Imputación de las celdas vacías de la matriz de Origen - Destino

María Eugenia Riaño, Fernando Massa, Antonio Rey, Martín Hansz y Andrés Castrillejo

Proyecto "Calibración de un modelo de interacción espacial para la modelización de la movilidad de viajes por trabajo en Montevideo", Fondo Sectorial de Investigación a partir de Datos, ANII 2017.

5 de junio de 2019

Introducción

Matriz de Origen - Destino (OD)

La matriz de Origen - Destino es una tabla de dos dimensiones, donde cada celda representa el volumen de tráfico de una región a otra.

En la matriz OD se oberva:

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

Ceros estructurales y ceros muestrales

Los ceros estructurales y muestrales son celdas con frecuencia cero en una tabla de contingencia.

	R_1	R_2	R_3	R_4
R_1	50	20	30	15
R_2	20	40	0	8
R_3	10	0	25	30
R_4	60	50	10	15

Ceros estructurales y ceros muestrales

Cero Estructural

El cero estructural se debe a la imposibilidad de observar una combinación dada de las categorías de la tabla de contingencia.

$\Rightarrow$ Modelos *Zero Inflated*, Tobit, etc.

Ceros estructurales y ceros muestrales

Cero Muestral

Los ceros muestrales se deben a que el tamaño de muestra es pequeño para la desagregación que contiene la tabla de contingencia.

$\Rightarrow$ Dependiendo del objetivo:

- Imputación.
- Modelos de Conteo, especificación Poisson o Binomial Negativa.

¿Qué es imputar?

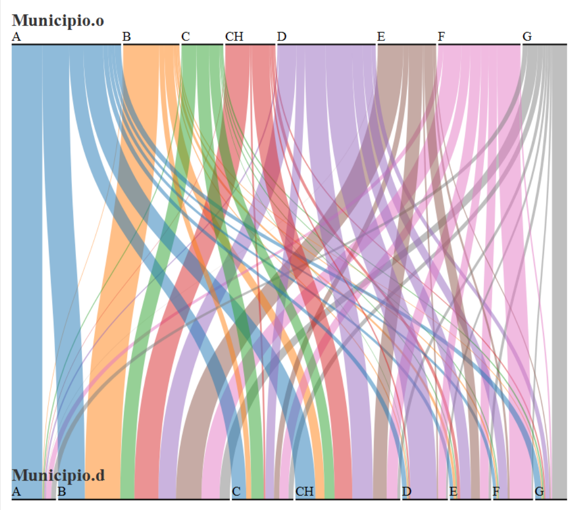
En Estadística, la imputación es la sustitución de valores no informados en una observación por otros.

En el contexto de matrices de Origen - Destino (OD) imputar es predecir el flujo de tráfico en las celdas vacías de la matriz.

Matriz OD para viajes por trabajo en Montevideo considerando Municipios

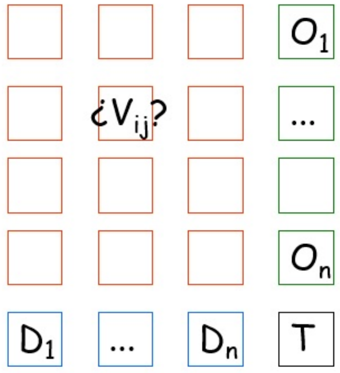
	A	B	C	CH	D	E	F	G
A	60	51	27	39	9	7	6	12
B	1	69	10	19	2	5	3	2
C	2	28	25	20	1	2	2	3
CH	1	50	4	36	3	6	0	3
D	3	35	17	39	50	21	17	8
E	0	51	11	27	3	18	4	2
F	11	37	18	22	15	17	41	3
G	10	22	10	6	4	5	3	31

Matriz OD para viajes por trabajo en Montevideo considerando Municipios



Métodos para estimación de la matriz OD

Los métodos para estimación de matrices OD parten de las marginales de la tabla y estiman el valor de las celdas.



Métodos para estimación de la matriz OD

- Determinísticos:
  - Factor de Crecimiento.
  - Biproporcional (Deming - Stephan).
  - Máxima Entropía (Wilson, 1967).
- Probabilísticos:
  - Modelos Gravitacionales.
  - Modelos Logit Multinomial.
  - Otros.

