viajes aparece la localidad de Kennedy, si bien no es céntrica como las anteriores, es una localidad con oferta de trabajo, en particular en industria y comercio, uno de los mayores mercados de abasto de la región se encuentra dentro de los límites de Kennedy. Dentro de las generadoras de viajes se encuentran Kennedy, Engativa y San Cristóbal. Las tres comparten la característica de ser localidades de gran superficie y alta población. Con respecto a esto último hay que notar que Bosa es una localidad de mayor población que San Cristóbal y aun así genera menos viajes, una diferencia entre ambas que explica la mayor cantidad de generación de viajes de San Cristobal es que se encuentra a menor distancia de las dos localidades más atractoras de viajes, incluso limita con la localidad de Santa Fe.

4.2. Viajes entre los CCZ de Montevideo

Las unidades geográficas consideradas serán los CCZ de la ciudad de Montevideo. Si bien la ciudad de Montevideo se encuentra dividida en 18 CCZ, se usaron 17 polígonos para construir la matriz OD ya que ningún elemento de la muestra quedo dentro del CCZ 15. Se creo un nuevo polígono con la unión los CCZ 3 y 15 (manteniendo el nombre ÇCZ 3") como se ve en el mapa a continuación:

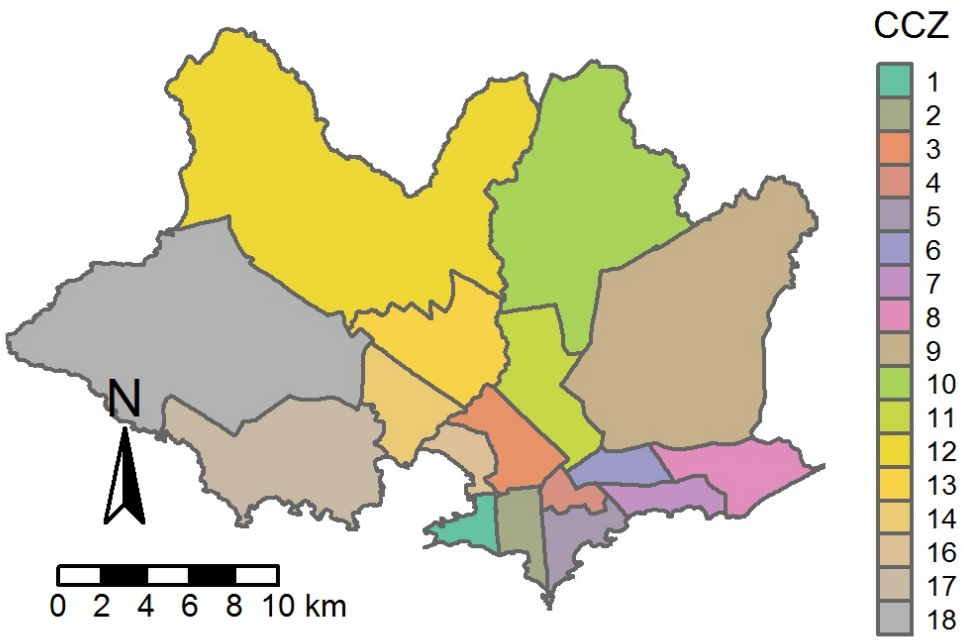

Figura 4.2: Unidades Geográficas de la matriz de Origen - Destino

Los datos a imputar provienen de la Encuesta de Movilidad en el Área Metropolitana de Montevideo. Se relevaron 2.230 hogares de los cuales se obtuvo información de 12.546 viajes. Se tomaran en cuenta solo los viajes por motivos laborales de los 17 CCZ. En total se tienen 2170 viajes por dicho motivo en la muestra.
La matriz OD es:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**16**	**17**	**18**
1	12	6	2	2	3	0	1	1	1	0	1	1	0	0	2	0	0
2	32	19	5	7	7	0	1	2	2	1	0	1	0	1	1	0	0
3	11	8	16	7	12	1	0	1	1	0	1	0	0	1	3	0	1
4	3	3	1	3	2	0	0	1	0	0	2	2	0	0	0	0	0
5	31	13	2	10	21	3	1	1	0	0	1	0	1	0	1	1	0
6	9	5	3	3	4	3	3	0	0	0	1	0	0	0	1	0	0
7	17	12	3	8	8	1	2	1	3	0	2	0	1	0	1	0	0
8	6	2	1	1	3	0	3	5	1	0	0	1	0	0	2	0	0
9	16	21	14	10	12	6	5	6	41	3	12	2	1	3	4	5	3
10	8	8	5	11	15	1	11	1	9	15	7	1	4	0	2	0	1
11	13	6	8	1	12	0	4	4	8	7	21	1	2	2	2	0	0
12	7	1	1	0	4	0	0	0	0	0	2	7	1	2	1	0	0
13	3	11	4	2	0	2	1	2	3	0	2	3	20	7	4	0	1
14	16	10	8	7	7	1	3	0	3	1	3	1	4	20	1	5	1
16	7	2	2	1	0	0	0	0	1	0	0	2	1	0	4	0	0
17	12	9	8	6	18	1	2	0	0	0	4	0	2	7	4	6	3
18	2	2	5	1	0	0	0	0	3	0	1	1	4	3	1	2	13

Tal como ocurre en la matriz OD de Bogotá, los mayores flujos de viajes están en la diagonal. La diagonal con mayor número de viajes corresponde al CCZ 9, esto se debe a ser uno de los CCZ de mayor tamaño y en el cual viven un mayor número de personas. Incluye los barrios de Manga, Flor de Maroñas, Villa Española, Piedras Blancas, entre otros. Centrándose en los destinos, el CCZ 1 es el que atrae más viajes. EL CCZ 1 está en la zona céntrica de la ciudad. Esto es esperable ya que los barrios Ciudad Vieja y Centro con una gran concentración de lugares de trabajo son parte de este CCZ. En atracción de viajes sigue el CCZ 2 (Cordón, Palermo, Parque Rodó) que comparte límite con el CCZ 1. Dentro de los generadores de viajes se encuentran en primer lugar el CCZ 9, por las mismas razones antes mencionadas y en segundo lugar el CCZ 10 que comparte algunos barrios con el CCZ 9.

Se observa que la matriz OD presenta 85 celdas vacías, un 29.4 % del total de sus valores. Los CCZ 12 y 17 tienen 12 celdas vacías como destinos (columnas de la matriz OD). Al ser CCZ ubicados en la periferia de Montevideo y con algunas zonas rurales, es esperable que tengan valores bajos (incuso algún cero) pero también es posible que esta cantidad de ceros se deba al tamaño de la muestra de la encuesta de movilidad.

Capítulo 5. Resultados

En este capítulo se presenta en una primer sección las definiciones previas necesarias para la implementación de la imputación de las celdas vacías de la matriz OD con los modelos gravitacionales considerados.

En las siguientes secciones se presentan los resultados de los modelos gravitacionales y los detalles del procedimiento para la simulación de las celdas vacías de la matriz OD.

