

Una aproximación a las técnicas de Filtrado Espacial

María Eugenia Riaño, Fernando Massa y Antonio Rey

Grupo de Investigación de Estadística Espacial Aplicada, Instituto de Estadística, Departamento de Métodos Cuantitativos

14 de mayo de 2019

Introducción

El Filtrado Espacial es un método que surge para el tratamiento de la autocorrelación espacial para Datos de Área.

En este caso el espacio se encuentra discretizado en unidades geográficas, y la distancia entre ellas se establece a través de una matriz de conectividad definida por el investigador.

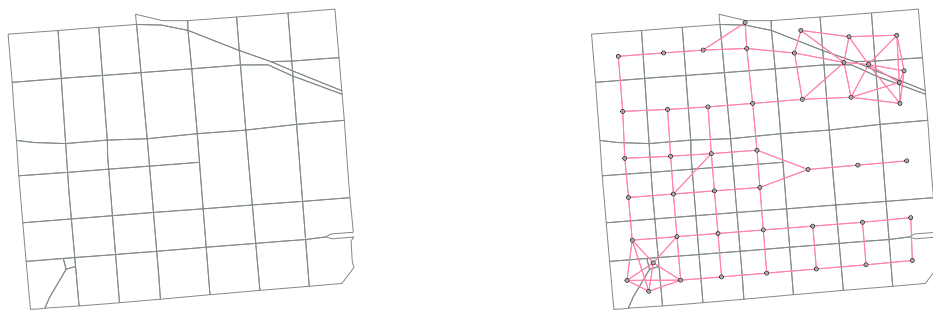


Figure 1: Datos de Área y Matriz de conectividad

¿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 independencia de las observaciones de la estadística clásica.

¿Qué es la autocorrelación espacial?

La autocorrelación espacial en un modelo de regresión puede originarse por:

- ▶ Existencia de un proceso espacial subyacente en las observaciones de la variable de interés.
- ▶ Variables exógenas con un patrón espacial no incluidas en el modelo.
- ▶ Una agregación no apropiada de las unidades geográficas.

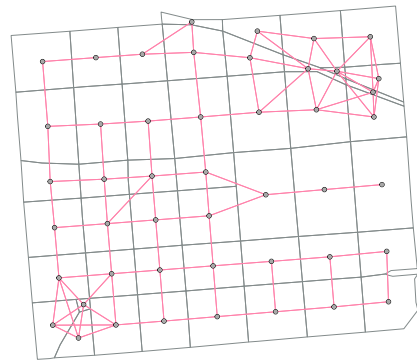
Ignorar la autocorrelación espacial lleva a estimaciones sesgadas de los parámetros de la regresión y de la varianza.

Matriz de Conectividad

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

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

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



Indice de Moran

La correlación espacial puede ser expresada en términos del coeficiente de correlación de Pearson, en donde la variable x corresponde a los valores vecinos de cada unidad.

$$\frac{\sum_{i=1}^n (x_i - \bar{X}) (y_i - \bar{Y}) / n}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2 / n} \sqrt{\sum_{i=1}^n (y_i - \bar{Y})^2 / n}} \Rightarrow$$
$$IM = \frac{n \sum_{i=1}^n \sum_{j=1}^n c_{ij} (y_i - \bar{Y}) (y_j - \bar{Y})}{\left(\sum_{i=1}^n \sum_{j=1}^n c_{ij} \right) \sum_{i=1}^n (y_i - \bar{Y})^2}$$

Índice de Moran

Si se interpreta la autocorrelación espacial como un patrón en el mapa, ésta puede asociarse a tendencias, o mosaicos en la superficie.

Índice de Moran

Si se interpreta la autocorrelación espacial como un patrón en el mapa, ésta puede asociarse a tendencias, o mosaicos en la superficie.

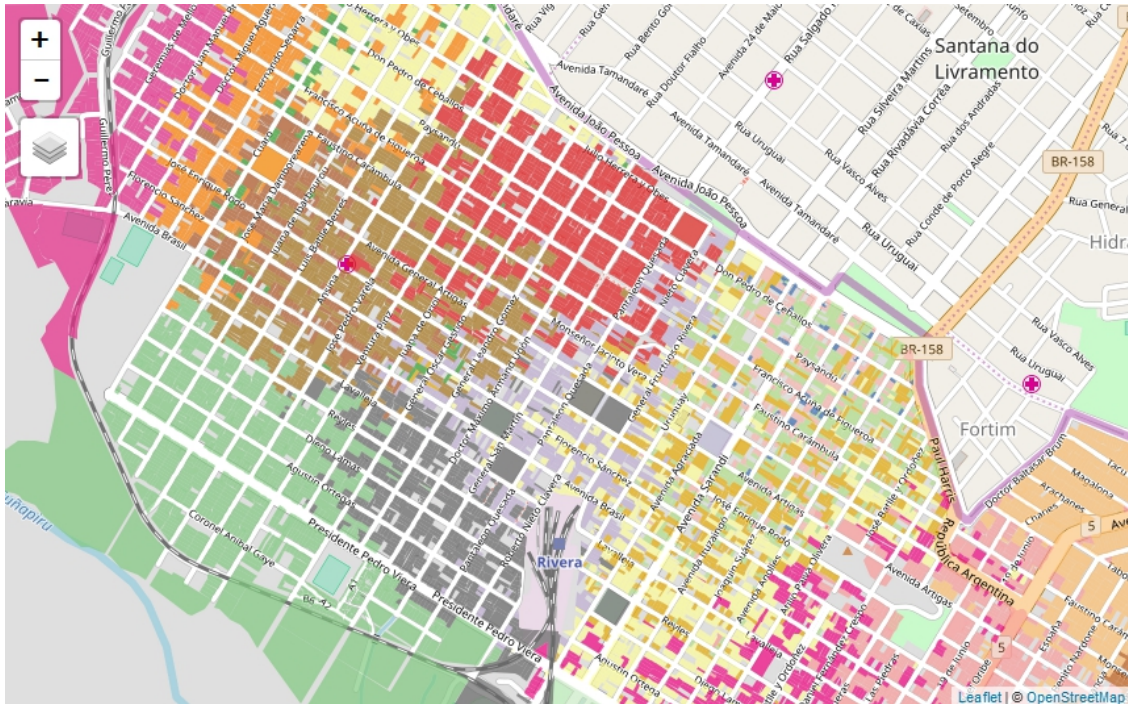


Figure 2: Ejemplo de tendencia espacial: Valores Catastrales del Centro de la ciudad de Rivera

Índice de Moran

Dado el numerador del Índice de Moran,

$$\mathbf{Y}^T (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n) \mathbf{C} (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n) \mathbf{Y}$$

Se tiene que los valores propios extremos de

$$(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n) \mathbf{C} (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)$$

determinan el rango del Índice de Moran.

