

# Regresión con Dependencia Espacial y GeoDa

Joshua Kunst

`jbkunst@gmail.com`

12 de Diciembre, 2007

Muy conocido entre estadísticos y no estadísticos es el modelo de regresión lineal, en donde se trata de explicar una variable de interés  $y$ , la cual es difícil o costosa de medir, a través de una combinación lineal de otras variables que están asociadas. El modelo matricialmente tiene la siguiente forma:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon.$$

Donde  $\epsilon$  es un vector aleatorio distribuido normalmente con media el vector de ceros y matriz de varianzas covarianzas la identidad de orden igual al tamaño muestral multiplicado por un escalar. Un supuesto implícito que proviene de lo anterior es que cada observación se asume independiente del resto debido a que la no correlación en el caso normal implica independencia entre las componentes.

¿Qué ocurre cuando no se se puede suponer este supuesto de independencia? Por ejemplo, en el caso de las Series de Tiempo un valor es dependiente de los anteriores. Remitiendonos al caso sencillo del modelo Autoregresivo de primer orden:

$$y_n = \phi y_{n-1} + \epsilon, \quad i = 1, 2, \dots, n.$$

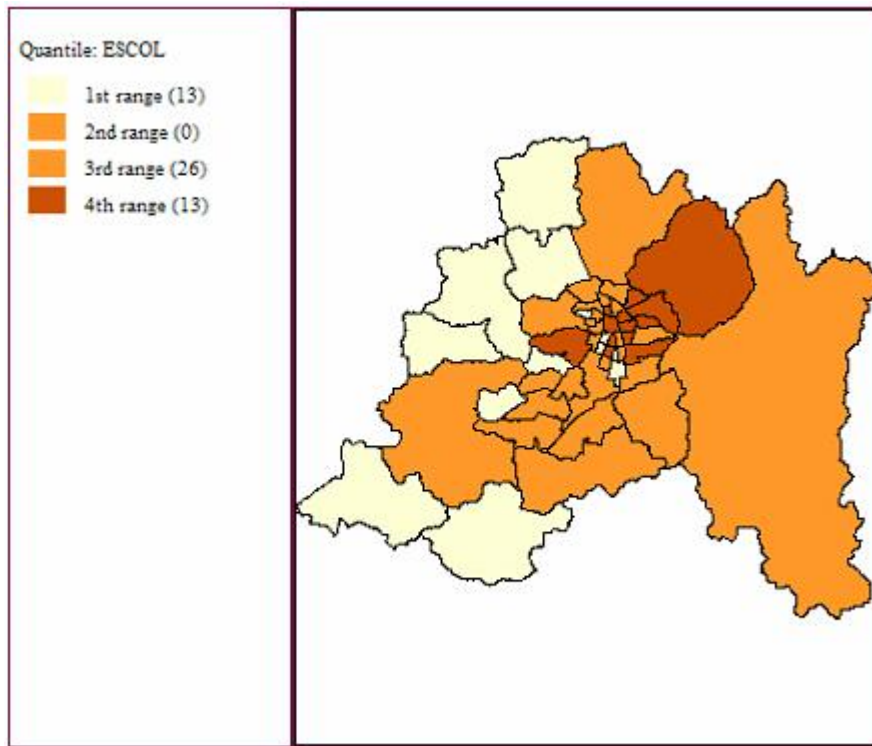
Acá  $\epsilon$  suele ser un ruido gaussiano y claramente los valores no son independientes. El modelo anteriormente señalado se puede reescribir matricialmente de la siguiente manera:

$$\mathbf{y} = \phi \mathbf{W} \mathbf{y} + \epsilon.$$

Donde la matriz  $W$  es cuadrada de orden  $n$  y posee ceros a excepción de los elementos  $w_{i,(i+1)}$  donde existen unos.

Notamos que la estructura de dependencia del modelo está contenido en la forma de la matriz  $\mathbf{W}$ , luego podemos modificarla, siempre de manera coherente, para que la dependencia de las observaciones sea distinta.  $\mathbf{W}$  asocia que valores dependen de otros.

Como acabamos de ver, en una Serie de Tiempo la dependencia es secuencial y es fácil expresar la matriz  $\mathbf{W}$ . Cuando hablamos de observaciones espaciales, por ejemplo, comunas las cuales están contiguas unas con otras, la intuición nos dice que debiera haber dependencia entre estas observaciones. En el siguiente dibujo vemos los percentiles (divididos por 4 rangos) del *Nivel de Escolaridad* por comunas de la Región Metropolitana de Santiago, Chile:



Como vemos los percentiles se agrupan claramente por zonas y por tanto un supuesto de independencia se cae al suelo. Existe *Dependencia Espacial*, dependencia entre unidades muestrales contiguas o cercanas en esta variable y por tanto un modelo de regresión común, para explicar tal variable, no sería apropiado.

El problema que surge es como integrar tal dependencia al modelo. Es necesario integrar un(os) parámetro(s), decidir que regiones dependen de que regiones, y en palabras técnicas: la forma debiese tener la matriz  $\mathbf{W}$ . Lo que se hace es agregar un parámetro que pondere la matriz  $\mathbf{W}$  (de la misma forma que veíamos en el modelo autoregresivo). Para crear la matriz  $\mathbf{W}$ , que en la jerga geográfica-estadística se le denomina *matriz de pesos* (de ahí la "w") *de contactos espaciales*, existen dos criterios:

de *contigüidad* (criterios de vecindad) y de *distancia*, y para cada uno de ellos existen tipos para definir la dependencia y por tanto la construcción de la matriz  $\mathbf{W}$ .

Ya establecido el criterio de dependencia y cercanía se pueden ajustar dos posibles modelos, a saber: el modelo con *Dependencia Espacial Sustantiva* y el modelo con *Dependencia Residual*, los cuales tiene una estrecha relación en la forma con los modelos autoregresivos y de medias móviles. En particular si el modelo a ajustar es el de Dependencia Espacial Sustantiva el modelo es de la forma:

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\beta + \epsilon.$$

Donde se supone:

$$\epsilon \sim N(0, \sigma^2 I).$$

Ajustando el modelo y teniendo estimaciones lo importante es testear la significancia de los parámetros y en este caso es muy importante docimar la significancia del parámetro  $\rho$ , pues en ello se basa el modelo de dependencia.

Todo lo anterior: especificación del modelo, construcción de la matriz de pesos de contactos espaciales y dójimas viene implementado en un *software* llamando *GeoDa*, el cual es desarrollado por el doctor Luc Anselin's del departamento de Geografía de la Universidad de Illinois. El programa es bien fácil de usar, la mayoría a un click y se puede descargar gratuitamente desde el link <https://www.geoda.uiuc.edu/downloadin.php>.