Instituto Tecnológico Autónomo de México

Maestría en Ciencia de Datos - Reporte de estancia

Miguel Ángel Castañeda Martínez

Tabla de contenidos

1 Contexto del problema

Dentro del mundo de problemas que es posible resolver con herramientas de ciencia de datos, existen algunos que destacan por el tipo de estructura inherente en los datos que se tienen para su resolución. Uno de ellos es el problema que involucra **estructura geoespacial**, mismo que será el tema central de este trabajo.

La forma más general de definir un conjunto de datos geoespaciales es todo aquel que tenga un componente geográfico, que por lo general son el par de coordenadas **longitud** y **latitud**. Estos pares de coordenadas tiene su abstracción matemática en vectores cuya unidad fundamental es el **punto**, un par de puntos definen una **polilínea**, y a su vez una serie de polilíneas definen un **polígono**.

Tomando esto en cuenta, existe una gran variedad de aplicaciones alineadas a explicar fenómenos geoespaciales como pudieran ser:

- Visualizar la distribución de algún índice de pobreza en una región dada.
- Predecir la producción de algún producto agrícola en función de su ubicación.
- Optimizar la ruta más corta y de menor tiempo de traslado dado un origen y un destino.

Los primeros dos ejemplos son de particular interés para este trabajo debido a que un enfoque de solución de estos es utilizando regresión lineal, misma a la que se le puede imponer estructura geoespacial como una componente adicional de la regresión.

Sin embargo, dado el costo computacional que esto implica, una práctica que se da en estos problemas es aproximar el componente espacial con funciones indicadoras de la región a la que pertenece un registro dado, lo cual no es propiamente un enfoque incorrecto pero sacrifica información del problema.

Adicionalmente, una de las implicaciones de utilizar métodos de regresión para la resolución de estos problemas es que a la hora de comparar modelos se usan métricas típicas de regresión, como lo son R^2, RMSE, entre otros, a la par que se dejan de lado métodos para diagnosticar el desempeño de modelos geoespaciales que serán explicados con mayor detalle más adelante.

Habiendo explicado lo anterior, el objetivo de este trabajo es detallar estos métodos de diagnóstico para comparar técnicas de regresión simplificadas con modelos que propiamente incorporan estructuras geoespaciales al problema.

2 Estrategia de solución

En términos generales, la comparación de modelos geoespaciales se sustenta en encontrar **aleatoriedad espacial** en los residuales del modelo, que también es conocidad como *Spatial Randomness* (SR), y es la ausencia de cualquier tipo de patrón espacial. Para que esta exista se deben satisfacer dos condiciones:

- Los residuales son equiprobables en cualquier ubicación en el espacio.
- El valor de un residual no depende del valor de sus vecinos.

Consecuentemente, la ubicación de las observaciones puede ser alterada sin afectar la información contenida en los datos bajo SR, esto se puede hacer mediante procesos de permutación de los datos, y el probarlo está apalancado del concepto de **autocorrelación espacial**, que busca medir la variación de una misma variable en ubicaciones distintas.

Esto lleva al planteamiento de la prueba de hipótesis fundamental para cualquier problema con estructura geoespacial:

- H_0: Los datos se distribuyen bajo SR.
- H_1: Los datos presentan estructura espacial.

Finalmente, si un modelo geoespacial es lo suficientemente bueno como para extraer esta estructura de los datos, entonces hacer una prueba de hipótesis espacial en sus residuales debería aceptar la hipótesis nula.

3 Modelos a comparar

Para compararlo, se utilizará como referencia el trabajo de investigación publicado en el artículo *The Assessment of Impacts and Risks of Climate Change on Agriculture (AIRCCA)*, mismo que plantea un modelo de regresión lineal para predecir la producción mundial por región de tres semillas fundamentales: maíz, arroz, y trigo.

Los modelos utilizados en dicha publicación hacen la simplificación descrita en la sección anterior, es decir, agrupa diversos puntos de coordenadas en una región dada para la cual se crea una variable indicadora que toma los valores:

- $I_R = 1$, si las coordenadas del registro pertenecen a esa región.
- $I_R = 0$, si no pertenecen a esa región.

Esto se hace para tantas regiones existan en la definición del problema, y los modelos que se van a comparar serán entonces:

- Regresión lineal de referencia planteada en el artículo.
- Regresión lineal con componente geoespacial.
- Regresión lineal con componente geoespacial y efectos fijos.
- Regresión con pesos geográficos, o *Geographically Weighted Regression* (GWR)
