Utrecht University

IDENTIFYING KEY DETERMINANTS OF PROPERTY SALES PRICES IN MEXICO CITY: A SPATIAL ANALYSIS USING MACHINE LEARNING AND SPATIAL STATISTICS TECHNIQUES

VELDE, E. VAN DER
LEAL CASTILLO, J. E.
VAZQUEZ SANCHEZ, I. A.

WORDCOUNT: 1540
HTTPS://GITHUB.COM/EWOUTVANDERVELDE/SPATIALCOURSE

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 km^2 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, accounting for spatial factors, such as access to amenities, crime rates, and overall neighbourhood characteristics, is inevitable when conducting research on this field (Chiang & Tsai, 2016), and ignoring these effects can lead to misleading conclusions (Ma & Liu, 2015).

Understanding the determinants of property sales prices in Mexico City is crucial for several reasons. Firstly, by examining the impact of factors such as accessibility to amenities and crime rates on property sales prices, policymakers can identify areas experiencing social exclusion or inequality, and design interventions to address these issues (Ziccardi, 2014). Secondly, given Mexico City's significance as a global financial centre (Graizbord, 2003), comprehending the factors that drive property sales prices is pertinent to real estate investors to make informed decisions about buying and selling properties (Greer & Kolbe, 2003).

Despite the significance of this topic, research on this subject is limited for Mexico City (Sobrino, 2014). While Sobrino (2014) proposed a hedonic price model to explore the role of quantitative variables in explaining housing prices in Mexico City, and a recent study attempted to use Automated Valuation Models to estimate the market value of real estate portfolios (Gamboa, 2022), there is still much to be learned about the factors that affect property sales prices, particularly from a spatial perspective.

This study aims to identify the key determinants of property sales prices in Mexico City and understand how they vary across different geographic locations. Thus, the following research questions will be addressed: what are the key determinants for house prices in Mexico City according to Spatial Random Forest (SRF), Geographically Weighted Regression (GWR) and Multiple Geographically Weighted Regression (MGWR)? Specifically, which are the main determinants for each method and how do these results compare with each other? The structure of this paper consists of the following sections: 1) methodology, 2) results and discussion, and 3) conclusion.

Commented [ev1]: Old version suggest the market is complex because the city is large

Commented [ev2]: There is no seccondly

Commented [ev3]: For this city?

Commented [ev4]: In mexico city

Methodology

The research was conducted in multiple stages, beginning with data collection and pre-processing. The study then proceeded with geographical analysis at a global and local level.

1. *Data collection*

A Mexican real estate agency was the source of the housing dataset, containing the location, amenities, and prices of 13,798 properties on sale between March to November 2022 throughout Mexico City, with an average price around $6.5 million MXN (approx. $360,000 USD). Additionally, spatial data such as the boundaries of the neighbourhoods and crime rates during 2022 were collected from official open data sources of the Mexican government (ADIP, 2022, 2023).

Commented [ev5]: This sentance does not really read nicely...

Commented [ev6]: Cititation

2. *Data pre-processing*

This step involved addressing missing variables, translating variable names into English, removing inconsistent data, outliers, and highly correlated variables like construction size and lot size. The number of crimes per neighbourhood was counted, divided by the neighbourhood area and spatially joined with the properties to capture the crime rate at each property location.

To incorporate spatial features, all coordinates were converted to EPSG 6362 and Euclidean distances to amenities, like public services and recreation zones were calculated. The locations of these targets were obtained through the OSM Overpass API. An additional filtering of the targets was manually done based on local knowledge.

Finally, the individual house data points were aggregated by neighbourhoods to obtain the median property prices and predictors for each neighbourhood. This resulted in 699 neighbourhood polygons with 36 predictors to be analysed.

3. *Global Analysis*

SRF, known for its ability to identify spatial predictors while considering spatial relationships (Belgiu & Drăguţ, 2016), was employed to identify determinants of sale prices using permutation of variable importance. This technique evaluates the predictive accuracy of a random forest model when each feature is randomly permuted (Strobl et al., 2008). The centroids of the neighbourhoods were used to create a distance matrix.

The input data was standardized with z-score normalization for the GWR and MGWR analysis. A fixed bandwidth of 115 nearest neighbours was identified based on the lowest AIC criterion using a Gaussian kernel for GWR, while MGWR used the same kernel and criterion to determine the optimal bandwidth for each feature. Both methods returned coefficients for each variable at a neighbourhood level.