Modelos Restringidos

Restringido en Origen: El modelo reproduce el total de viajes producidos.

Restringido en Destino: El modelo reproduce el total de viajes atraídos.

Doble Restringido: El modelo reproduce ambas marginales.

Sin Restricciones: El modelo no reproduce ninguna de las marginales.

Modelo Gravitacional

$$y_{ij} = A_i B_j {X_{o_i}}^{\beta_o} {X_{d_j}}^{\beta_d} c_{ij}^{\phi} \tag{1}$$

- y_{ij} es la cantidad de viajes del origen i al destino j.
- A_i y B_j son factores de balance.
- X_{o_i} y X_{d_j} son características asociadas al origen y al destino respectivamente.
- c_{ij}^{ϕ} es la función de impedancia o disuasión, donde c_{ij} es una medida del costo (distancia, tiempo) entre el origen i y el destino j.
- β_o, β_d y ϕ son los parámetros a estimar.

El modelo gravitacional asume INDEPENDENCIA entre los flujos de origen - destino.

¿Qué es la autocorrelación espacial?

Es la correlación entre los valores de una variable, estrictamente atribuible a la proximidad de las unidades geográficas en el espacio.

Tobler's First Law of Geography (1970)

"Everything is related to everything else, but near things are more related than distant things."

La autocorrelación espacial implica que no se cumpla el supuesto de independecia de las observaciones de la estadística clásica.

¿Cómo se incorpora la correlación espacial al Modelo Gravitacional?

- Modelos Endógenos (Rezagados): SAR, LagSar (espaciales autorregresivos).

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

- Filtrado Espacial.

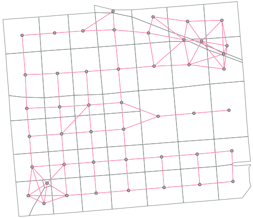
$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{E} \boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

Filtrado Espacial

Sea n la cantidad de unidades geográficas en el mapa. Los valores "vecinos" de identifican con una matriz de conectividad de dimensión $n \times n$ binaria, a la que denominaremos $\mathbf{C}$, tal que

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

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix} \Rightarrow$$



La matriz $\mathbf{W}$ de pesos espaciales, es la matriz $\mathbf{C}$ ponderada.

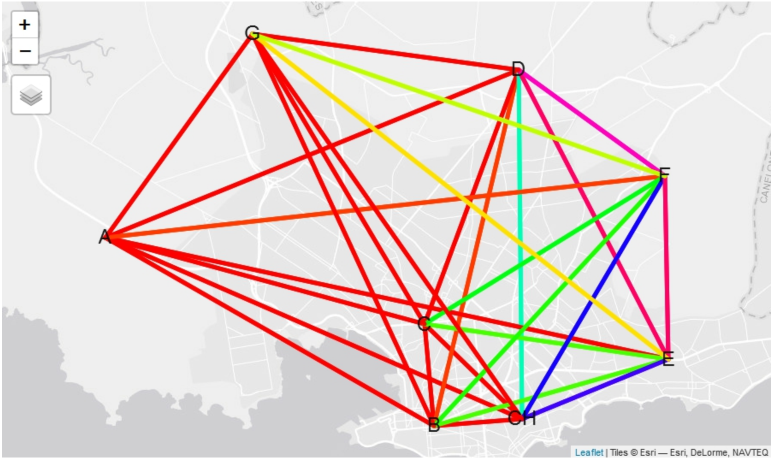
¿Cómo se obtiene el conjunto de variables proxy $\mathbf{E}$?

A partir de los vectores propios de una transformación de la matriz $\mathbf{W}$ se define un subconjunto de vectores que aproxima a la matriz $\mathbf{E}$.

- El subconjunto se elige mediante un proceso *stepwise forward*, utilizando como función objetivo al Índice de Moran de los residuos del modelo.
- Cada vector $\mathbf{E}_i$ define un patrón en el mapa.
- Los primeros vectores captan fenómenos globales, los últimos captan fenómenos a escala micro.