Por último se comparan los resultados de ambos modelos y se realiza la predicción para la matriz OD de los viajes por trabajo de Montevideo con el método que presenta menor ARMSE (*Average Root Mean Square Error*).

5.1. Definiciones

El primer paso consiste en definir el área de influencia de las unidades geográficas de la matriz OD. En este caso se usará el criterio de adyacencia, y w_{ij} se define de acuerdo a la ecuación (2.20). La matriz W además se estandariza por filas.

Luego de definir la forma de la matriz W se procede a determinar el tipo de interacción entre flujos. En este caso se utiliza la definición de W_{od} de acuerdo a la ecuación $1/2(W_o + W_d)$.

Del conjunto total de vectores propios de la transformación propuesta en la ecuación 2.30 de la matriz W_{od} se obtiene el subconjunto de vectores propios candidatos a "filtrar" la auto-correlación espacial. Ese subconjunto se obtiene agregando los vectores propios de a uno al modelo hasta que no se rechace la hipótesis nula del test de Moran. En cada paso, el vector que se agrega es aquel, del conjunto total de vectores propios, que maximiza el valor p de la prueba. Dado que los vectores propios son de dimensión $1 \times N$ y la matriz de diseño al tener datos faltantes, es de menor dimensión, es necesario modificar la dimensión del vector para que incorporarlo como variable al modelo. Esto se realiza eliminando la fila del vector propio que se corresponda con la celda del flujo faltante.

La única variable explicativa que tendrán los modelos sera la distancia. Como medida de distancia para Bogotá se usó el tiempo de viaje en vehículo privado entre los centros (centroides) de las localidades, que se obtiene evaluando el tiempo que toma en llegar desde el centro de la localidad de origen hasta el centro de la localidad de destino, teniendo en cuenta la ruta posible más rápida. Para realizar dicha medición se utilizó la aplicación "Distancias Himmera" (*Distancias Himmera*, s.f.).

Para Montevideo la distancia entre los CCZ se mide con el promedio entre el tiempo de transporte en ómnibus y en auto entre los centros de los CCZ. Estos centros se fijan a partir de las "centralidades" de la ciudad definidas por la Intendencia de Montevideo (*SIG - Intendencia de Montevideo*, s.f.). Los tiempos en transporte público se miden con la aplicación "Cómo Ir" (*Intendencia de Montevideo*, s.f.) y los tiempos en transporte privado se calcula utilizando Google Maps (Google, s.f.). Los cálculos se realizan a la misma hora del día, de forma de evitar variaciones debidas al tránsito.

5.2. Elección de la distribución para el modelo gravitacional.

Si bien la especificación Binomial Negativa presenta ventajas sobre la especificación Poisson, se evaluará como se ajustan a los datos ambos modelos usando el criterio AIC, para evaluar la magnitud de la diferencia entre ambas especificaciones. Como puede observarse en las tablas 5.1 y 5.2, con ambos métodos se llega a un valor negativo de β_1, lo que es razonable con los supuestos del modelo gravitacional.

Tabla 5.1: Resultados del ajuste con la especificación Poisson

Variable	Parámetro	Estimación	p-valor
Intercept	β_0	5.09	2e-16
distancia	β_1	-0.089	2e-16
IM Residuos		0.52	2.2e-16
AIC		10977	
BIC		10984	

El AIC obtenido con la especificación Binomial Negativa es notoriamente menor el AIC que se obtiene en la especificación Poisson. En las siguientes secciones del capítulo se continuará entonces con el modelo gravitacional con la especificación Binomial Negativa.

5.3. Simulación de datos faltantes

Para la simulación de datos faltantes se comienza sustituyendo de a una celda a la vez el valor del flujo por cero, obteniendo así matrices OD con una celda vacía. Luego se imputan con

Tabla 5.2: Resultados del ajuste con la especificación Binomial Negativa

Variable	Parámetro	Estimación	p-valor
Intercept	β_0	4.89	2e-16
distancia	β_1	-0.071	1.63e-12
IM Residuos		0.55	2.2e-16
AIC		2243	
BIC		2253	

ambos métodos en cada paso hasta recorrer la matriz entera. Para simular matrices con dos o más celdas vacías se sortean cien grupos de celdas (para simular matrices con dos celdas vacías se sortean cien pares, de tres celdas vacías se sortean cien ternas, y así sucesivamente). Para cada grupo se dejan vacías las celdas y se imputan los valores con cada uno de los modelos: gravitacional clásico (G) y gravitacional con incorporación de filtros espaciales (GE).

Es importante notar en este punto que para que las dimensiones en el modelo sean conformables es necesario quitar la fila i a los vectores propios, correspondientes al flujo ij vacío. Esto hace que a medida que aumenten los datos faltantes se vaya perdiendo parte de la información contenida en cada vector propio. Es esperable que el modelo GE vaya perdiendo eficiencia al aumentar las celdas vacías.

Para la comparación entre los modelos se usa el ARMSE. Una vez obtenida la predicción para cada k-úpla se calcula el RMSE. A partir de los 100 valores del RMSE se calcula el promedio, obteniendo así el ARMSE para cada método según la cantidad K de datos faltantes.

- $RMSE_r^{(K)} = \sqrt{\frac{\sum_{t=1}^{K}(\hat{y}_t - y_t)^2}{K}}$ con $r = 1, ..., 100$
- $ARMSE^{(K)} = \frac{\sum_{r=1}^{100} RMSE_r^{(K)}}{100}$

5.4. Comparación de resultados de acuerdo a la cantidad de celdas faltantes.

En la Tabla 5.3 se presentan los valores del $ARMSE$ para cada uno de los modelos, para algunos valores de porcentajes de celdas vacías. Se observa que cuando el porcentaje de celdas faltantes es menor o igual al $30\,\%$, el método basado en la inclusión de los filtros espaciales presenta un menor ARMSE, y por ende un mayor poder de predicción. La máxima diferencia entre los métodos se da cuando se anula el $10\,\%$ de flujos en la matriz OD: el modelo gravitacional basado en filtros espaciales presenta un valor del ARMSE $36\,\%$ más bajo respecto al modelo gravitacional tradicional. A medida que aumenta el porcentaje de celdas vacías, esta eficiencia se reduce, hasta llegar a valores similares, cuando el porcentaje de flujos eliminados se encuentra en el entorno del $40\,\%$. En este caso se está frente a una situación en donde se debería analizar la regionalización utilizada, ya que implicaría imputar un alto porcentaje de los datos totales.

Tabla 5.3: Comparación de Resultados de la Imputación según la cantidad de celdas faltantes.

	$ARMSE$		
Porcentaje de celdas faltantes	Gravitacional (G)	Gravitacional + $\mathbf{E}$ (GE)	$ARMSE_{GE}/ ARMSE_G$
$10\,\%$	58	37	$0{,}64$
$20\,\%$	63	49	$0{,}78$
$30\,\%$	64	57	$0{,}89$
$40\,\%$	64	68	$1{,}06$

En la Figura 5.1 se puede apreciar lo mencionado anteriormente: el promedio por cantidad de celdas faltantes de los errores de ambos métodos convergen cuando la matriz OD se acerca aproximadamente a un $40\,\%$ de celdas vacías.