To compare the global results from all models, the average of the absolute values of the local regression coefficients of GWR and MGWR, and the absolute value of the importance score for SRF were calculated. Next, the proportion of importance for each feature was calculated by dividing its value by the sum of the weights of the remaining variables.

Commented [ev7]: Not clear

4. Local Analysis

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

Results and Discussion

Figure 1 displays the 15 main predictors, sorted by the mean proportion of importance between the three models. Each technique produced a different importance for each variable. For instance, access to sport facilities was deemed relatively unimportant by MGWR, while SRF considered it an important predictor. Nonetheless, the findings indicate that lot size, number of bedrooms and (half) bathrooms, distance to sport facilities, subway stations, industrial zones, and high-impact crime rates are the overall primary determinants of property prices. Accuracy for the three models is shown in Appendix A.

Commented [ev8]: We don't mension why we keep high impact crime, and we draw different conclusion than in conclusion

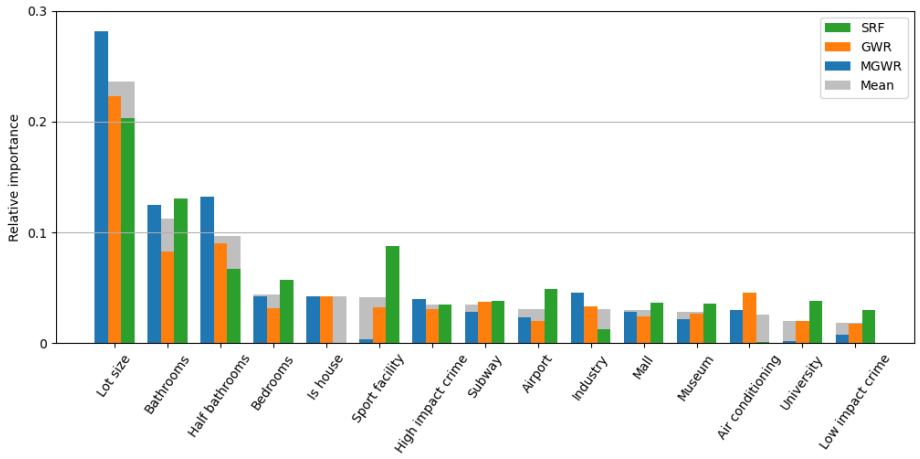


Figure 1. Relative importance scores per variable for Multiscale Geographically Weighted Regression (MGWR), Geographically Weighted Regression (GWR) and Spatial Random Forest (SRF). Importance is normalized so all scores add up to one. The 15 variables with the highest average score are shown and ordered by their mean value.

Commented [ev9]: Suggestion to move this to later in the results when we have shown the results

Figure 2 shows, for both GWR and MGWR, the local regression coefficients for lot size in the central and northern regions have higher coefficients compared to the southern region. This implies that increasing the lot size in these areas would result in higher property prices compared to the south. The higher coefficient for lot size in the north may be attributed to the greater value given to space in densely populated areas.

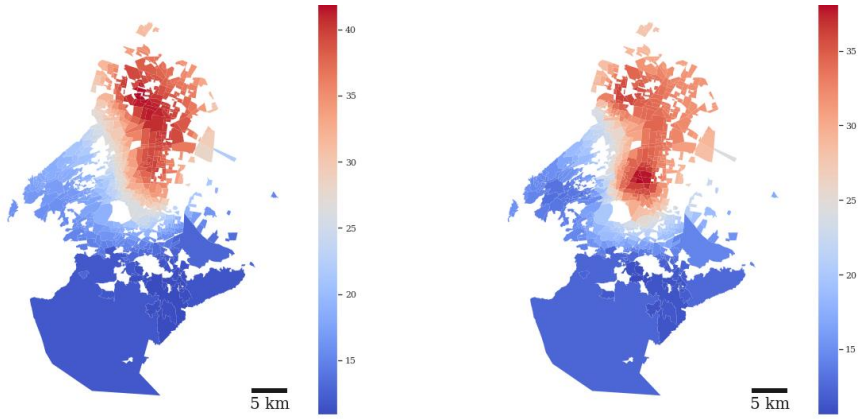


Figure 2. GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for lot size across Mexico City

Figure 3 depicts a higher regression coefficient for distance to industry in the North for both methods. Notably, most industries in the city are situated in the northern neighbourhoods, which could account for the greater value of this factor in that region. Our hypothesis is noise, pollution, and other externalities associated with industrial activities influence the preference for homes farther away from the industrial zones, but this will need further research.