Algunos vectores resultantes del Filtrado Espacial

Figure 1: Primer Vector del Filtrado Espacial



Algunos vectores resultantes del Filtrado Espacial

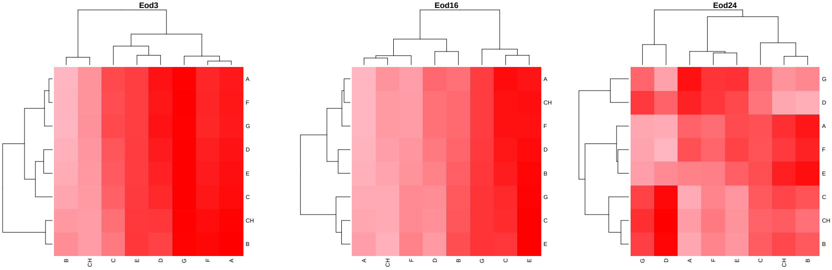
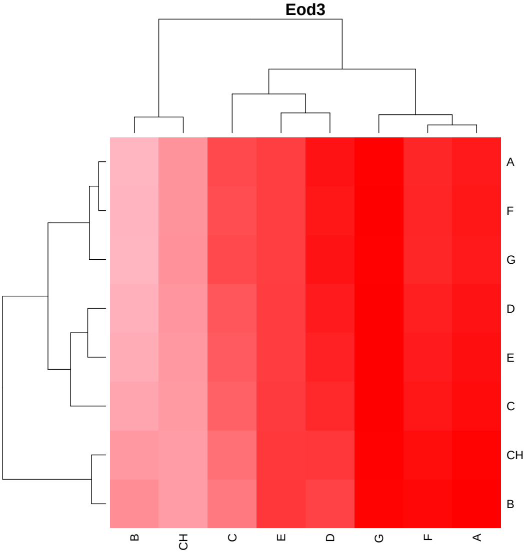
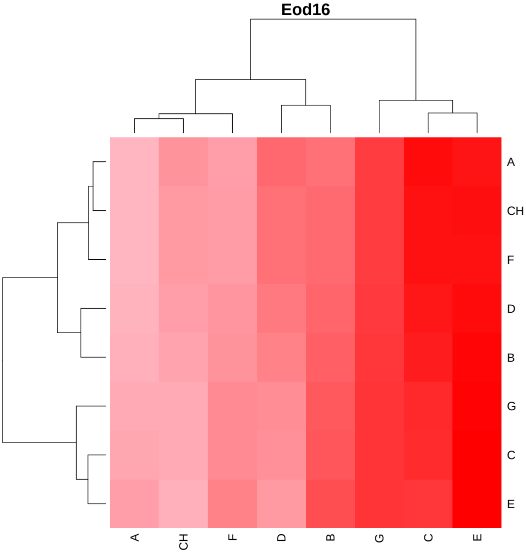


Figure 2: Heatmaps para los vectores propios del Filtrado Espacial

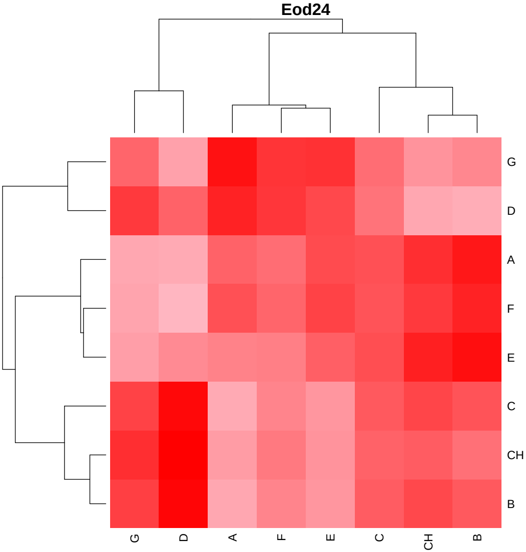
Vector 3



Vector 16



Vector 24



Metodología

Para Municipios:

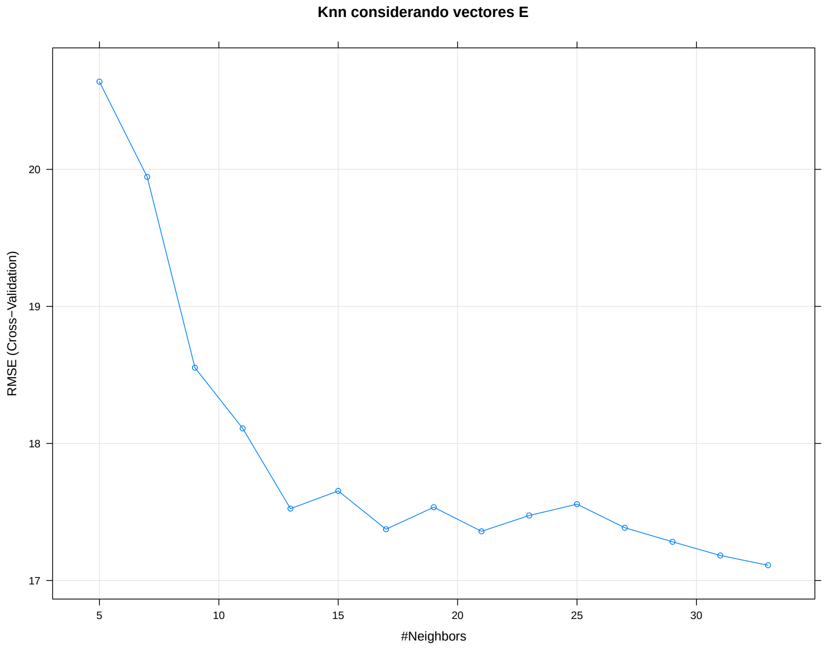
- Imputación modelo dependiente:
  - Modelo Gravitacional.
  - Modelo Gravitacional considerando correlación espacial.
  - Modelo Gravitacional considerando correlación espacial y un efecto en la diagonal.
- Imputación modelo independiente:
  - KNN con vectores propios (filtros) como input.
  - KNN con vectores propios (filtros) y distancia como input.
  - KNN con vectores propios (filtros), distancia y efecto en la diagonal como input.
- Comparación de resultados.