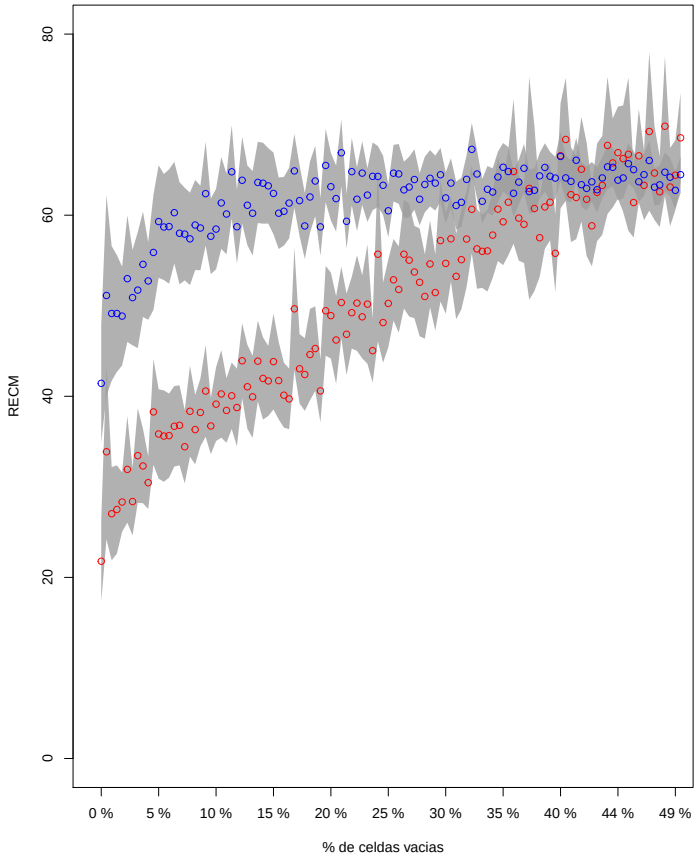

Figura 5.1: ARMSE según porcentaje de celdas faltantes

5.5. Comparación de resultados de acuerdo a diferentes patrones de celdas faltantes.

En la sección anterior se analizaron los resultados sin tomar en cuenta el tamaño de los flujos faltantes. Sin embargo, es razonable esperar que se encuentren celdas vacías cuando la cantidad de viajes entre dos regiones sea baja. De cualquier manera, con el fin de comparar ambos métodos en distintos escenarios, se midió el desempeño para tres patrones de datos faltantes. Para flujos pequeños (menos de 15 viajes), para flujos medianos (15 o más viajes y 69 o menos viajes) y para flujos altos (más de 69 viajes). Si bien se compara para cada cantidad de datos faltantes (como puede verse más adelante en la gráfica), mostramos en la tabla 5.4 el escenario para un 30 % de celdas vacías ya que es el más asimilable a los datos de Montevideo.

Tabla 5.4: Comparación de Resultados de la Imputación según el tamaño de los flujos para 30 % de celdas faltantes.

	$ARMSE$		
Tamaño del flujo	Gravitacional (G)	Gravitacional + $\mathbf{E}$ (GE)	$ARMSE_{GE}/$ $ARMSE_G$
Pequeños	41	11	0,27
Medianos	29	33	1,14
Grandes	104	96	0,92
Total	64	57	0,89

El modelo gravitacional que considera la autocorrelación espacial es levemente mejor cuando el patrón de celdas vacías abarca los flujos grandes, por otro lado, el modelo gravitacional tradicional es levemente mejor cuando los flujos son medianos. Por último, cuando se contrastan ambos métodos para flujos pequeños es donde el método que considera la autocorrelación espacial tiene un desempeño mejor, incluso cuando la cantidad de celdas vacías es de un 30 % de la matriz OD. Es importante apreciar que éste es el escenario más factible, ya que son los flujos pequeños los que podrían faltar cuando los datos se obtienen mediante una muestra.

En el siguiente gráfico se puede observar con claridad la mejor eficiencia del modelo con los vectores propios por sobre el modelo gravitacional clásico cuando se evalúa en un patrón de flujos faltantes pequeños. Para matrices OD con hasta un 50 % de celdas vacías, el método que considera la autocorrelación espacial siempre es más eficiente que el modelo gravitacional tradicional.

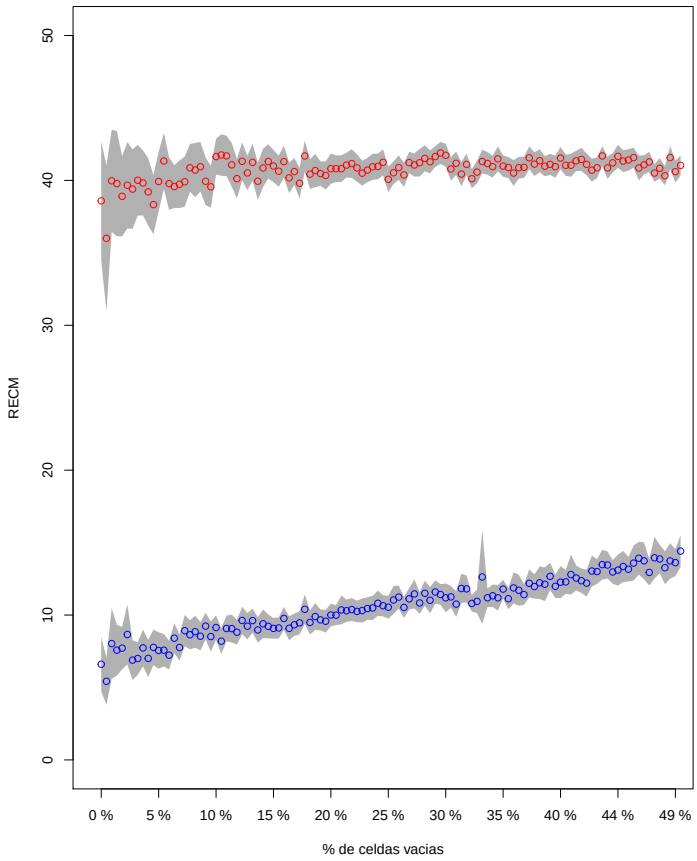

Figura 5.2: ARMSE para flujos pequeños según porcentaje de celdas faltantes

En la Tabla 5.5 se resumen algunos valores del gráfico anterior, también, como se aprecia en el gráfico, a medida que aumenta la cantidad de celdas vacías, disminuye levemente la eficiencia del modelo que considera los filtros espaciales, pero con valores notoriamente menores del $ARMSE$.

Tabla 5.5: Comparación de Resultados de la Imputación para los flujos pequeños según el porcentaje de celdas faltantes.