Tiefelsdorf y Boots (1995) demuestran que los n vectores propios representan un caleidoscopio de patrones incorrelacionados, ortogonales entre sí, de posible autocorrelación espacial.

Teorema (Griffith, 2003)

El primer vector propio, \mathbf{e}_1^* es el conjunto de valores numéricos que tiene el mayor IM alcanzable por cualquier conjunto, para la configuración espacial definida por C .

El segundo vector propio \mathbf{e}_2^* es el conjunto de valores que tienen el mayor IM alcanzable para cualquier conjunto incorrelacionado con \mathbf{e}_1^* , y así sucesivamente.

Los primeros vectores propios se asocian con tendencias espaciales, mientras que los últimos se asocian a fenómenos de escala local.

Modelos Espaciales Autorregresivos

La especificación genérica de un modelo espacial autorregresivo asocia un rezago espacial con la variable endógena y y rezagos para cada variable exógena x .

$$\mathbf{y} = \rho_y \mathbf{V} \mathbf{y} + (\mathbf{I} - \rho_1 \mathbf{V}) \mathbf{x}_1 \beta_1 + \cdots + (\mathbf{I} - \rho_k \mathbf{V}) \mathbf{x}_k \beta_k + \boldsymbol{\varepsilon} \quad (1)$$

con $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ y $k + 1$ parámetros a estimar.

\mathbf{V} es la matriz de pesos espaciales. En el caso de que sea binaria coincide con la matriz \mathbf{C} .

Modelos Espaciales Autorregresivos

- a) Si $\rho \equiv \rho_1 = \dots = \rho_k$ se obtiene el modelo espacial Simultáneo Autorregresivo (SAR)

$$\mathbf{y} = \rho \mathbf{V} \mathbf{y} + (\mathbf{I} - \rho \mathbf{V}) \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2)$$

- b) Si $\rho \equiv \rho_y$ y $\rho_1 = \dots = \rho_k = 0$ se obtiene el modelo Espacial Rezagado (*Lag Model*).

$$\mathbf{y} = \rho \mathbf{V} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (3)$$

Especificación errónea del modelo

La perspectiva de una especificación errónea del modelo de regresión espacial asume que el modelo

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \quad (4)$$

tiene errores autocorrelacionados $\boldsymbol{\varepsilon}^*$, cuyos componentes son:

- ▶ Un ruido blanco, ε .
- ▶ Un conjunto de variables exógenas no especificadas (o no conocidas), \mathbf{E} , que conjuntamente presentan un patrón espacial con respecto a la estructura espacial subyacente \mathbf{V} .

El modelo con especificación errónea tiene la siguiente estructura:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \quad (5)$$

Especificación errónea y modelo SAR

Si el proceso espacial subyacente es estacionario, se puede expandir el término $(\mathbf{I} - \rho \mathbf{V})^{-1}$ de la ecuación (2) obteniendo:

$$(\mathbf{I} - \rho \mathbf{V})^{-1} = \sum_{k=0}^{\infty} \rho^k \mathbf{V}^k \quad (6)$$

Especificación errónea y modelo SAR

Operando,

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

$$\mathbf{y} = (\mathbf{I} - \rho \mathbf{V})^{-1} (\mathbf{X} \boldsymbol{\beta} - \rho \mathbf{V} \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon})$$

$$\mathbf{y} = \sum_{k=0}^{\infty} \rho^k \mathbf{V}^k (\mathbf{X} \boldsymbol{\beta} - \rho \mathbf{V} \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon})$$

$$\mathbf{y} = \sum_{k=0}^{\infty} \rho^k \mathbf{V}^k \mathbf{X} \boldsymbol{\beta} - \sum_{k=0}^{\infty} \rho^{k+1} \mathbf{V}^{k+1} \mathbf{X} \boldsymbol{\beta} + \sum_{k=0}^{\infty} \rho^k \mathbf{V}^k \boldsymbol{\varepsilon}$$

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \underbrace{\sum_{k=1}^{\infty} \rho^k \mathbf{V}^k \mathbf{X} \boldsymbol{\beta} - \sum_{k=1}^{\infty} \rho^{k+1} \mathbf{V}^{k+1} \mathbf{X} \boldsymbol{\beta}}_{=0} + \sum_{k=0}^{\infty} \rho^k \mathbf{V}^k \boldsymbol{\varepsilon}$$

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \underbrace{\sum_{k=1}^{\infty} \rho^k \mathbf{V}^k \boldsymbol{\varepsilon}}_{\text{misspecification term}} + \boldsymbol{\varepsilon}$$

Especificación errónea y modelo SAR

Si ε y las variables exógenas \mathbf{X} están incorrelacionadas, el estimador $\hat{\beta}_{MCO}$ es insesgado para el modelo

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$$

Pero las varianzas siguen siendo estimadas incorrectamente.

Especificación errónea en el modelo Espacial Rezagado

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

Especificación errónea en el modelo Espacial Rezagado

En este caso el término $\sum_{k=1}^{\infty} \rho^k \mathbf{V}^k (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$ se encuentra correlacionado con las variables exógenas \mathbf{X} , y bajo esta condición las estimaciones bajo *MCO* son sesgadas.

Filtrado Espacial

El objetivo del filtrado espacial es incorporar al modelo variables que puedan aproximar los términos mal especificados en los modelos autorregresivos (2) y (3).

