Leal Castillo, J. E. | Vazquez Sanchez, I. A. | Velde, E. van der

Applied Data Science

Finding key determinants of property prices in Mexico City: *A machine learning approach*

Introduction

Mexico City is one of the largest and most dynamic metropolises in Latin America (Aguilar et al., 2003), with a covering area of 1,485 km2 and a population of approximately 22.1 million people (INEGI, 2022). The city presents a complex and challenging real estate market. While the size, age, and condition of a property are typical factors affecting house sales prices, spatial factors, such as access to amenities, crime rates, and overall neighbourhood characteristics should not be ignored when researching determinants for housing prices (Chiang & Tsai, 2016), as ignoring these effects can lead to biased or even misleading conclusions (Ma & Liu, 2015). Understanding the determinants of property sales prices in Mexico City is crucial for policymakers and real estate agencies as they gain a better understand the socioeconomic conditions of different neighbourhoods in the city. Furthermore, comprehending the factors that drive property prices is particularly interesting to real estate investors to make informed decisions on buying and selling properties (Greer & Kolbe, 2003).

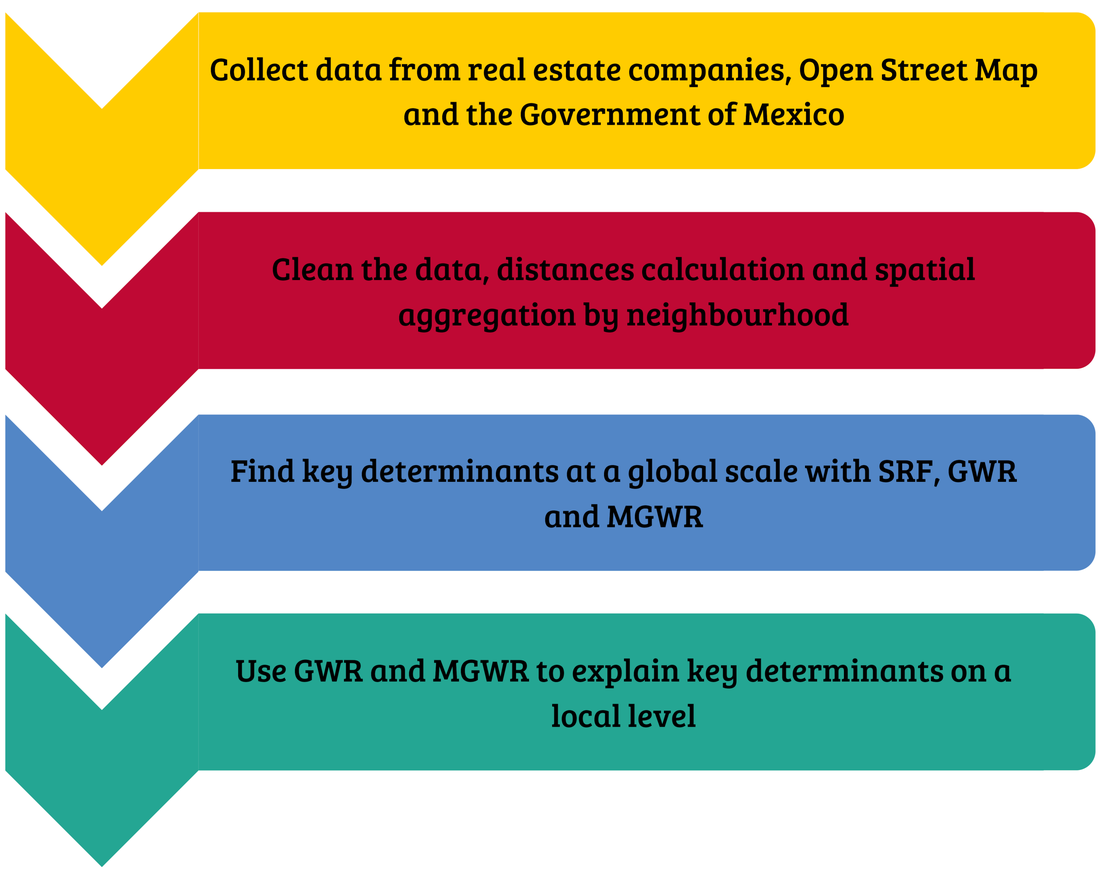
Aim and research questions

This study aims to identify the key determinants of property sales prices in Mexico City and understand how they vary across different geographic locations. To fulfil this aim, the following research questions will be answered:

- What are the key determinants for house prices in Mexico City according to Spatial Random Forest (SRF), Geographically Weighted Regression (GWR) and Multiscale Geographically Weighted Regression (MGWR)?
- Specifically, what are the main determinants for each method and how do these results compare with each other?