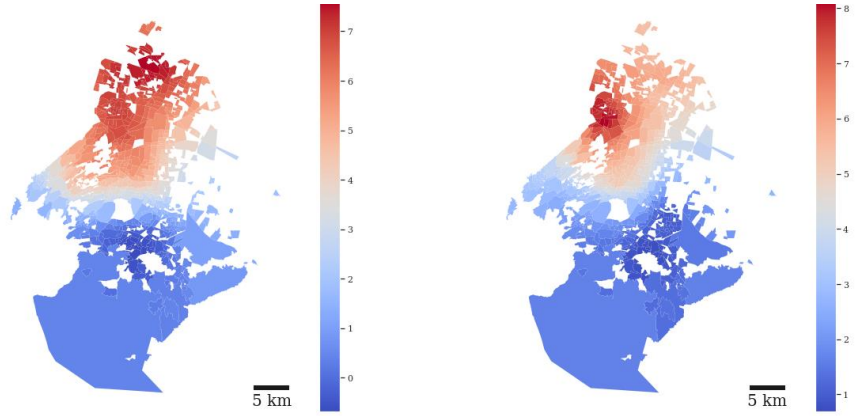


Figure 3. GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to industrial zones across Mexico City.

Figure 4 shows a discrepancy in the outcomes of the GWR and MGWR model regarding the high impact crime rate, The GWR model exhibits that variability across neighbourhoods and crime rates have less influence on property prices near the City Centre. Contrarily, the MGWR model generates coefficients that are almost uniform throughout the city, implying that while the high impact crime rate is an important global predictor.

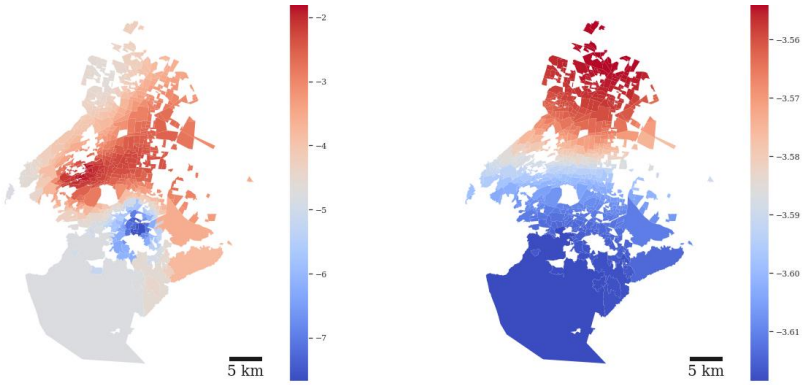


Figure 4. GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for high impact crimes ratio across Mexico City.

The local analyses conducted with GWR and MGWR suggest that the impact of predictors on property sales prices vary across neighbourhoods, highlighting the presence of spatial autocorrelation. Appendix B displays the visualizations of the results for all predictors.

It is important to recognize that this study has certain limitations that should be considered. Firstly, to reduce the number of points to be processed, the data was aggregated by neighbourhood using median and mean values, depending on the variable. This approach has two issues: the mean can be sensitive to outliers, and the median ignores the variability of the data. Furthermore, the models assume all variables are equal for each property in a neighbourhood, leading to ecological fallacy.

Secondly, Euclidean distances between centroids were calculated. To obtain more accurate results, the shortest path when travelling over the road network of Mexico City could be used instead. Additionally, an improvement in distance calculations would be to get the distance from the house location to the edge of an area (such as parks) instead of from centroid to centroid.

Finally, it is worth noting that property owners can choose not to disclose the exact location of their property, leading to two potential issues: some properties may be assigned to an incorrect neighbourhood; and the accuracy of the computed distances may be compromised.

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 bathrooms, half 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.