Luego de introducir estas variables al modelo los residuos $\hat{\varepsilon}$ se transforman en Ruido Blanco, y así los modelos podrían llegar a estimarse utilizando *MCO*.

Especificación de las variables proxy utilizando vectores propios

Consideremos dos matrices de proyección

$$\mathbf{M}_{(\mathbf{1})} \equiv \mathbf{I} - \mathbf{1}(\mathbf{1}^T \mathbf{1})^{-1} \mathbf{1}^T \quad (7)$$

y

$$\mathbf{M}_{(\mathbf{X})} \equiv \mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (8)$$

Especificación de las variables proxy utilizando vectores propios

El conjunto de vectores propios $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}_{SAR}$ que se extrae de la siguiente forma cuadrática:

$$\{\mathbf{e}_1, \dots, \mathbf{e}_n\}_{SAR} \equiv \text{evec} \left[\mathbf{M}_{(\mathbf{X})} \frac{1}{2} (\mathbf{V} + \mathbf{V}^T) \mathbf{M}_{(\mathbf{X})} \right] \quad (9)$$

es, por diseño, ortogonal a la variable exógena \mathbf{X} .

Especificación de las variables proxy utilizando vectores propios

El conjunto de vectores propios $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}_{LAG}$ que se extrae de la siguiente forma cuadrática:

$$\{\mathbf{e}_1, \dots, \mathbf{e}_n\}_{LAG} \equiv \text{evec} \left[\mathbf{M}_{(1)} \frac{1}{2} (\mathbf{V} + \mathbf{V}^T) \mathbf{M}_{(1)} \right] \quad (10)$$

está potencialmente correlacionado con la variable exógena \mathbf{X} .

Estos dos conjuntos de vectores propios establecen la base para las variables proxy espaciales.

Variables proxy en el Modelo SAR

Sea \mathbf{E}_{SAR} una matriz cuyos vectores son un subconjunto de $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}_{\text{SAR}}$. Una combinación lineal de este subconjunto aproximará el término mal especificado en el modelo SAR.

$$\mathbf{E}_{\text{SAR}}\gamma \approx \sum_{k=1}^n \rho^k \mathbf{V}^k \varepsilon$$

El término $\mathbf{E}_{\text{SAR}}\gamma$ es ortogonal a las variables exógenas \mathbf{X} y a los residuos $\hat{\varepsilon}$, por lo que el estimador $\hat{\beta}_{MCO}$ es insesgado y se reduce la varianza de las estimaciones respecto al modelo erróneamente especificado.

$$\mathbf{y} = \mathbf{X}\hat{\beta} + \mathbf{E}_{\text{SAR}}\hat{\gamma} + \hat{\varepsilon}$$

Variables proxy en el Modelo Lag

Sea \mathbf{E}_{Lag} una matriz cuyos vectores son un subconjunto de $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}_{Lag}$. La aproximación al término mal especificado es

$$\mathbf{E}_{\text{Lag}}\boldsymbol{\gamma} \approx \sum_{k=1}^n \rho^k \mathbf{V}^k \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

En este caso el término $\mathbf{E}_{\text{Lag}}\boldsymbol{\gamma}$ se encuentra correlacionado con las variables \mathbf{X} y su incorporación en el modelo corrige el sesgo del estimador $\hat{\beta}_{MCO}$, así como también se reducen las varianzas. Sin embargo existe correlación entre el componente de tendencia sistemática $\mathbf{X}\hat{\boldsymbol{\beta}}$ y la señal estocástica $\mathbf{E}_{Lag}\hat{\boldsymbol{\gamma}}$.

- ▶ El conjunto $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}_{Lag}$ se calcula independientemente de las variables \mathbf{X} . Únicamente depende de la matriz \mathbf{V} .
- ▶ El conjunto $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}_{SAR}$ depende de las variables exógenas. Cualquier cambio en la estructura subyacente del modelo requiere recalcular el conjunto de vectores propios.

¿Cómo se identifican los subconjuntos de vectores \mathbf{E}_{SAR} y \mathbf{E}_{Lag} ?

Los conjuntos \mathbf{E}_{SAR} y \mathbf{E}_{Lag} deben cumplir dos condiciones:

- ▶ Los residuos del modelo filtrado tienen que ser un Ruido Blanco.
- ▶ Tienen que ser el menor conjunto de vectores posible (parsimonia).

Estrategias de búsqueda

Griffith (2003) sugiere utilizar procedimientos *stepwise* sobre el conjunto \mathbf{E} completo o un subconjunto preseleccionado de vectores \mathbf{E}^c :

Correlación Positiva

Si los residuos ε^* presentan correlación positiva, \mathbf{E}^c es el conjunto de vectores propios tal que

$$\lambda_1 \geq \lambda_i \geq (\alpha)\lambda_1 \text{ con } 0 < \alpha \leq 1$$

El parámetro α controla la amplitud de la búsqueda. Se sugiere $\alpha \approx 0.5$, incluyendo sólo los vectores con correlación moderada o alta.

Estrategias de búsqueda

En los procedimientos *stepwise* en una regresión estándar, el objetivo es minimizar el término $\boldsymbol{\varepsilon}^{*T} \boldsymbol{\varepsilon}^*$. Este criterio ignora la información espacial en los residuos $\boldsymbol{\varepsilon}^*$.

Tiefelsdorf y Griffith (2007) proponen que la función objetivo sea el Índice de Moran estandarizado de los residuos.

$$z [IM (\boldsymbol{\varepsilon}^*)] \equiv \frac{\left\{ \frac{\mathbf{y}^T \mathbf{M}_{\mathbf{X}|\mathbf{E}} \mathbf{V} \mathbf{M}_{\mathbf{X}|\mathbf{E}}}{\mathbf{y}^T \mathbf{M}_{\mathbf{X}|\mathbf{E}} \mathbf{y}} - E [IM (\boldsymbol{\varepsilon}^*)] \right\}}{\{var [IM (\boldsymbol{\varepsilon}^*)]\}^{1/2}} \quad (11)$$

Estrategias de búsqueda

Paso 1: Se obtiene el p - valor para la prueba:

$$\begin{aligned} H_0) IM &= 0 \\ H_1) IM &> 0 \end{aligned} \tag{12}$$

realizada para los residuos ϵ^* del modelo inicial.