	$ARMSE$		
Porcentaje de celdas faltantes	Gravitacional $+ \mathbf{E}(G)$	Gravitacional (GE)	$ARMSE_{GE}/ ARMSE_G$
10 %	40	9	0,22
20 %	41	10	0,24
30 %	41	11	0,27
40 %	41	12	0,29

Centrándose en el comportamiento de los modelos dentro del grupo de los flujos pequeños, donde el método que incorpora los filtros espaciales muestra la mayor diferencia de eficiencia

con el método tradicional, resulta interesante comparar ambos métodos desglosando aún más el porcentaje de celdas vacías para estos viajes. Para el caso de 30 % de celdas vacías (cuadro 5.6), se puede observar como se repite nuevamente la efectividad mayor del método que incorpora filtros espaciales cuanto menores son los flujos. En particular este método alcanza la máxima diferencia por sobre el método tradicional cuando los flujos son menores a 6 viajes.

Tabla 5.6: Comparación de Resultados de la Imputación para los flujos pequeños, con un 30 % de celdas faltantes, desagregado por tamaño.

	$RMSE$		
Tamaño del flujo	Gravitacional (G)	Gravitacional + $\mathbf{E}$ + $\mathbf{E}$ (GE)	$RMSE_{GE}/RMSE$
Menos de 6	38	6	$0,16$
Entre 6 y 9	43	12	$0,28$
Entre 9 y 14	40	12	$0,3$
Total	64	57	$0,89$

5.6. Matriz imputada para Montevideo con el modelo GE

La matriz OD imputada para Montevideo utilizando el modelo que considera filtros espaciales es la siguiente:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**16**	**17**	**18**
1	12	6	2	2	3	0	1	1	1	0	1	1	0	1	2	1	0
2	32	19	5	7	7	2	1	2	2	1	3	1	1	1	1	1	0
3	11	8	16	7	12	1	3	1	1	1	1	0	1	1	3	1	1
4	3	3	1	3	2	1	3	1	1	0	2	2	0	1	1	0	0
5	31	13	2	10	21	3	1	1	1	0	1	0	1	1	1	1	0
6	9	5	3	3	4	3	3	1	2	1	1	0	0	1	1	0	0
7	17	12	3	8	8	1	2	1	3	1	2	0	1	1	1	0	0
8	6	2	1	1	3	1	3	5	1	1	1	1	0	0	2	0	0
9	16	21	14	10	12	6	5	6	41	3	12	2	1	3	4	5	3
10	8	8	5	11	15	1	11	1	9	15	7	1	4	3	2	1	1
11	13	6	8	1	12	4	4	4	8	7	21	1	2	2	2	1	1
12	7	1	1	2	4	1	1	0	1	1	2	7	1	2	1	1	1
13	3	11	4	2	2	2	1	2	3	1	2	3	20	7	4	1	1
14	16	10	8	7	7	1	3	1	3	1	3	1	4	20	1	5	1
16	7	2	2	1	2	0	1	0	1	0	1	2	1	3	4	1	1
17	12	9	8	6	18	1	2	1	1	1	4	2	2	7	4	6	3
18	2	2	5	1	1	1	1	0	3	1	1	1	4	3	1	2	13

Los ceros de la tabla imputada corresponden a valores menores a 0.5. Como se trata de valores de conteo, los valores de las predicciones se redondearon al entero más cercano.

A continuación se presentan los histogramas de los valores de la matriz OD antes y después de la imputación:

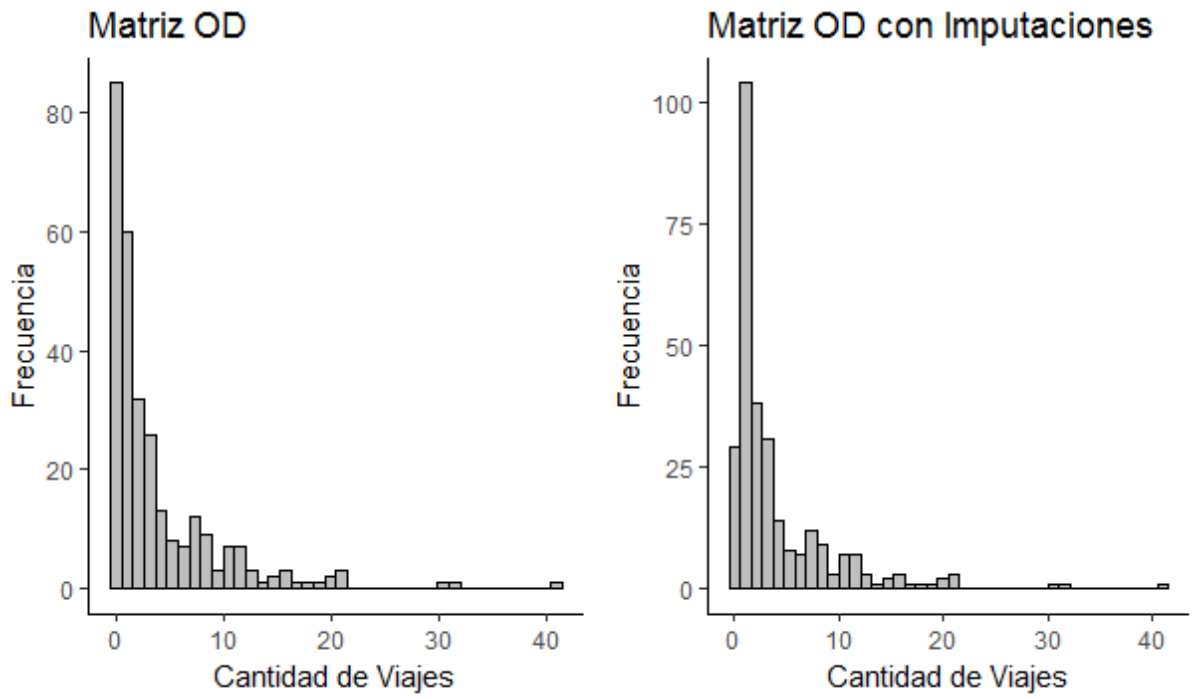

Figura 5.3: Frecuencias de flujos antes y después de la imputación

Los valores imputados corresponden todos a flujos pequeños (de hasta 4 viajes). Esto es esperable debido a que el tamaño de muestra no logra capturar los flujos de viajes pequeños. El mayor numero de imputaciones fue de un viaje, con una frecuencia de 44 imputaciones.

El mayor valor imputado fue de 4 viajes, corresponde al flujo con origen en el CCZ 11, correspondiente a los barrios de Villa Española, Cerrito y parte de la Unión, y como destino el CCZ 6, correspondiente principalmente al barrio de la Unión y Malvín Norte. Estos dos CCZ son vecinos y el barrio de la Unión es parte de ambos. Cabe destacar que la Unión es uno de las centralidades de la ciudad, siendo un atractor de viajes para las zonas aledañas.