Based on the analysis, the models demonstrated consistent results for their top determinants, with lot size, number of bathrooms, half bathrooms, and bedrooms being identified as key predictors in all models. Additionally, distance to sport facilities, subway stations, industrial zones, and high-impact crime rates were found to be important predictors by at least one of the methods. Notably, the last three variables had a consistently high relative importance across all models.

Since local regression coefficients were varying considerably between GWR and MGWR, the next step in this research would be to investigate whether there are spatial effects such as autocorrelation, heterogeneity, stationarity, and multicollinearity present in the data. By answering these questions, a better understanding on how each predictor contributes spatially can be gained.

References

Agencia Digital de Innovación Pública (ADIP). (2022). Portal de Datos Abiertos de la CDMX, Fiscalía General de Justicia (FGJ) de la Ciudad De México. Retrieved from https://datos.cdmx.gob.mx/dataset/carpetas-de-investigacion-pgj-cdmx.

Agencia Digital de Innovación Pública. (2023). Portal de Datos Abiertos de la CDMX, Instituto Electoral de la Ciudad de México. Retrieved from https://datos.cdmx.gob.mx/dataset/carpetas-de-investigacion-pgj-cdmx.

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.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing, 114, 24-31.

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.

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.

Gamboa, M. (2022). Automated Valuation Models and Real Estate Pricing in Mexico. Mexico Business News. Retrieved from https://mexicobusiness.news/tech/news/automated-valuation-models-and-real-estate-pricing-mexico.

Graizbord, B., Rowland, A., & Guillermo Aguilar, A. (2003). Mexico City as a peripheral global player: The two sides of the coin. The Annals of Regional Science, 37, 501-518.

Greer, G. E., & Kolbe, P. T. (2003). Investment analysis for real estate decisions (Vol. 1). Dearborn Real Estate.

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

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.

Sobrino, J. (2014). Housing prices and submarkets in Mexico City: A hedonic assessment. Estudios económicos, 57-84.

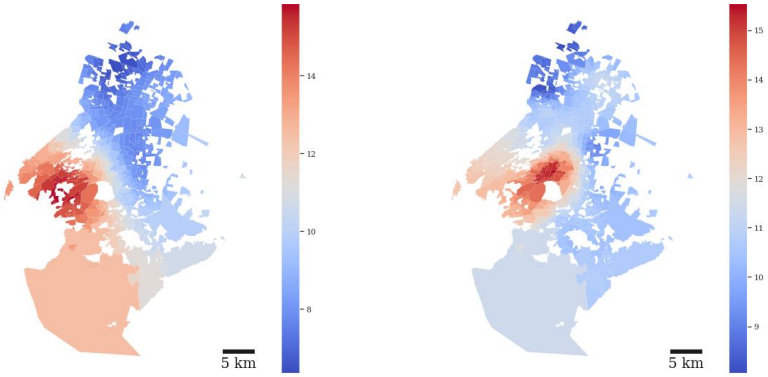
Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. BMC bioinformatics, 9, 1-11.

Ziccardi, A. (2014). Poverty and urban inequality: the case of Mexico City metropolitan region. international social science Journal, 65(217-218), 205-219.