Si p - valor $> \alpha$ (predeterminado) se detiene el algoritmo.

Si p - valor $< \alpha$ (predeterminado) va al Paso 2.

Estrategias de búsqueda

Paso 2: Se elige el vector propio que al agregarlo en el modelo maximice el p - valor de la prueba (12).

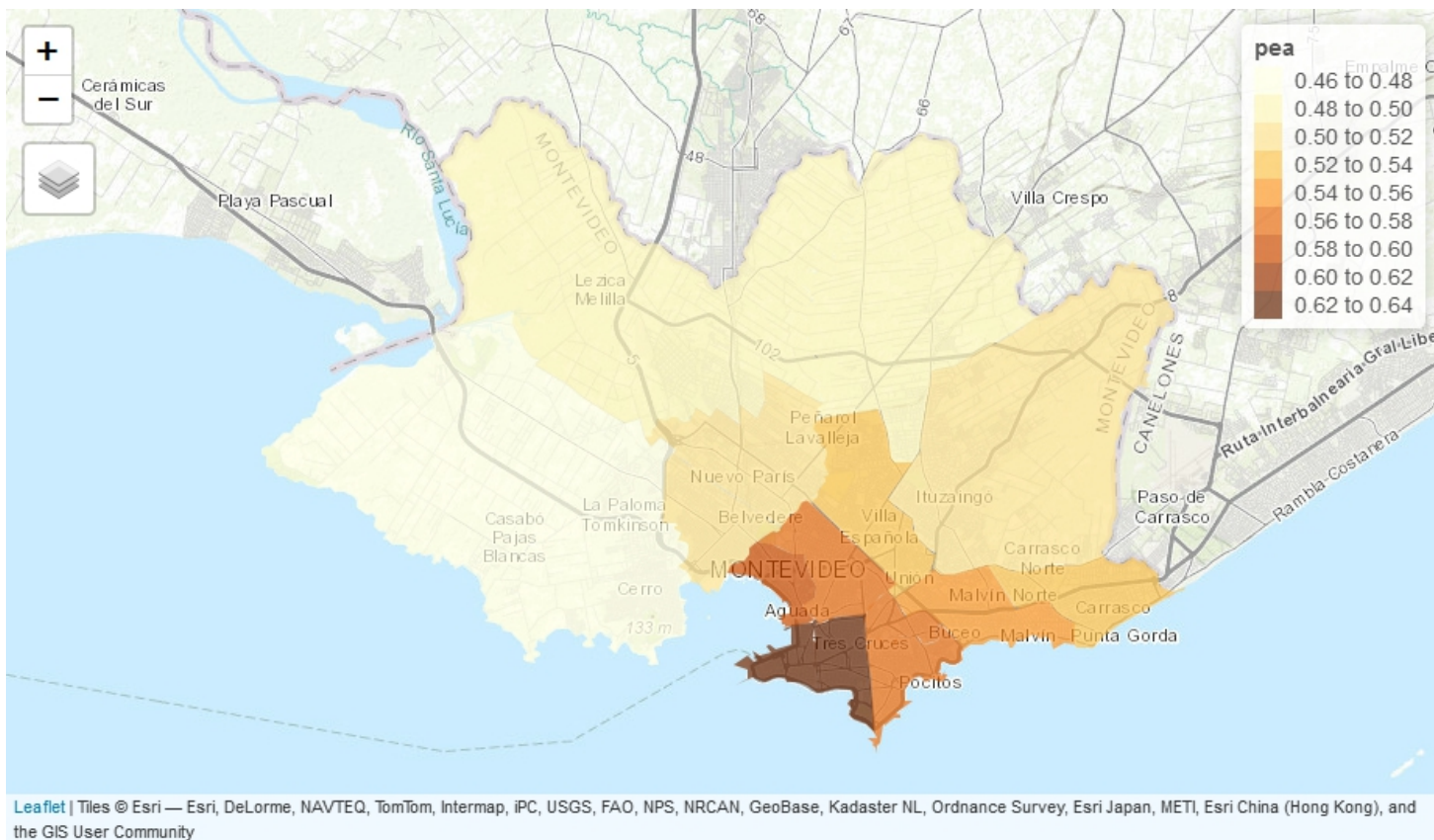
Si el p- valor $> \alpha$ se detiene el algoritmo.

Si p - valor $< \alpha$ se quita el vector seleccionado del conjunto inicial de vectores, y se va al Paso 3.

Paso 3: Se repite el Paso 2 hasta que p- valor $> \alpha$.

Ejemplo: Modelización de la PEA por CCZ

Se quiere ajustar un modelo espacial que explique la PEA, tomando como unidad geográfica los CCZ de Montevideo.



Ejemplo: Modelización de la PEA por CCZ

Se utiliza como variable explicativa al porcentaje de población con nivel educativo universitario o superior (*Edu*). Las estimaciones para el modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$ son:

Table 1: Resultados del ajuste del Modelo de Regresión Lineal

Variable	Parámetro	Estimación	p-valor	Significación
Intercept	β_0	0.48804	2e-16	***
Edu	β_1	0.27509	2.67e-05	***
IM Residuos		0.27414	0.009901	