Como se dijo anteriormente, la matriz OD imputada sigue teniendo ceros: con la imputación se logró reducir la cantidad de ceros de un 29.4 % a un 10 %. Analizando desde el origen de los viajes, se observa que el CCZ 1 es el que tiene el mayor porcentaje de celdas vacías en origen, luego de la imputación. Este CCZ corresponde a los barrios Ciudad Vieja y Centro, más característicos por ser atractores que generadores de viajes.

Por otro lado, desde el punto de vista de los destinos, el CCZ 18 es el que retiene más ceros. Este CCZ, ubicado en la periferia oeste de Montevideo, es tan poco generador como atractor de viajes, concentrándose el 50 % de ellos en la diagonal de la matriz.

Además de conseguir con la imputación la reducción de ceros a un nivel menor, es importante contar con estas estimaciones de viajes ya que considerar que no hay viajes entre los flujos

imputados puede llevar a la desatención de los temas de movilidad entre los CCZ involucrados. Máxime teniendo en cuenta que se trata en su mayoría de CCZ periféricos, los cuales suelen tener algún tipo de déficit en cuanto al transporte entre ellos. De esta manera se evita el posible error de no considerarlos dentro de las políticas de movilidad cuando la ausencia de flujos se debe a un tema de limitaciones de la muestra y no a la inexistencia de viajes entre estos CCZ.

Capítulo 6. Conclusiones y Líneas futuras de Investigación.

El modelo gravitacional que incorpora los filtros espaciales muestra un desempeño superior frente al modelo gravitacional tradicional cuando el tamaño de los flujos faltantes es pequeño. Si sólo se considera la cantidad de celdas vacías, el modelo con filtros espaciales es mejor, hasta que el porcentaje de celdas vacías es aproximadamente de un 40 %.

El escenario más realista, es el que corresponde al de una matriz OD con faltantes en celdas en donde la cantidad de viajes se presume que es pequeña. De no ser así, ya no se estaría afrontando un problema de imputación de la matriz OD, sino un problema de muestreo o de regionalización inadecuada de la matriz OD. Desde el punto de vista del análisis del desempeño de los modelos es de interés analizar el poder predictivo ante la falta de flujos medianos o grandes, pero si este fuera el caso, imputar los valores faltantes no sería la mejor solución.

Se parte de una matriz con un 29.4 % de celdas vacías, y luego de la imputación este porcentaje se reduce a un 10 % de las celdas de la matriz. Esto implica que para estas regiones se podrían tratar de ceros estructurales, es decir, que entre esos CCZ la frecuencia de viajes es casi nula.

Una primera limitación del trabajo realizado es que debido a la implementación de los filtros espaciales, a medida que las celdas vacías aumentan se pierde información, al tener que eliminar vectores para que los modelos sean conformables. Una alternativa a considerar sería la propuesta por LeSage y Pace (2008), en donde se parte de un modelo endógeno, y se definen "rezagos" espaciales a nivel de los flujos. Hasta el momento este enfoque puede ser aplicado desde el punto de vista computacional asumiendo una distribución Normal, lo que no es correcto asumir para flujos de transporte. No se encuentran desarrolladas librerías que trabajen con datos de flujos y en donde se pueda asumir una distribución Binomial Negativa o Poisson.

Una segunda limitación, es que no se tienen en cuenta los pesos muestrales de la encuesta. La incorporación de los pesos muestrales es un desafío a futuro en el sentido de que los modelos gravitacionales se construyen sobre una estimación (la de la matriz OD), no sobre los datos originales de la encuesta, y el cálculo de la varianza de la estimación de los parámetros

no pueden realizarse actualmente con las librerías convencionales de muestreo. Una forma de considerar los pesos de la muestra sería implementando un procedimiento *bootstrap*, en donde se puedan obtener diferentes estimaciones de la matriz OD, y así poder llegar a una estimación de la varianza. Los filtros espaciales deberían incorporarse en cada paso de este proceso.

Por ultimo, el uso del ARMSE como medida de error da una idea global de como funciona cada método de acuerdo a la cantidad y magnitud de cada simulación de datos faltantes. Se podría agregar una medida de error relativo para cada una de las simulaciones de n-uplas de tal manera de poder ver específicamente en cada imputación de celda faltante la diferencia entre el valor real y la imputación con cada modela. Con este error relativo se podría tener una idea de que tan alejadas han sido las imputaciones en si mismas, más allá de que modelo resulta mejor según la comparación por ARMSE.

Además de las posibilidades antes mencionadas, se visualizan otras líneas de investigación a futuro. La primera tiene que ver con la forma de medir las distancias. En cuanto a las distancias se podrían utilizar las que surgen de la propia encuesta, o definir algún otro tipo de medida. Esto puede hacer que los resultados del modelo varíen, principalmente en la estimación del parámetro de la función de impedancia.

La segunda se refiere a la definición de la matriz W y la matriz W_{od}. La primera ha sido definida por contigüidad a pesar de que se podrían haber usado otras definiciones, como el k-esimo vecino más cercano o una distancia mínima entre centroides. A su vez los pesos espaciales podrían haber sido ponderados, por ejemplo, por el inverso de la distancia entre los centroides, penalizando aquellas regiones contiguas pero con sus centroides distanciados. La matriz W_{od} por su parte también puede ser definida de otras maneras, como propuso, por ejemplo, LeSage and Pace (2008). Como línea futura de investigación, se podría introducir al análisis otros formatos para las matrices W y W_{od} para medir el impacto de estas diferentes definiciones en la estimación de los modelos.

Una tercera linea a seguir en futuras investigaciones consiste en agregar, además de los escenarios según cantidad de celdas vacías y tamaño de las mismas, la distribución de estas en las regiones de la matriz O-D. Generando diferentes patrones simulados de celdas vacías en regiones especificas de la ciudad.

Finalmente, una cuarta línea de investigación es la inclusión de variables explicativas al modelo. En un contexto de imputación de datos faltantes, el uso de variables auxiliares podría ser crucial para mejorar la efectividad de la imputación. Además, es importante analizar si la presencia de variables auxiliares pueden capturar, al menos en parte, la autocorrelación espacial en los flujos, reduciendo así algunos de los vectores propios agregados al modelo. En el caso de que algunas de las variables auxiliares incluidas en el modelo muestren una alta correlación con alguno de

los filtros espaciales, se podría asociar estas variables a un patrón espacial, encontrando la causa de la autocorrelación espacial, al menos de manera parcial.

Referencias

Referencias

Bivand, R. S., Pebesma, E., y Gómez-Rubio, V. (2013). *Applied spatial data analysis with r*. New York: Springer.

Cameron, A. C., y Trivedi, P. (2013). *Regression analysis of count data*. New York, NY: Cambridge University Press.

Chambers, J., Hastie, T., y Pregibon, D. (1990). Statistical models in s. En *Momirović k., mildner v. (eds) compstat.* Physica-Verlag HD.