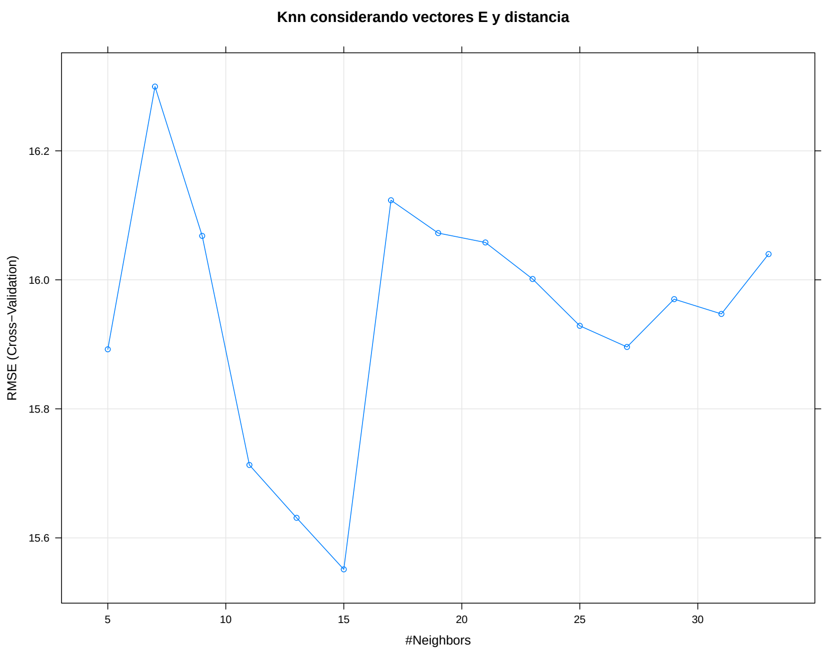
Resultados

Table 1: Comparación de Resultados de la Imputación

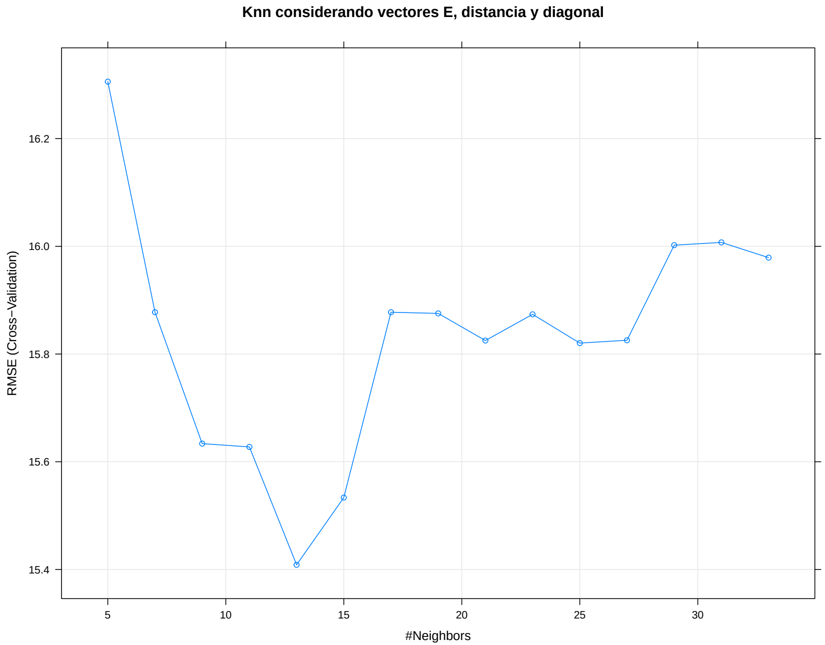
Modelo	Valor CH, F	Valor E, A	$RMSE$
Gravitacional	10.53	4.5	13.81
Gravitacional + **E**	2.04	0.84	10.34
Gravitacional + **E** + Diag	2.28	1.81	$\boxed{9.17}$
KNN **E**	17.5	20.3	16.8
KNN **E** + c_{ij}	11.7	11.4	14.21
KNN **E** + c_{ij} + Diag	11.15	11.07	$\boxed{13.65}$

Obs: cada observación corresponde aproximadamente a 400 viajes.


Knn considerando vectores E
RMSE (Cross−Validation)
#Neighbors
17
18
19
20
5
10
15
20
25
30



Knn considerando vectores E y distancia
RMSE (Cross−Validation)
16.2
16.0
15.8
15.6
5
10
15
20
25
30
#Neighbors


Knn considerando vectores E, distancia y diagonal
RMSE (Cross−Validation)
#Neighbors
16.2
16.0
15.8
15.6
15.4
5
10
15
20
25
30

Conclusiones

- El mejor resultado para la imputación se obtiene con el modelo considerando correlación espacial y un efecto para la diagonal.
- El desempeño del método KNN iguala al del modelo gravitacional cuando se consideran las distancias y el efecto diagonal.
- Se observan grandes diferencias entre los valores imputados, por lo que se debe considerar el mejor método de acuerdo a si se quiere trabajar con un modelo dependiente, o con un método independiente del modelo.

Lineamientos a Futuro

- Considerar pesos muestrales en la predicción.
- Analizar la perfomance de los métodos en matrices con una proporción alta de celdas vacías.

Bibliografía

- Griffith, D. A. (2003) *Spatial Autocorrelation and Spatial Filtering*, Advances in Spatial Science, Springer - Verlag.
- LeSage, J. y Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Chapman Hall/CRC, Taylor Francis Group.
- Tiefelsdorf, M., Boots, B. (1995). The Exact Distribution of Moran's I. *Environment and Planning A: Economy and Space*, 27(6), 985–999. https://doi.org/10.1068/a270985
- Tiefelsdorf, M., Griffith, D.A. (2007). Semiparametric filtering of spatial autocorrelation: the eigenvector approach. *Environment and Planning*, vol. 39, 1193 - 1221.
- Ten Have, T. R. (2005). *Structural and Sampling Zeros*, Encyclopedia of Biostatistics, Wiley. https://doi.org/10.1002/0470011815.b2a10059.

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

Yaldi, G., Taylor, M. Yue, W. L. (2011). Forecasting origin-destination matrices by using neural network approach: A comparison of testing performance between back propagation, variable learning rate and levenberg-marquardt algorithms. *Australasian Transport Research Forum 2011 Proceedings*, http://www.patrec.org/atrf.aspx.

GRACIAS!!