Ejemplo 1: Modelización de la PEA por CCZ

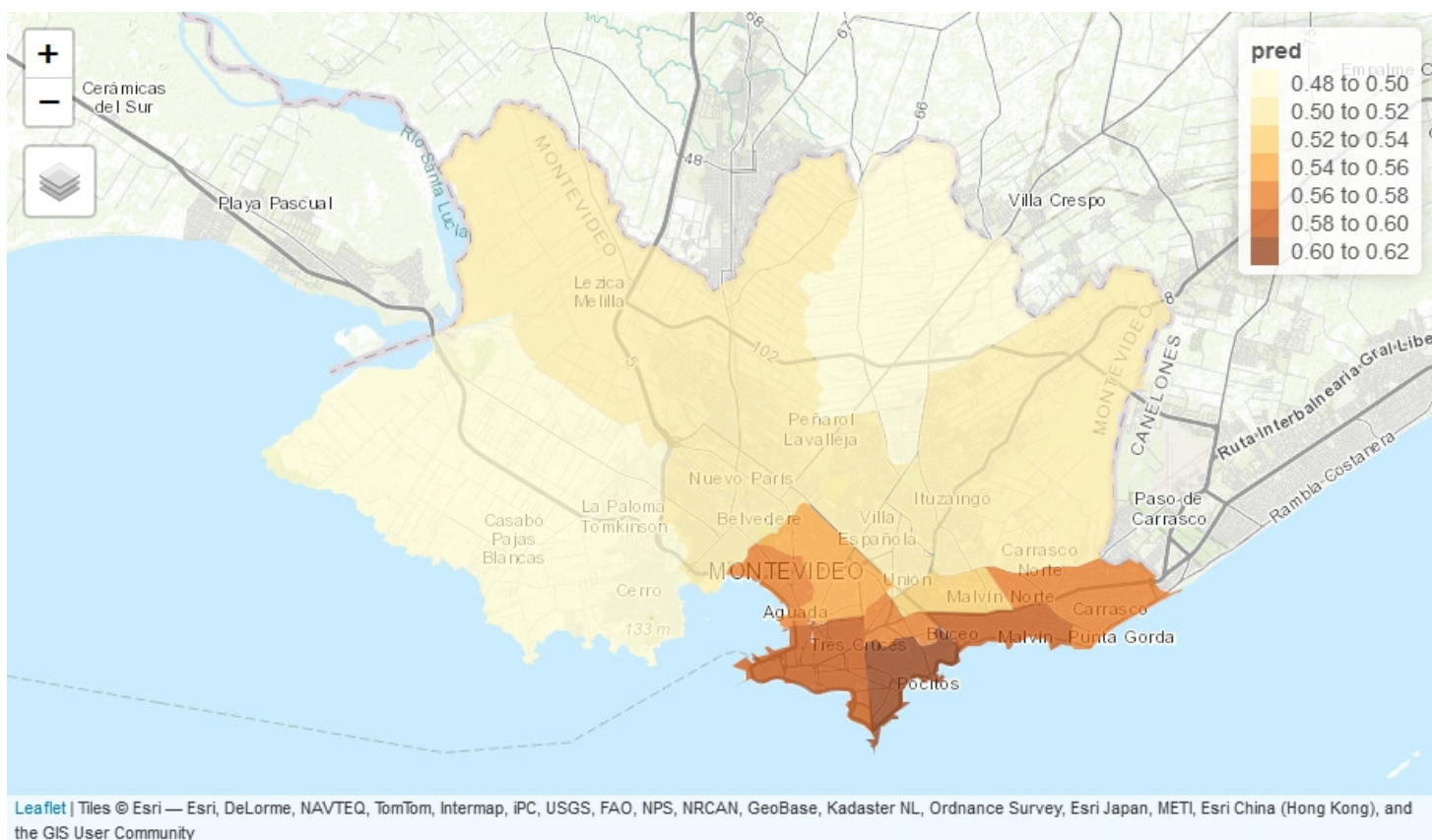


Figure 4: Predicción ($X\hat{\beta}$) del modelo de regresión lineal

Ejemplo 1: Modelización de la PEA por CCZ

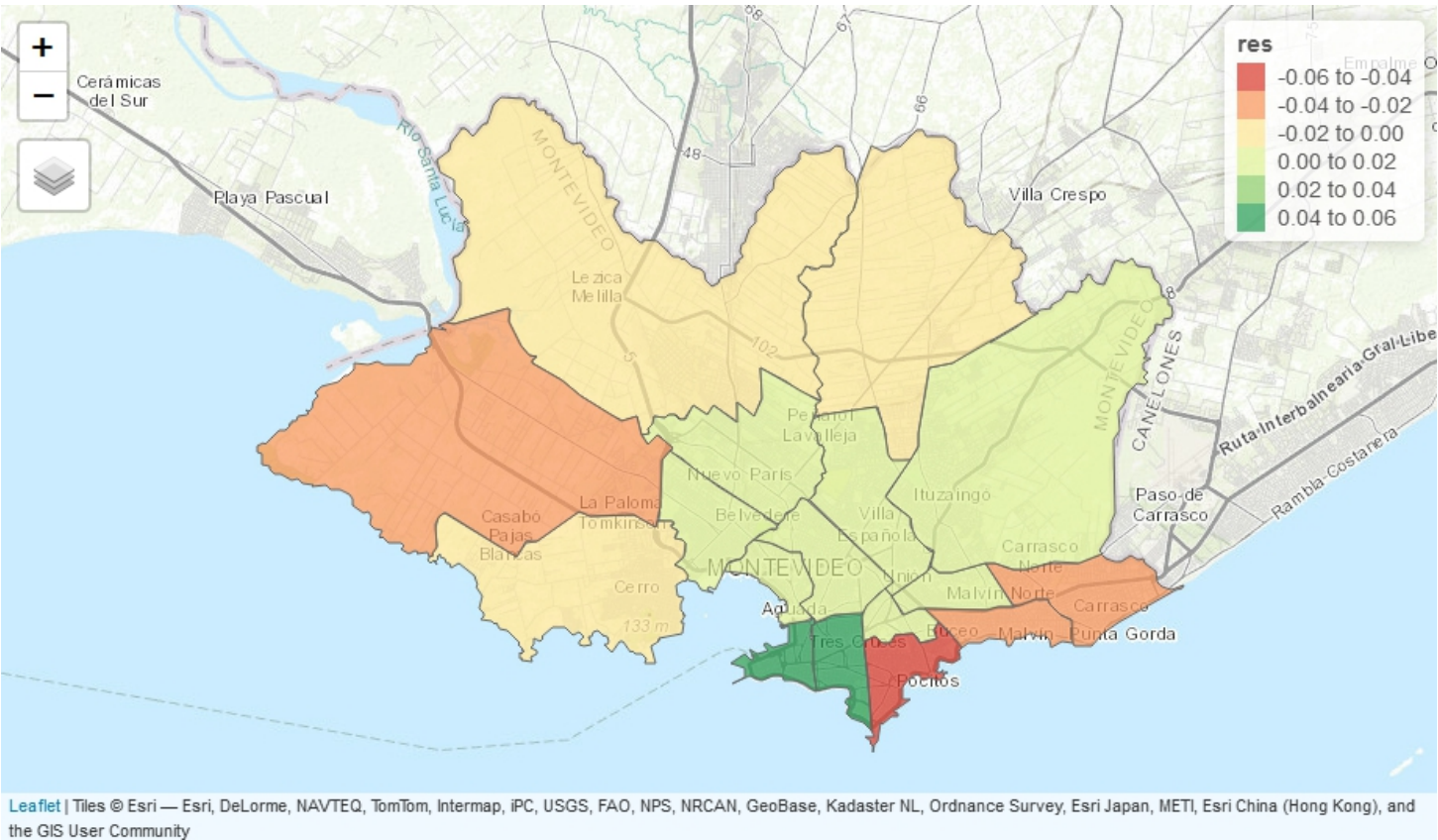