Chun, Y. (2008). Modeling network autocorrelation within migration flows by eigenvector spatial filtering. *Journal of Geographical Systems*, *10*, 317–344.

Distancias himmera. (s.f.). `http://es.distancias.himmera.com/`.

Fischer, M., y Wang, J. (2013). *Spatial data analysis: Models, methods, and techniques*. New York: Springer.

Flowerdew, R., y Aitkin, M. (1982). A method of fitting the gravity model based in the poisson distribution. *Journal of Regional Science*, *22*(2), 191–202.

Fotheringham, A. S. (1983). A new set of spatial-interaction models: the theory of competing destinations. *Environment & Planning A*, *15*, 15–36.

Furness, K. P. (1970). Time function interaction. *Traffic Engineering and Control*, *7*(7), 19–36.

Gaetan, C., y Guyon, X. (2010). *Spatial statistics and modeling*. New York: Springer.

Google. (s.f.). *Rutas sugeridas para conducir entre los centroides de los cczs.* Descargado 2018-11-01, de `https://www.google.com.uy/maps`

Griffith, D. A. (2009). Modeling spatial autocorrelation in spatial interaction data: empirical evidence from 2002 germany journey-to-work flows. *Journal of Geographical Systems*, *11*(2), 117–140.

Griffith, D. A., y Jones, K. G. (1980). Explorations into the relationship between spatial structure and spatial interaction. *Environment and Planning*, *12*, 187–201.

Intendencia de montevideo. (s.f.). `https://montevideo.gub.uy/aplicacion/como-ir`.

Jayet, H. (1990). Spatial search processes and spatial interaction. 1. sequential search, intervening opportunities, and spatial search equilibrium. *Environment & Planning A*, *22*(5), 583–599.

Jou, Y. J., Cho, H. J., Lin, P. W., y Wang, C. Y. (2006). Incomplete information analysis

for the origin - destination survey table. *Journal of Urban Planning and Development*, *132*(4). doi: 10.1061/(ASCE)0733-9488(2006)132:4(193)

LeSage, J., y Pace, R. K. (2008). Spatial econometric modeling of origin- destination flows. *Journal of Regional Science*, *48*(5), 941–967.

LeSage, J., y Pace, R. K. (2009). *Introduction to spatial econometrics*. United Kingdom: Chapman & Hall.

LeSage, J. P., y Fischer, M. (2010). Spatial econometric methods for modeling origin-destination flows. En *Handbook of applied spatial analysis.* Berlin, Heidelberg: Springer - Verlag.

O'Kelly, M. E. (2009). Spatial interacion models. En *International encyclopedia of human geography.* Elsevier.

Sig - intendencia de montevideo. (s.f.). `https://sig.montevideo.gub.uy/`.

Ten Have, T. R. (2005). Structural and sampling zeros. En *Encyclopedia of biostatistics.* New York, NY: Wiley.

Wilson, A. G. (1967). A statistical theory of spatial distribution models. *Transportation Research*, *1*, 253–269.

Anexos

Descripción de algunos filtros

Los filtros espaciales seleccionados por el modelo capturan la correlación espacial de los flujos, y cada uno de ellos muestra un patrón asociado a dicha correlación. Como se mencionó en los párrafos anteriores es de interés explorar la estructura de los filtros espaciales para desarrollar un método de imputación de las celdas vacías de la matriz OD.

Los vectores seleccionados por el modelo son en total 23[1]. Los vectores de menor orden se asocian a un patrón global, mientras que los de mayor orden se asocian a patrones locales. A modo de ejemplo se presentan los *heatmap* para los vectores 2, 10 y 98.
Cuando los colores del Heatmap se asocian a barras verticales, implica una correlación en destino. Las barras horizontales se asocian al efecto espacial del origen, y cuando se observan ambas a efectos de origen - destino. En nuestro caso mayormente se observan asociaciones en destino y en origen - destino.

Para el Vector 2 (Figura 6.1) la asociación es en destino, y los CCZ se dividen en dos grandes grupos: los CCZ 12, 13, 14, 16, 17 y 18 (Noroeste de Montevideo) y el resto de los CCZ (Centro - Este de Montevideo). Dentro de este último grupo se observa un subgrupo (CCZ 1, 2, 3, 4, 5, 14 y 16) que constituyen una franja central en el departamento.

El Vector 10 (Figura 6.2) es un ejemplo de asociación en origen - destino, mostrando un patrón que distingue cuatro grupos de flujos. En origen distingue una región céntrica (CCZ 1, 2, 3, 4, 5, 14, 16), que divide los flujos entre dos regiones en destino, una céntrica y otra periférica.
El Vector 98 (Figura 6.3)muestra un mosaico con efectos "micro" en origen - destino.

[1]En orden de selección, vectores 4, 23, 2, 30, 73, 28, 98, 27, 74, 43, 71, 56, 225, 80, 37, 47, 49, 125, 202, 32, 10, 12, 40