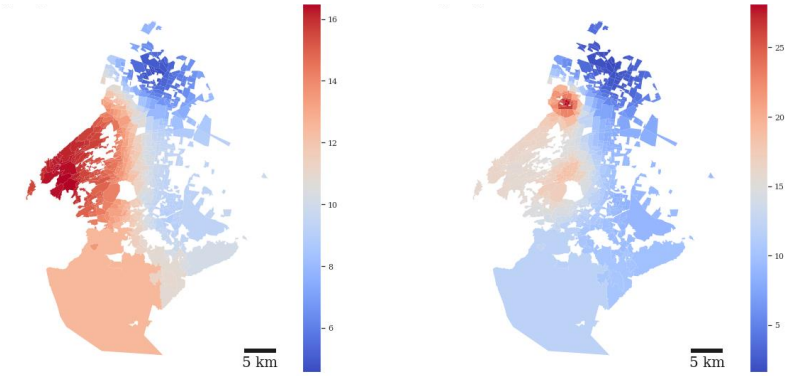
Appendix A

Model	R^2	AIC
GWR	0.733	3219.48
MGWR	0.743	3115.61
SRF	0.676	-

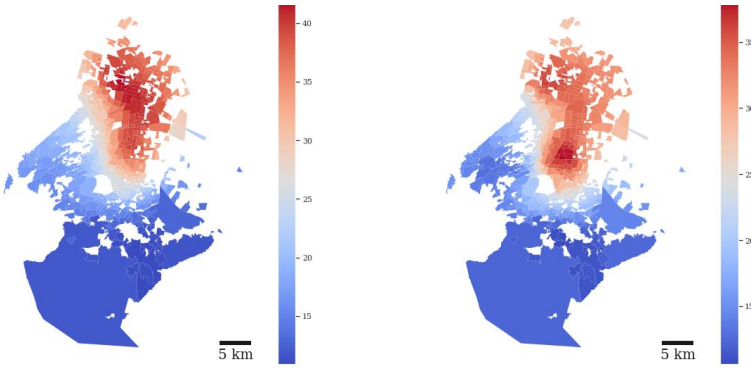
Appendix B



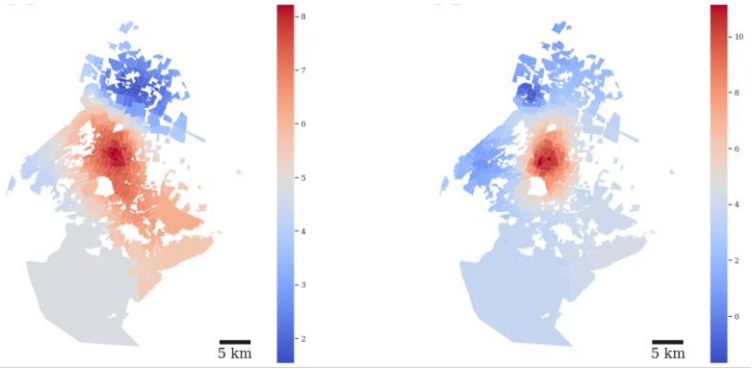
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for number of bathrooms across Mexico City.



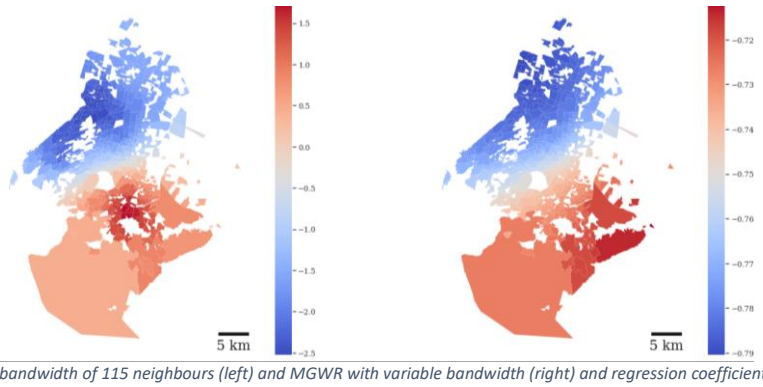
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for number of half-bathrooms across Mexico City.



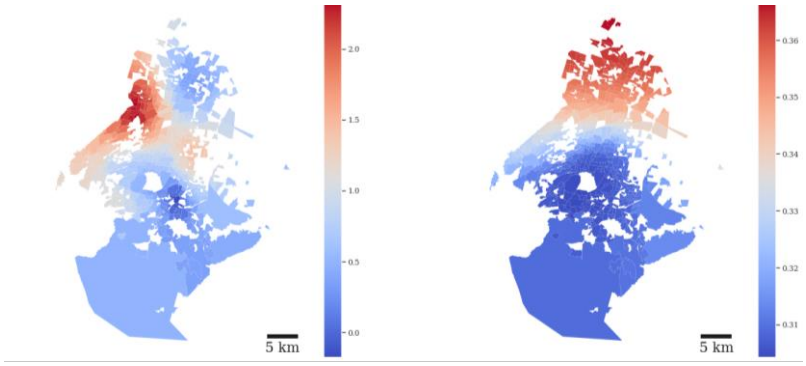
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for lot size across Mexico City.