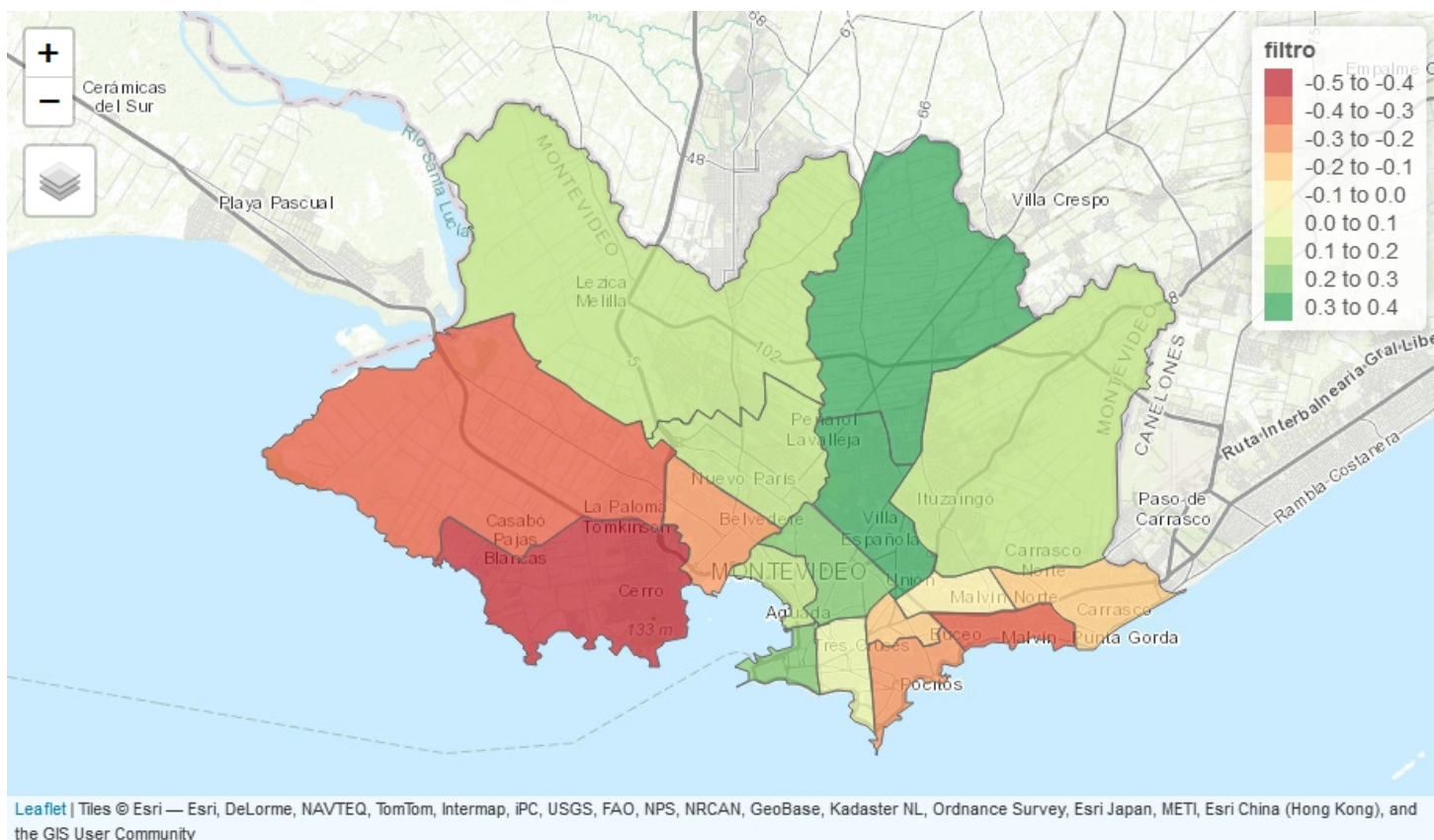


Figure 5: Residuos del modelo de regresión lineal

Ejemplo 1: Modelización de la PEA por CCZ

Del Filtrado Espacial surge un vector propio (el tercero) para introducir al modelo.