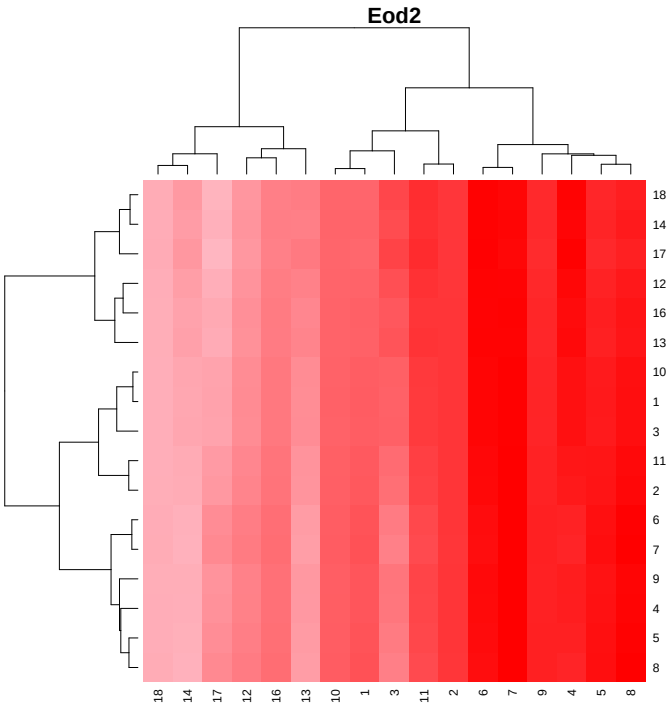

Figura 6.1: Heatmap para el Vector 2 del Filtrado Espacial

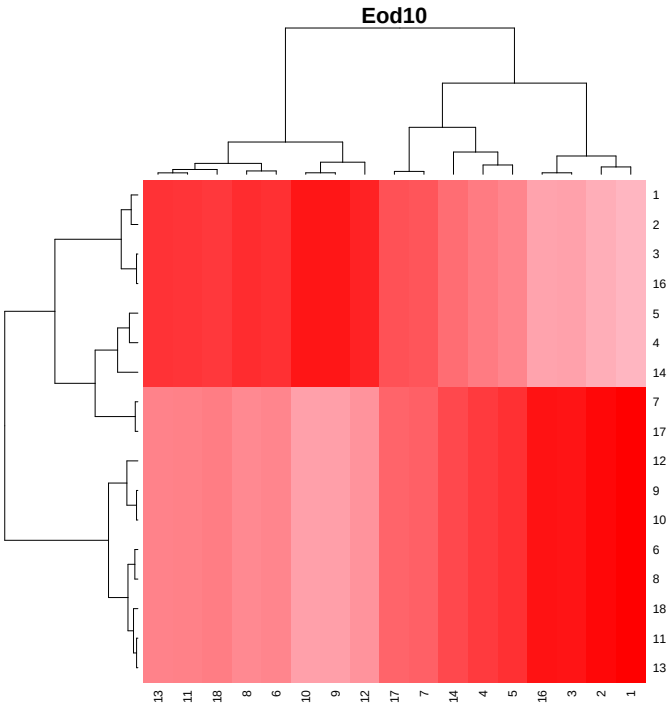

Figura 6.2: Heatmap para el Vector 10 del Filtrado Espacial

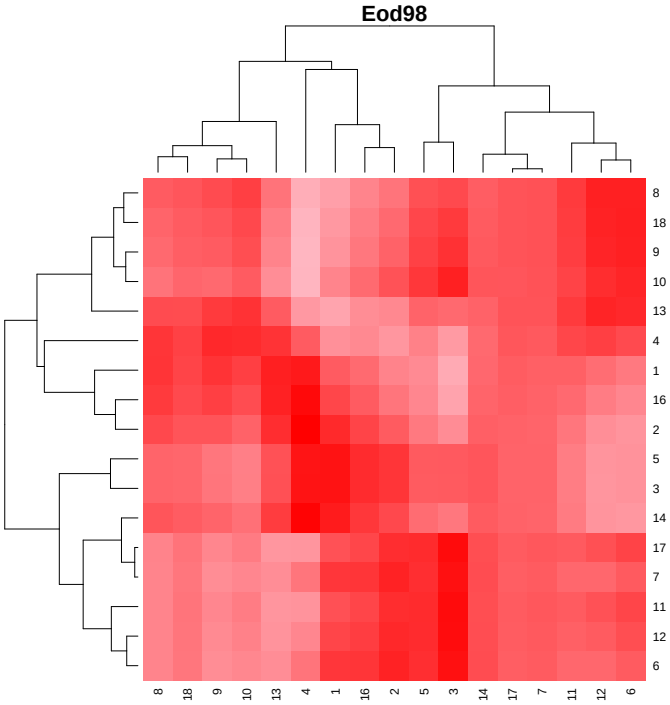

Figura 6.3: Heatmap para el Vector 98 del Filtrado Espacial

Función para el filtrado espacial

```
# funcion spfod (spatial filtering origin destination)
# tiene como argumentos:
#  1) un modelo lm o glm (ajustado a los flujos)
#  2) matriz w (de los poligonos, no de los flujos)
#  3) alfa (nivel de significacion)
#  4) tipo indica si se centra con respecto a los regresores o a la media
# requiere de la libreria spdep para correr en cada paso moran.test
# devuelve:
#  1) el modelo con los vectores propios seleccionados
#  2) el historial de valores (y p-valores) del indice de moran

spfod<-function(mod,w,alfa=0.05,metodo='chun',tipo='residuos',sacar=null){
  stopifnot(any(class(mod) %in% c('lm','glm','glmermod','lmermod')))
  if(!tipo%in%c('residuos','centrar')) stop('tipo debe ser residuos o centrar')
  if(!metodo%in%c('chun','ww')) stop('tipo debe ser chun o ww')
  if(any(c('glmermod','lmermod')%in%class(mod)))
  datos<-mod@frame else datos<-mod$model
  # carga spdep
  if(! 'spdep' %in% loadednamespaces()) library(spdep)

  n <- nrow(datos)
  n <- nrow(w)
  x <- model.matrix(mod)
```

```
# matrices w
wd  <- diag(n) %x% w
wo  <- w %x% diag(n)
if (metodo=='chun'){

wod=1/2*(wd+wo)

if (!is.null(sacar)){
  wod=wod[-sacar,-sacar]
} #para cv

if (tipo=='residuos') {
  m <- diag(n) - x%*%solve(t(x)%*%x)%*%t(x)
} else {
  m <- diag(n) - matrix(1,n,1)%*%matrix(1,1,n)/n
}
#mwodm <- m %*% wod %*% m
mwodm <- m %*% (1/2*(wod+t(wod))) %*% m
# vectores propios
vpod <- eigen(mwodm)$vectors
# saco el asociado al valor propio cero
vpod<-vpod[,-which.min(apply(vpod,2,var))]
colnames(vpod)<-paste('eod',1:ncol(vpod),sep='')

# paso 0
moran0od<- moran.test(residuals(mod),mat2listw(wod))
m_od<-data.frame(paso=0,i=as.numeric(moran0od$statistic),
p.value=moran0od$p.value)
pvod<-m_od$p.value
if (pvod>alfa) {
  cat('\n','no se detecto autocorrelacion espacial','\n')
  return(list(mod=mod,m=null))
}
cat('\n');print(m_od)

k<-0
minpv<-pvod
columnas=ncol(vpod)
sig<-rep(0,columnas)
vpmod=c()

while (minpv<=alfa) {
  k<-k+1
  # origen-destino
  if (m_od$p.value[nrow(m_od)]<=alfa){
    for (i in 1:ncol(vpod)){
      f.i<-as.formula(paste('.~.',colnames(vpod)[i],sep='+'))
      d.i<-data.frame(datos,vpod[,i,drop=false])
```

```
      m.i<-update(mod,f.i,data=d.i)
      sig[i]<-1-pchisq(-2*(loglik(mod)[1]-loglik(m.i)[1]),1)
    }
    vpmax<-which.min(sig)
    datos<-data.frame(datos,vpod[,vpmax,drop=false])
    f<-as.formula(paste('.~.',colnames(vpod)[vpmax],sep='+'))
    mod<-update(mod,f,data=datos)
    vpmod=cbind(vpmod,vpod[,vpmax])
    colnames(vpmod)[k]=colnames(vpod)[vpmax]

    mk<-moran.test(residuals(mod),mat2listw(wod))
    m_od<-rbind(m_od,c(k,as.numeric(mk$statistic),mk$p.value))
    cat('\n','origen-destino','\n');print(m_od)

    vpod<-vpod[,-vpmax]
    pvod<-mk$p.value
  }
  columnas=columnas-1
  sig<-r2<-rep(0,columnas)
  minpv<-pvod
}
return(list(mod=mod,m=list(od=m_od),v=list(vpmod)))
} else {
wod <- w %x% w
if (!is.null(sacar)){
  wod=wod[-sacar,-sacar]
}

if (tipo=='residuos') {
  m <- diag(n) - x%*%solve(t(x)%*%x)%*%t(x)
} else {
  m <- diag(n) - matrix(1,n,1)%*%matrix(1,1,n)/n
}
mwom  <- m %*% (1/2*(wo+t(wo))) %*% m
mwdm  <- m %*% (1/2*(wd+t(wd))) %*% m
mwodm <- m %*% (1/2*(wod+t(wod))) %*% m
# vectores propios
vpo  <- eigen(mwom)$vectors
vpd  <- eigen(mwdm)$vectors
vpod <- eigen(mwodm)$vectors
# saco el asociado al valor propio cero
vpo<-vpo[,-which.min(apply(vpo,2,var))]
vpd<-vpd[,-which.min(apply(vpd,2,var))]
vpod<-vpod[,-which.min(apply(vpod,2,var))]
colnames(vpo)<-paste('eo',1:ncol(vpo),sep='')
colnames(vpd)<-paste('ed',1:ncol(vpd),sep='')
colnames(vpod)<-paste('eod',1:ncol(vpod),sep='')
