Escuela de Ciencia y Tecnología - UNSAM
Licenciatura en Ciencia de Datos

ESTIMACIÓN DE PRECIOS DEPARTAMENTOS Y PRECIO POR METRO CUADRADO EN EL AMBA MEDIANTE REGRESSION KRIGING

Materia: Análisis de Datos Espaciales

Profesor: Rafael Grimson

Integrantes: Nazareno Magallanes, Julián Rolando

Año: 2024

Objetivos:

- Entrenar dos modelos utilizando Regression Kriging, uno para la estimación del precio de la propiedad y otro para la del precio por metro cuadrado.
- Investigar cómo varían el precio de un departamento y del metro cuadrado dependiendo de la ubicación geográfica dentro del AMBA.
- Determinar la influencia de las covariables en estos precios.

Herramientas:

- Python como lenguaje de programación para el procesamiento de datos y el desarrollo de modelos.
- NumPy y Pandas para el manejo de datos.
- Matplotlib y Seaborn para la visualización de datos.
- PyKrige y scikit-learn para implementar el modelo de regresión kriging.

Introducción:

En el Área Metropolitana de Buenos Aires (AMBA) hay mucha variabilidad en el precio de las propiedades. La estimación de estos precios es de suma importancia para el mercado inmobiliario. Para poder calcularlos proponemos utilizar el método de interpolación espacial llamado Regression Kriging para su modelado. Este método tomará en cuenta tanto la tendencia espacial como las siguientes variables predictoras o covariables: número de habitaciones, número de dormitorios, número de baños, superficie cubierta y total, distancia a estaciones de subtes, a paradas de colectivo, a estaciones de trenes y cercanía a espacios verdes.

Este análisis tiene como propósito entender de manera más profunda la distribución espacial de los precios y la manera en que los distintos factores inciden en estos. Al utilizar regression kriging, un enfoque geoespacial sofisticado, no solo se consigue una estimación fundamentada en la separación entre los puntos de interés, sino que también se incluyen otras variables predictivas, lo que podría contribuir a una mayor exactitud y utilidad del modelo en la valoración de precios en el sector de bienes raíces.

Regresión Kriging:
Es un método geoestadístico que combina un modelo lineal con ordinary kriging aplicado a los residuos.

La variable que queremos modelar se descompone en dos partes:

$$Z(s) = \mu(s) + \varepsilon(s)$$

donde $\mu(s)$ es la parte determinística o "deriva" que se asume como una función de la ubicación y de covariables adicionales. Se puede expresar como:

$$\mu(s) = \beta_0 + \sum_{i=1}^{p} \beta_i \cdot X_i(s)$$

Y donde $\varepsilon(s)$ corresponde a la parte estocástica, la cual modela la tendencia espacial. Aquí es donde el semivariograma obtiene importancia. Si los residuos muestran dependencia espacial (es decir, los valores en ubicaciones cercanas son más similares que los valores en ubicaciones lejanas), esta correlación puede ser modelada mediante un semivariograma.

El semivariograma mide cómo varían los residuos en función de la distancia entre puntos espaciales. Se calcula como:

$$\gamma(h) = \frac{1}{2N(h)} \cdot \sum_{i=1}^{Nh} [Z(x_i + h) - Z(x_i)]^2$$

Donde:

h: Distancia entre dos ubicaciones.

N(h): Número de pares de puntos separados por la distancia h.

Una vez calculado el semivariograma empírico, se ajusta un modelo teórico (por ejemplo, esférico, exponencial o gaussiano) que representa la correlación espacial de los residuos.

El semivariograma afecta directamente la matriz de covarianza espacial (C) que se utiliza en la estimación Kriging. Esta matriz define los pesos asignados a cada punto de datos para predecir los residuos en una nueva ubicación.

- Si dos puntos están **cercanos** y el semivariograma indica una fuerte correlación, se asignan pesos más altos a esos puntos.
- Si los puntos están **lejanos** o fuera del rango, el semivariograma indica que no están correlacionados, y sus pesos disminuyen.

La fórmula general para la predicción es:

$$\hat{\varepsilon}(s) = \sum_{i=1}^{N} \lambda_i \cdot \varepsilon(s_i)$$

Donde:

- λi: Pesos calculados mediante el sistema Kriging usando el semivariograma.
- ε(si): Residuos conocidos en las ubicaciones muestreadas.