Ejemplo 1: Modelización de la PEA por CCZ

Table 2: Resultados del ajuste del Modelo de Regresión Lineal con vectores resultantes del Filtrado Espacial

Variable	Parámetro	Estimación	p-valor	Significación
Intercept	β_0	0.484826	2e-16	***
Edu	β_1	0.29193	1.38e-06	***
Vec3	γ_1	0.066861	0.00585	**
IM Residuos		0.079037	0.2574	

Ejemplo 1: Modelización de la PEA por CCZ

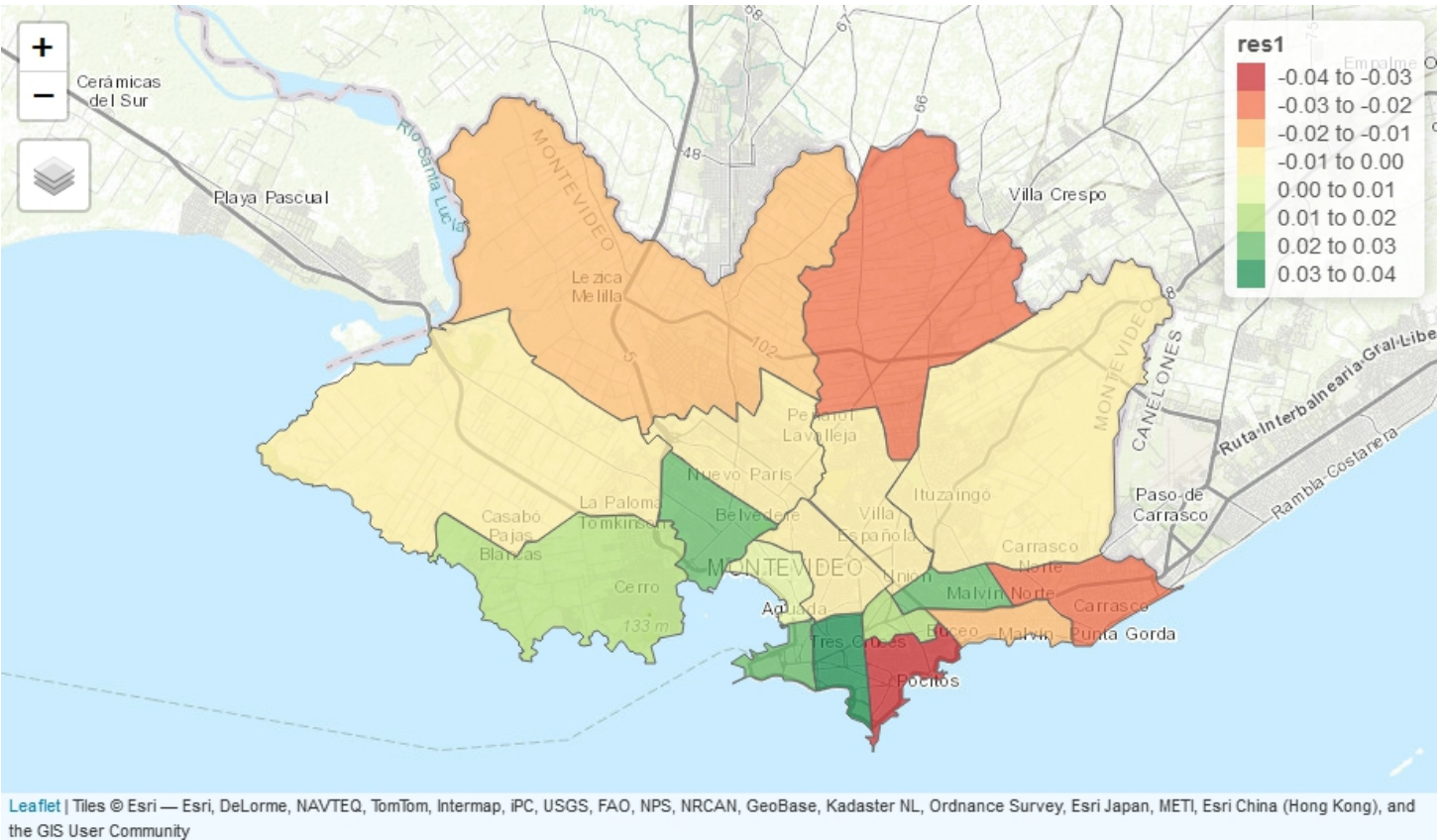


Figure 7: Residuos luego del Filtrado Espacial

Ejemplo 2: Filtrado Espacial aplicado a flujos

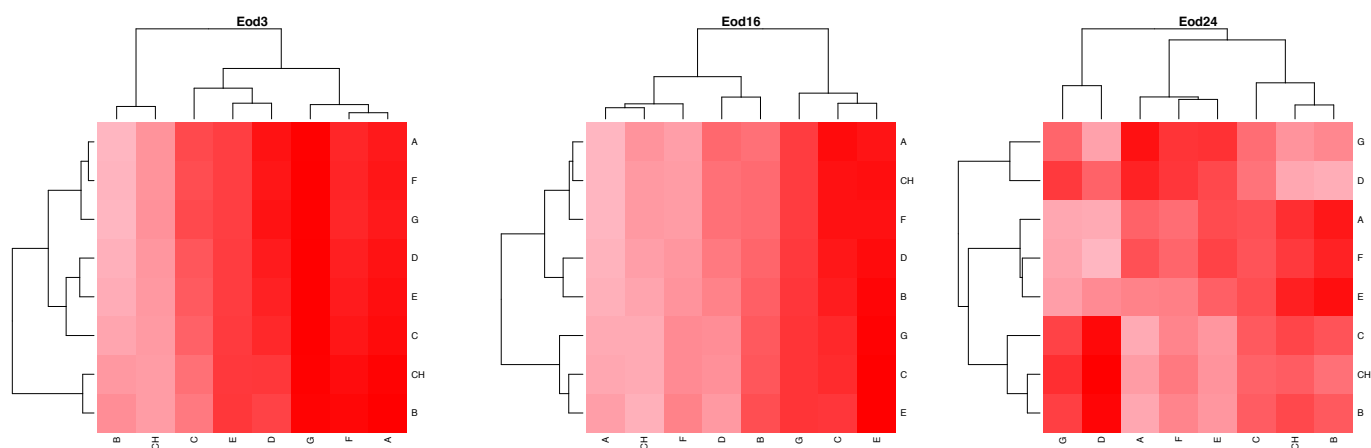