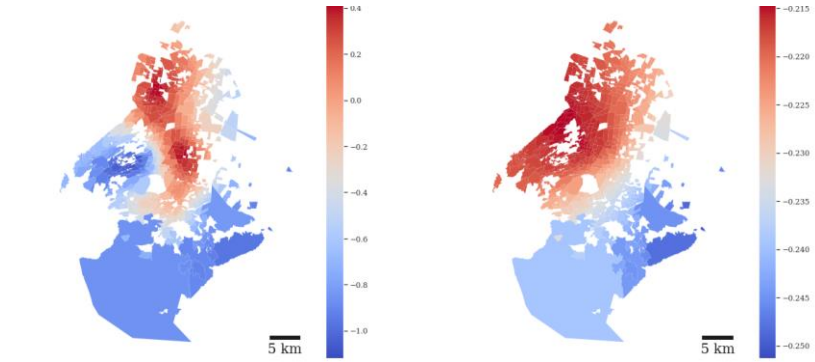
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for is-house.



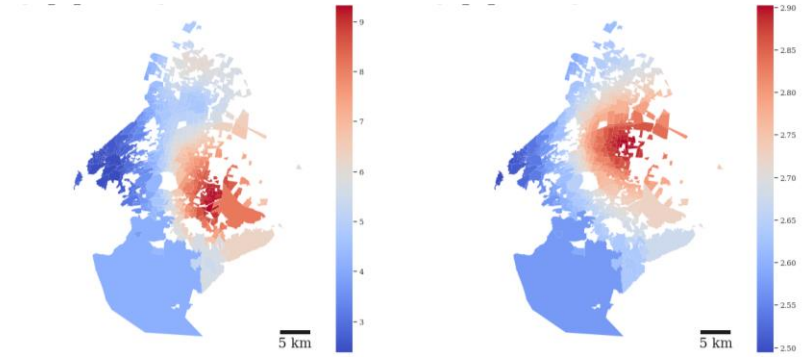
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for property age.



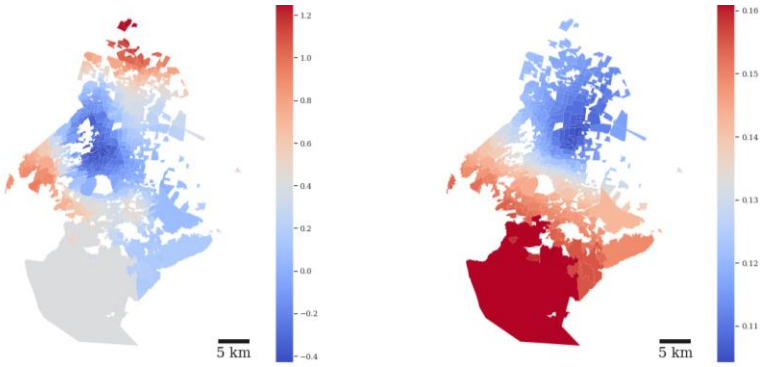
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for balcony.



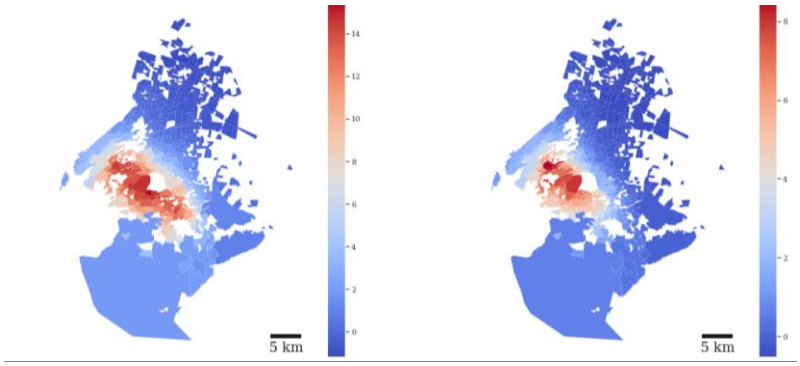
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for water storage.



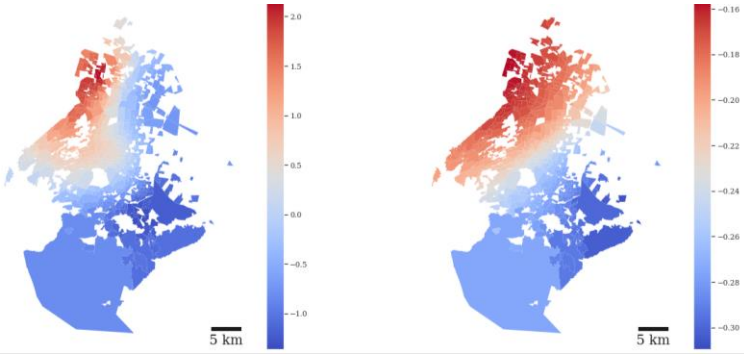
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for air conditioning.