Para el valor estimado de Z(s) en un nuevo punto:

$$\hat{Z}(s) = \hat{\mu}(s) + \hat{\varepsilon}(s)$$

Procedimiento:

1. Obtención de los datos:
   Los datos para el proyecto se consiguieron de las siguientes maneras:
   - Precio de los departamentos en venta con las covariables - material del curso, fuente: web de Properati.
   - Dataset de estaciones de Subte - extraído de BA Data.
   - Dataset de paradas de colectivo - A través de QGIS con OpenStreetMap.
   - Dataset de estaciones de trenes - extraido de Datos Argentina
   - Espacios Verdes: A través de QGIS con OpenStreetMap.
   - Polígonos de AMBA: extraídos de Datos abiertos PBA

2. Preparación de los datos:
   Con QGIS, convertimos todos los csv a Spatialite, para poder manejar SQL espacial.
   Usando QGIS importamos las capas como tablas para manejar la base de datos.

Con esto pudimos hacer intersección de los datos con el amba con *intersects on* y con la función *Distance* se calcularon las distancias a las bocas de subte, espacios verdes públicos, paradas de colectivos, y estaciones de tren. Para ésto se creó un archivo alternativo con menos cantidad de datos.

3. Análisis exploratorio de los datos:
    - Descripción del conjunto de datos, tipos de variables.
    - Histogramas y boxplots para analizar la distribución de las variables.
    - Diagramas de dispersión para estudiar relaciones entre variables, como precio vs. superficie.
    - Identificación de valores ausentes (NaN).
    - Casi siempre eliminación de registros para valores faltantes o outliers.
    - Las diferencias en distancias lejanas tienen un impacto marginal en el precio, mientras que las diferencias en distancias cercanas son más significativas. Por eso se decidió truncar las distancias a un valor máximo, establecido en **2000 metros (2 km)**

4. Modelado:
    - Selección de atributos: Se entrenó un modelo de regresión lineal con regularización Lasso usando el conjunto de datos reducido con el objetivo de ver la importancia tentativa de cada una de las variables predictoras.

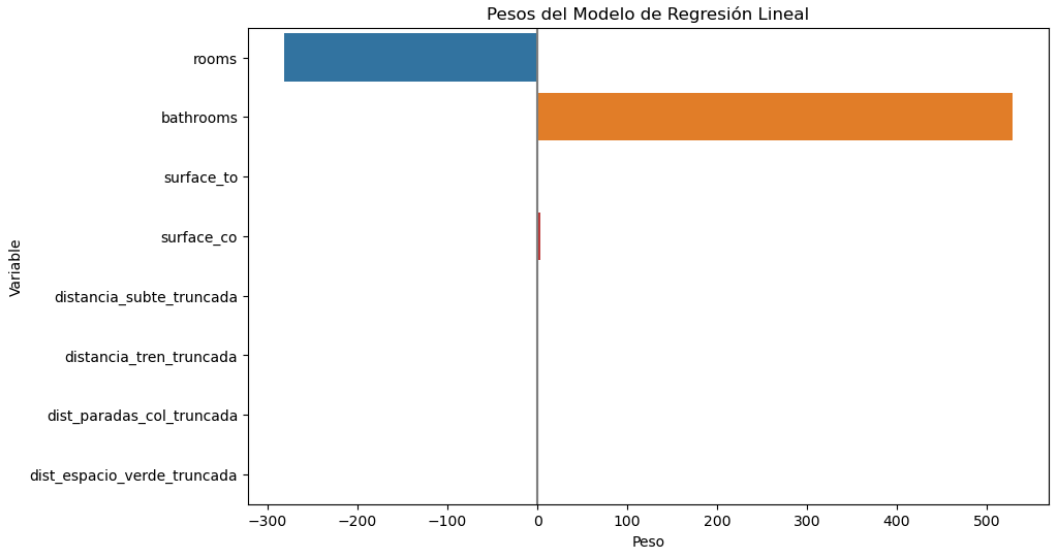

Fig 1

Se determinó descartar bedrooms, superficies y distancias. Éstas variables podrían haber sido descartadas desde el comienzo, la variable objetivo claramente no debería depender de las superficies de los departamentos, y la cantidad de dormitorios está correlacionada con la cantidad de ambientes.

- Selección del modelo:
  Para la selección del modelo de regresión kriging se utilizó la estrategia de búsqueda por grilla de hiper parámetros.
  Como modelo de regresión para la parte determinística se utilizó el método de regularización *elastic net*.

- Entrenamiento del modelo:
  Se dividió el conjunto de datos en dos partes (conjuntos de entrenamiento y de prueba). El modelo seleccionado anteriormente fue entrenado usando el conjunto correspondiente. Se hicieron predicciones sobre el conjunto de prueba para reportar las métricas de rendimiento o error.