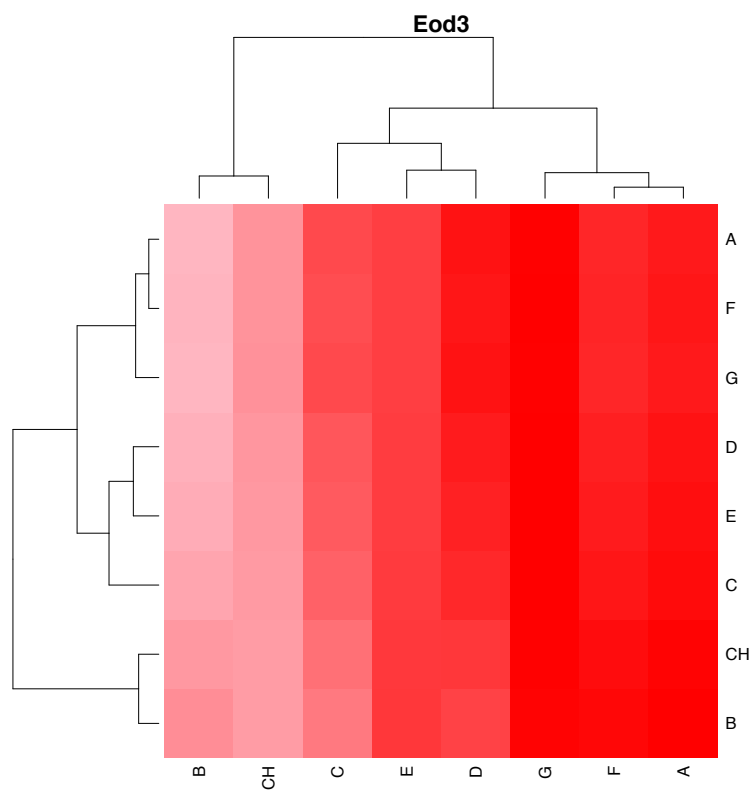
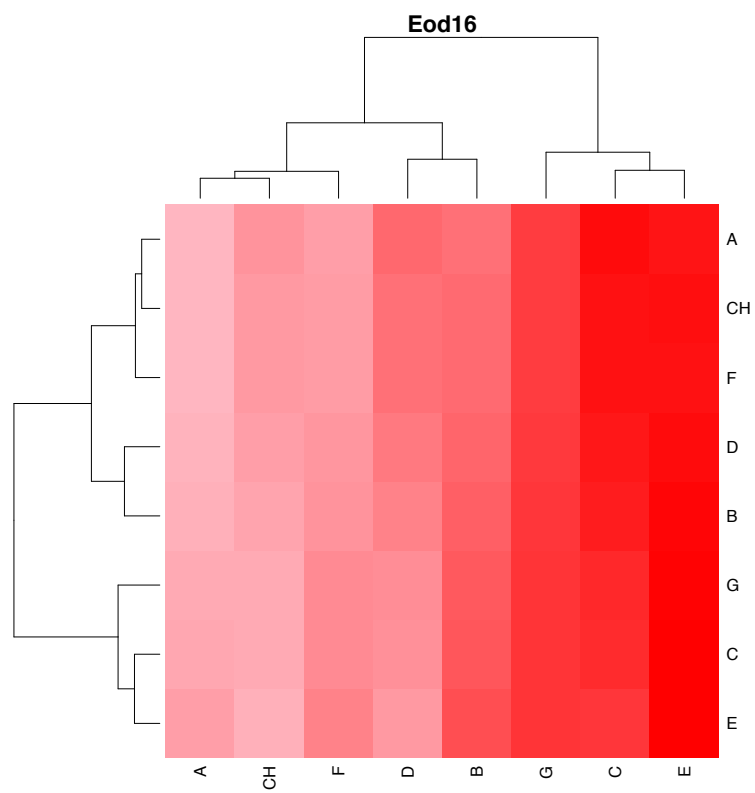


Figure 8: Heatmaps para los vectores propios del Filtrado Espacial

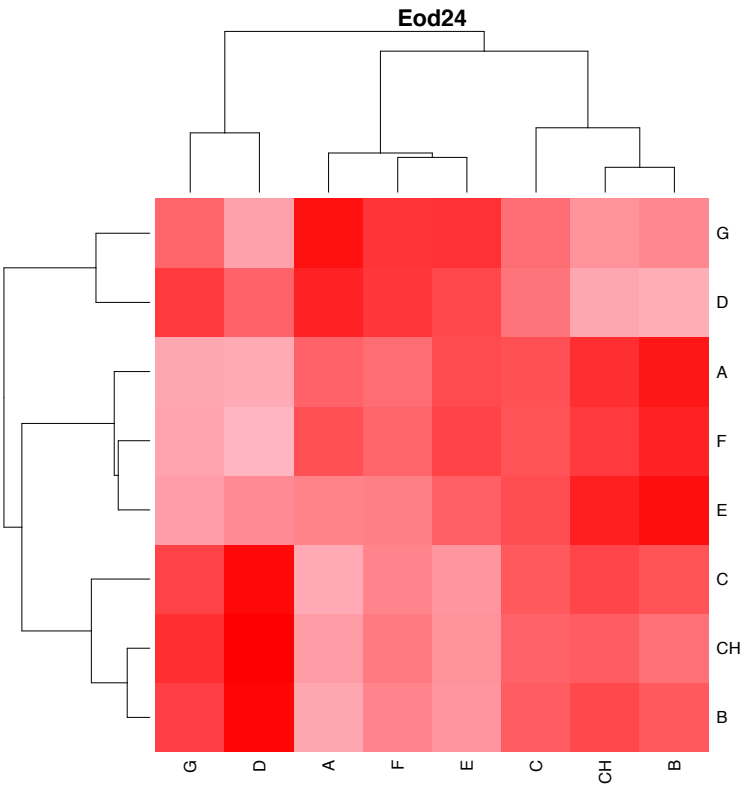
Ejemplo 2: Filtrado Espacial aplicado a flujos






Ejemplo 2: Filtrado Espacial aplicado a flujos



Ejemplo 2: Filtrado Espacial aplicado a flujos



Bibliografía

-  Griffith, D. A. (2003) *Spatial Autocorrelation and Spatial Filtering*, Advances in Spatial Science, Springer - Verlag.
-  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.