Methodology

The research was conducted in multiple stages, listed in the flowchart below.



Housing data was retrieved from a real estate company in Mexico City. This dataset contains locations from 13,798 properties listed on their site between March and November 2022 with basic information as lot size, number of bathrooms and pricing. The dataset was enriched with data from Open Street Map by including distance to amenities and public services. Finally, information on crime rates were added and the data was aggregated to a neighbourhood level. The final dataset contains 36 predictors for a total of 699 neighbourhoods.

SRF, GWR and MGWR were used to find the main determinants of sale prices. For the SRF the variable importance was used as the importance score per variable. Since GWR and MGWR are local models, and do not calculate importance scores by default, the coefficients per predictor variable were used. For GWR and MGWR the predictor variables were normalised with z-score normalisation, since the coefficients obtained by the model are sensitive to the scale of their input. To determine the variable importance, the absolute values of the importance scores were calculated and divided by the sum of these scores as a normalisation.

Finally, local spatial patterns were analysed. This step exclusively employed GWR and MGWR due to their capacity to provide insights into the behaviour of prediction coefficients at a local level (Chen et al., 2019). The coefficients for each variable were mapped, enabling the analysis of their variation across neighbourhoods.

Results

The top 10 predictors influencing the sale price of properties are shown in Figure 1, ranked by the average proportion of explanation across the three models. Each technique yielded varying importance values for each predictor. For instance, MGWR did not consider access to sports facilities to be very important, whereas SRF identified it as an important predictor.

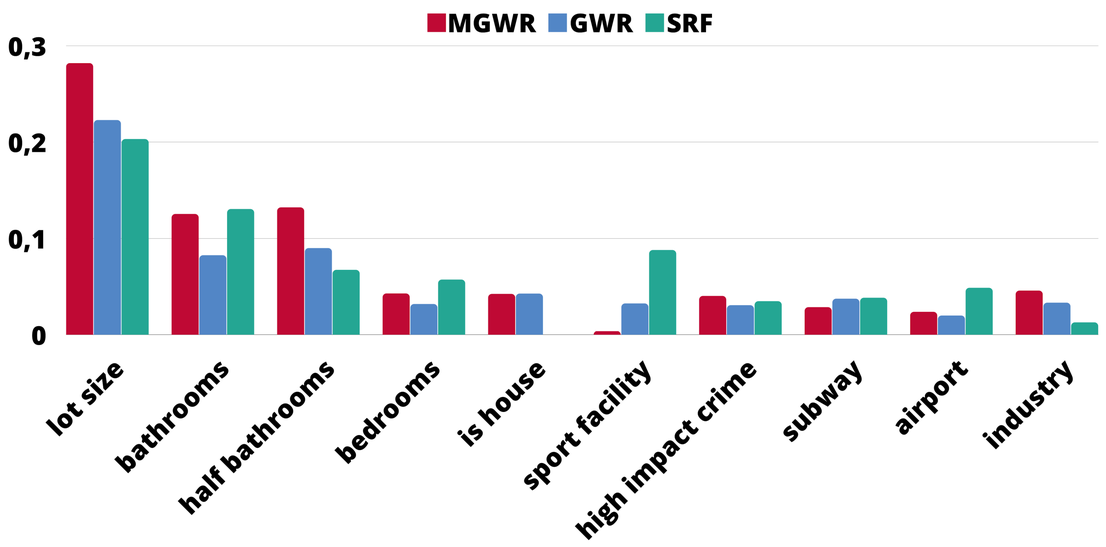


Figure 1. Importance scores per variable for Multiscale Geographically Weighted Regression (MGWR), Geographically Weighted Regression (GWR) and Spatial Random Forest (SRF). Importance is normalised so all scores add up to one. The ten variables with the highest average score are shown.

Looking at the local coefficients for lot size, the most important global variable, GWR and MGWR give results shown in Figure 2. Clusters of high coefficients can be seen near the centre of the city.