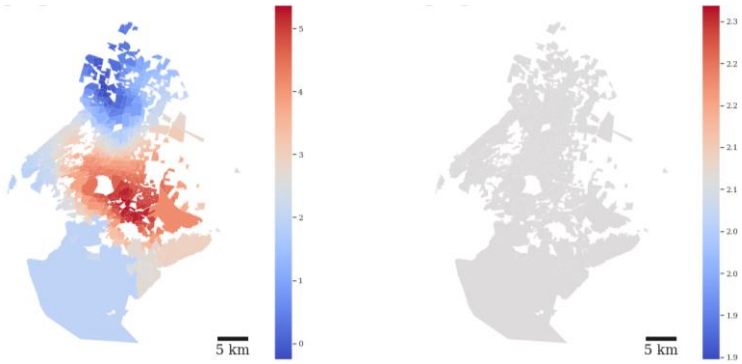
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for pool.



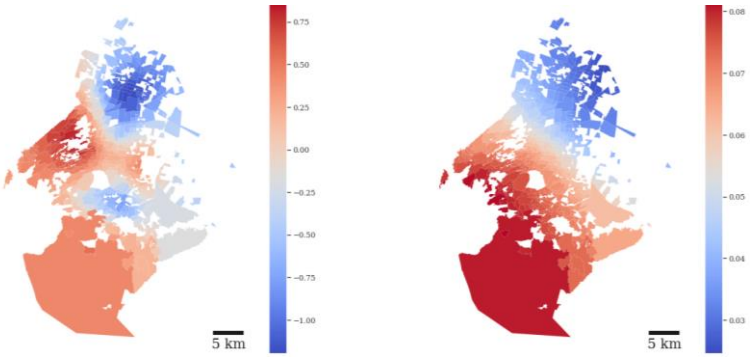
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for furniture.



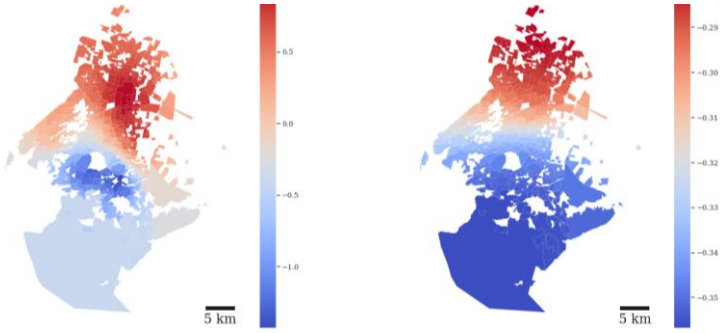
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for handicap accessible.



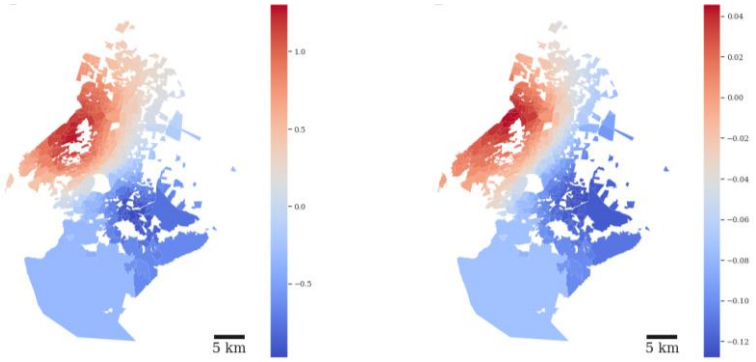
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for first floor.