- Visualización de resultados:
  Se creó una grilla que cubra la superficie del AMBA y con casillas de 200 x 200 metros usando la herramienta “Crear cuadrícula” de QGIS.
  Se utilizó el modelo obtenido anteriormente para predecir el precio por metro cuadrado de monoambientes en el centro de cada casilla.
  El resultado fue cargado nuevamente a QGIS para hacer una presentación final de los resultados.

Resultados:

- Descripción del Modelo:
  La componente determinística del modelo es una regresión lineal con regularización elastic net con parámetros:
  - alpha = 10
  - l1_ratio = 0,1

  Los pesos de las variables del modelo son:
  - ambientes ≅ 3,12
  - baños ≅ 23,94

  La componente espacial es modelada por un Ordinary Kriging con parámetros:
  - variogram_model = gaussian

- weight = False
- n_closest_points = 5

El modelo gaussiano ajustado al variograma tiene las siguientes propiedades:

- nugget ≅ 279746
- sill ≅ 1150762,52
- partial_sill ≅ 871015,53
- range ≅ 0,00025

- Semivariograma:

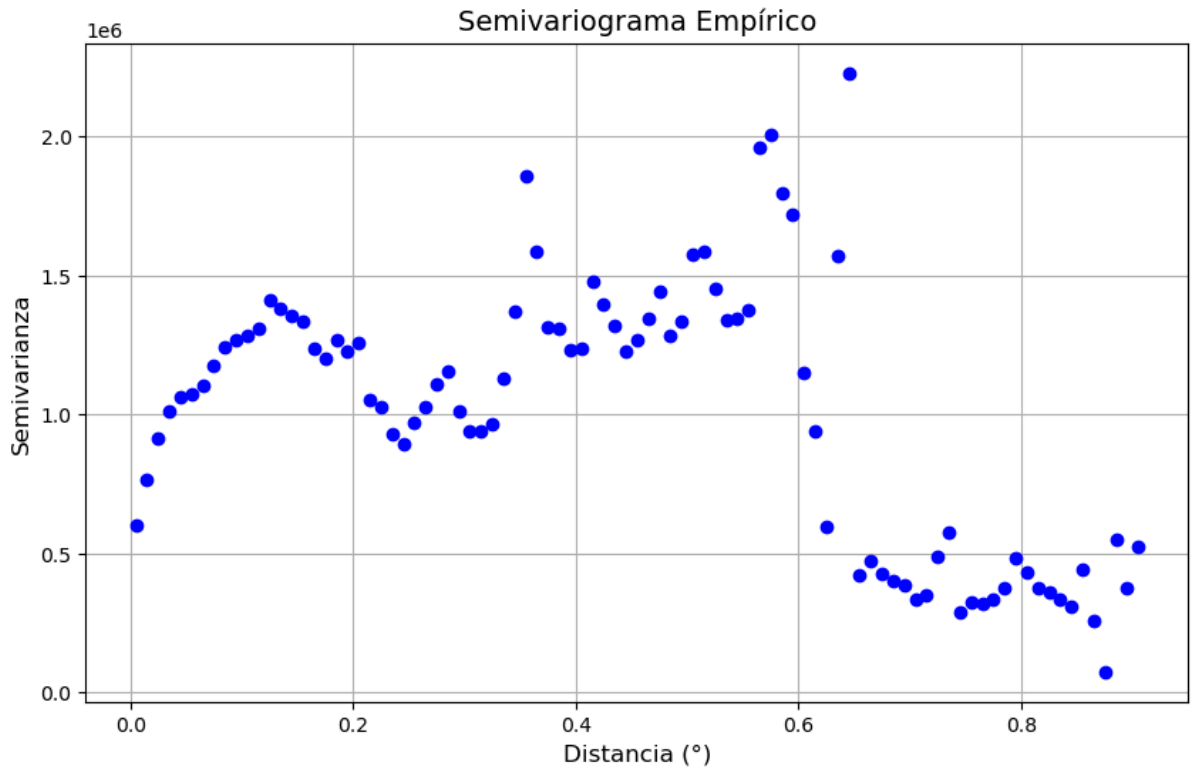

Fig 2: Predicción del precio por metro cuadrado para monoambientes en el AMBA.

Parece que el semivariograma indica una tendencia espacial.