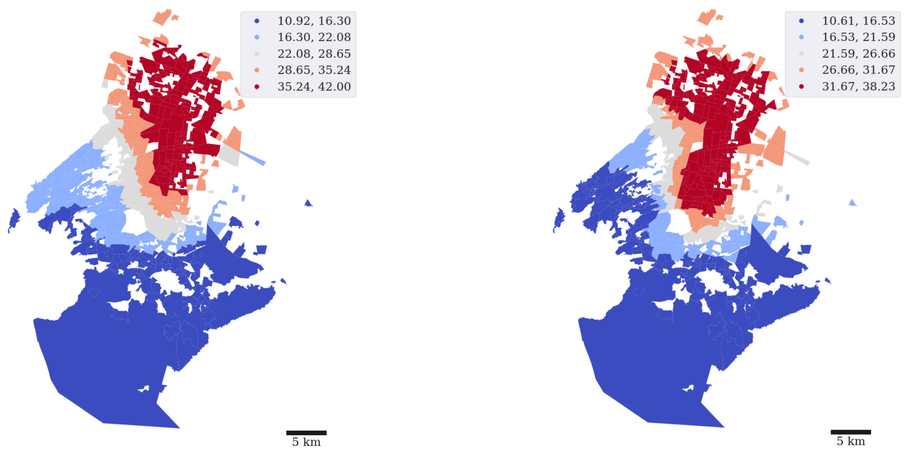


Figure 2. Local regression coefficients for lot size when predicting the price. Left results from Geographically Weighted Regression (GWR), right results from Multiscale Geographically Weighted Regression (MGWR).

Discussion

The results from Figure 1 show lot size, number of (half) bathrooms and bedrooms are the four most important determinants for property prices. This means that when the property is larger and has more (half) bathrooms or bedrooms, the price will be higher. High impact crime is the next most important factor, as this factor has similar and relative high importance in all separate models.

To better understand how predictors vary at the neighbourhood level, GWR and MGWR results were compared. The coefficient values for lot size across the city are shown in Figure 2, both methods depict higher coefficients for the central and northern regions compared to the southern region. This implies that increasing the lot size in these areas would lead to higher property prices compared to the south. The higher coefficient for lot size in the north may be attributed to the higher value placed on space in densely populated areas.

Conclusion

Determinants for house prices in Mexico City were obtained using three machine learning techniques.

1. Geographically Weighted Regression found lot size, the presence of air conditioning, number of bathrooms, half bathrooms and bedrooms to be the key determinants.
2. Multiscale Geographically Weighted Regression found lot size, number of half bathrooms, bathrooms and bedrooms, and distance to industrial zones to be the key determinants.
3. Spatial Random Forest found lot size, number of bathrooms, half bathrooms and bedrooms, and distance to a sports facility to be the key determinants.

It was shown that the models obtained near identical results for the top determinants and conclude that the top determinants are lot size, number of bathrooms, half bathrooms, bedrooms as they are shared key determinants in the models separately. Furthermore, high impact crime is argued to be the fifth key determinant.

The next step in this research would be to investigate whether there are spatial effects, such as autocorrelation and heterogeneity.

Bibliography

- Aguilar, A. G., Ward, P. M., & Smith Sr, C. B. (2003). Globalization, regional development, and mega-city expansion in Latin America: Analyzing Mexico City's peri-urban hinterland. Cities, 20(1), 3-21.
- INEGI: Instituto Nacional de Estadística, Geografía e Informática. (2022). Anuario de estadísticas por entidad federativa. Retrieved from https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=889463909415
- Greer, G. E., & Kolbe, P. T. (2003). Investment analysis for real estate decisions (Vol. 1). Dearborn Real Estate.
- Chiang, M. C., & Tsai, I. C. (2016). Ripple effect and contagious effect in the US regional housing markets. The Annals of Regional Science, 56, 55-82.
- Ma, L., & Liu, C. (2015). Is there long-run equilibrium in the house prices of Australian capital cities. International Real Estate Review, 18(4), 503-521.
- Chen, S., Fang, M., & Zhuang, D. (2019, July). Spatial non-stationarity and heterogeneity of metropolitan housing prices: The case of Guangzhou, China. In IOP Conference Series: Materials Science and Engineering (Vol. 563, No. 4, p. 042008). IOP Publishing.