```

```
# paso 0
moran0o <- moran.test(residuals(mod),mat2listw(wo))
moran0d <- moran.test(residuals(mod),mat2listw(wd))
moran0od<- moran.test(residuals(mod),mat2listw(wod))
m_o <-data.frame(paso=0,i=as.numeric(moran0o$statistic),
p.value=moran0o$p.value)
m_d <-data.frame(paso=0,i=as.numeric(moran0d$statistic),
p.value=moran0d$p.value)
m_od<-data.frame(paso=0,i=as.numeric(moran0od$statistic),
p.value=moran0od$p.value)
pvo<-m_o$p.value
pvd<-m_d$p.value
pvod<-m_od$p.value
if (min(pvo,pvd,pvod)>alfa) {
  cat('\n','no se detecto autocorrelacion espacial','\n')
  return(list(mod=mod,m=null))
}
cat('\n');print(m_o);print(m_d);print(m_od)

k<-0
minpv<-min(pvo,pvd,pvod)
columnas=ncol(vpo)
sig<-rep(0,columnas)
vpmod=c()
vpmo=c()
vpmd=c()

while (minpv<=alfa) {
  k<-k+1
  # origen
  if (m_o$p.value[nrow(m_o)]<=alfa){
    for (i in 1:ncol(vpo)){
      f.i<-as.formula(paste('.~.',colnames(vpo)[i],sep='+'))
      d.i<-data.frame(datos,vpo[,i,drop=false])
      m.i<-update(mod,f.i,data=d.i)
      sig[i]<-1-pchisq(-2*(loglik(mod)[1]-loglik(m.i)[1]),1)
    }
    vpmax<-which.min(sig)
    datos<-data.frame(datos,vpo[,vpmax,drop=false])
    f<-as.formula(paste('.~.',colnames(vpo)[vpmax],sep='+'))
    mod<-update(mod,f,data=datos)
    vpmo=cbind(vpmo,vpo[,vpmax])
    colnames(vpmo)[k]=colnames(vpo)[vpmax]

    mk<-moran.test(residuals(mod),mat2listw(wo))
    m_o<-rbind(m_o,c(k,as.numeric(mk$statistic),mk$p.value))
    cat('\n','origen','\n');print(m_o)
```

```
    vpo<-vpo[,-vpmax]
    pvo<-mk$p.value
  }
  sig<-sig*0
  # desetino
  if (m_d$p.value[nrow(m_d)]<=alfa){
    for (i in 1:ncol(vpd)){
      f.i<-as.formula(paste('.~.',colnames(vpd)[i],sep='+'))
      d.i<-data.frame(datos,vpd[,i,drop=false])
      m.i<-update(mod,f.i,data=d.i)
      sig[i]<-1-pchisq(-2*(loglik(mod)[1]-loglik(m.i)[1]),1)
    }
    vpmax<-which.min(sig)
    datos<-data.frame(datos,vpd[,vpmax,drop=false])
    f<-as.formula(paste('.~.',colnames(vpd)[vpmax],sep='+'))
    mod<-update(mod,f,data=datos)
    vpmd=cbind(vpmd,vpd[,vpmax])
    colnames(vpmd)[k]=colnames(vpd)[vpmax]

    mk<-moran.test(residuals(mod),mat2listw(wd))
    m_d<-rbind(m_d,c(k,as.numeric(mk$statistic),mk$p.value))
    cat('\n','destino','\n');print(m_d)

    vpd<-vpd[,-vpmax]
    pvd<-mk$p.value
  }
  sig<-sig*0
  # origen-destino
  if (m_od$p.value[nrow(m_od)]<=alfa){
    for (i in 1:ncol(vpod)){
      f.i<-as.formula(paste('.~.',colnames(vpod)[i],sep='+'))
      d.i<-data.frame(datos,vpod[,i,drop=false])
      m.i<-update(mod,f.i,data=d.i)
      sig[i]<-1-pchisq(-2*(loglik(mod)[1]-loglik(m.i)[1]),1)
    }
    vpmax<-which.min(sig)
    datos<-data.frame(datos,vpod[,vpmax,drop=false])
    f<-as.formula(paste('.~.',colnames(vpod)[vpmax],sep='+'))
    mod<-update(mod,f,data=datos)
    vpmod=cbind(vpmod,vpod[,vpmax])
    colnames(vpmod)[k]=colnames(vpod)[vpmax]

    mk<-moran.test(residuals(mod),mat2listw(wod))
    m_od<-rbind(m_od,c(k,as.numeric(mk$statistic),mk$p.value))
    cat('\n','origen-destino','\n');print(m_od)

    vpod<-vpod[,-vpmax]
    pvod<-mk$p.value
```

```
    }
    columnas=columnas-1
    sig<-r2<-rep(0,columnas)
    minpv<-min(pvo,pvd,pvod)
  }
  vpmod=cbind(vpmo,vpmd,vpmod)
  return(list(mod=mod,m=list(origen=m_o,destino=m_d,od=m_od),v=list(vpmod)))
}}
```

Programa para imputación de las celdas vacías de la matriz de Origen - Destino

```
Fmetccz <- function(arch,vecp)
{
  nom <- paste('datos/temp_BOG/',arch,sep='')
  load(file=eval(nom)) #tAUX

  cuales <- which(is.na(tAUX$counts))
  mod <- glm.nb(counts~dist,data=tAUX) # modelo con valores iniciales

  mod.E <- spfod(mod,W=w,alfa=0.05,metodo=T,sacar=cuales,residuos=F)
  vecp=vecp[,which(colnames(vecp) %in% colnames(mod.E$v))]

  tAUX.E <- cbind(tAUX,vecp) #agrego los vec prop a la tabla

  ##prediccion
  aux_pred<- predict(mod.E$mod,newdata=tAUX.E,type="response")
  imputacion.E <- aux_pred[cuales];attributes(imputacion.E) <- NULL
 }
```