- Visualización de resultados:

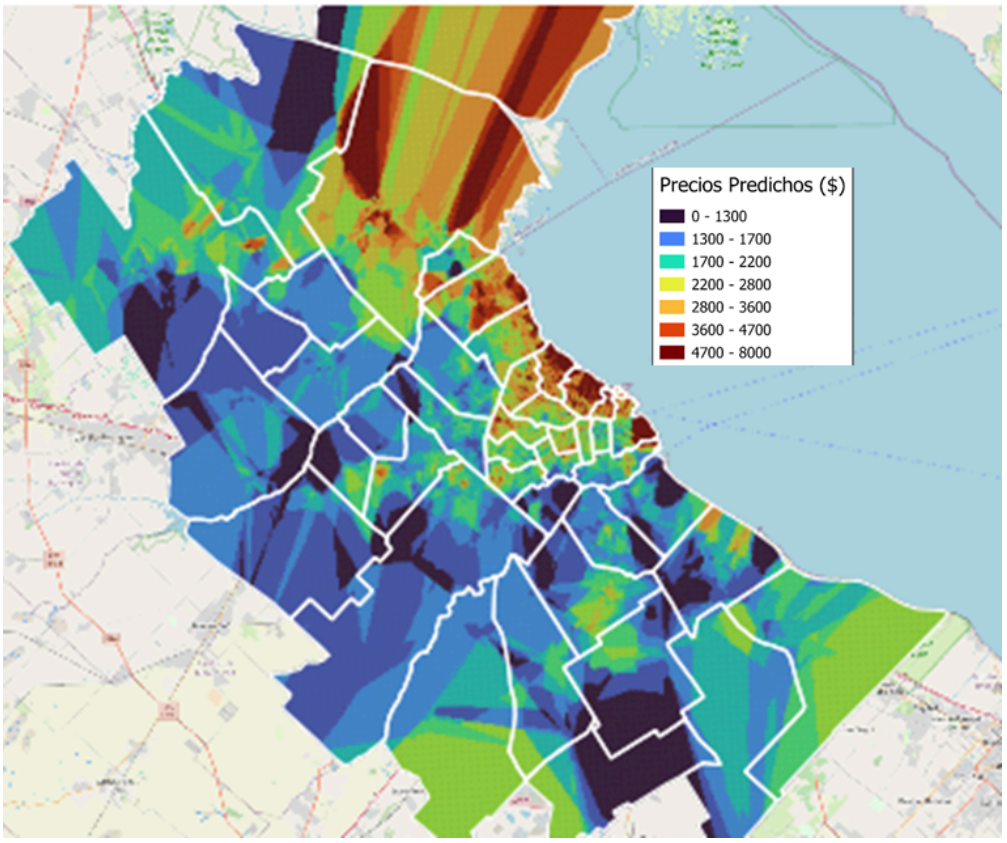

Fig 3: Predicción del precio por metro cuadrado para monoambientes en el AMBA.

El modelo predijo precios de departamentos mostrando mayor precisión en áreas con mayor densidad poblacional.

- Métricas de evaluación:
  El modelo predice con un error medio absoluto (MAE) de aproximadamente 375,51 dólares por metro cuadrado. Otras métricas evaluadas fueron: el error cuadrático medio (MSE) con un valor aproximado de 367707,84 y el R cuadrado con un valor aproximado de 0,69.

- Limitaciones del modelo:
  En consecuencia al alto costo computacional, la selección del modelo se realizó con una muestra de 15000 datos del conjunto original y sin utilizar validación cruzada. El modelo seleccionado fue entrenado teniendo en cuenta la totalidad de los datos.

Conclusiones:

El estudio de los elementos que afectan el precio por metro cuadrado mostró que, aunque elementos como la proximidad al transporte público (subtes, trenes y paradas de colectivos) y la cercanía a áreas verdes ejercen cierto impacto, no son los factores más determinantes.

Al poner en práctica Regression Kriging, se evidenció que este método es eficaz para capturar tanto el efecto de las covariables citadas como los patrones espaciales que subyacen. Este enfoque fusiona la habilidad de los modelos de regresión para calcular efectos con la habilidad del kriging para modelar las correlaciones espaciales locales, facilitando una representación más exacta de los precios basándose en los atributos y la localización. Sin embargo, se detectaron restricciones significativas vinculadas con la disponibilidad de los datos. En algunas zonas del territorio estudiado, la escasez de información suficiente resultó en proyecciones menos exactas, lo que pone de manifiesto la necesidad de incrementar la cobertura de datos en zonas determinadas.

Este problema también resalta la relevancia de disponer de un conjunto de datos sólido y adecuadamente distribuido espacialmente para mejorar la habilidad predictiva del modelo. En futuros estudios, resultaría beneficioso investigar el empleo de datos extra que representen la calidad del ambiente, tales como índices socioeconómicos más exhaustivos, rasgos de los edificios (edad, amenidades).

Finalmente, la investigación demuestra que la mezcla de métodos estadísticos y geoespaciales no solo potencia la comprensión de los elementos que determinan los precios de bienes raíces, sino que también simplifica la toma de decisiones más fundamentadas para urbanistas, promotores de bienes raíces y políticos.