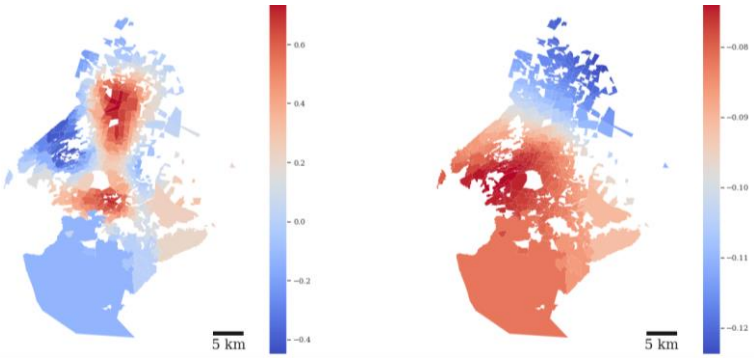
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for security.



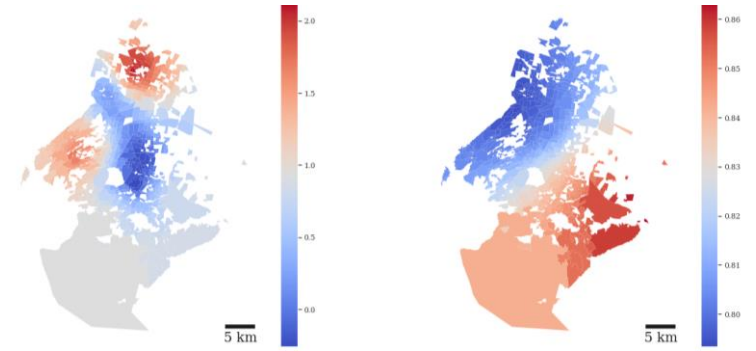
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for parking space.



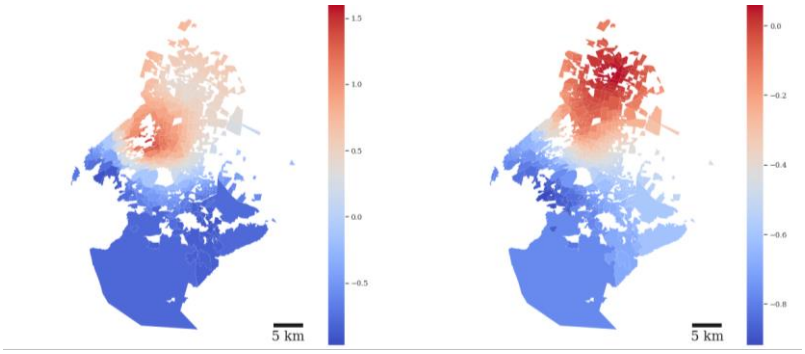
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for kitchen.



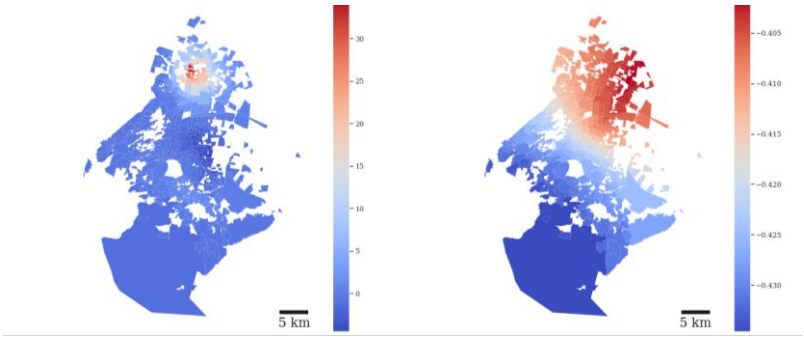
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for garden.



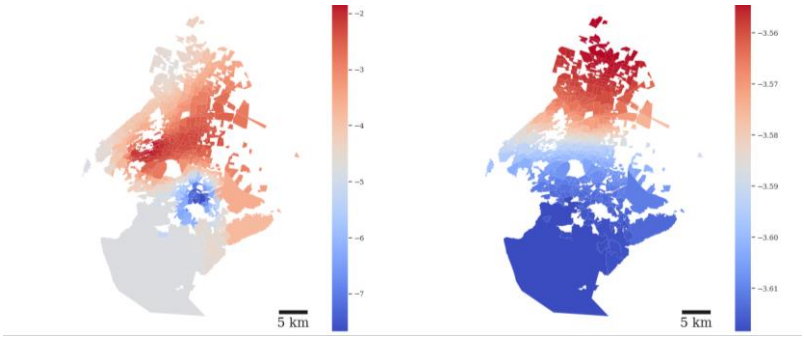
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for roof-garden.



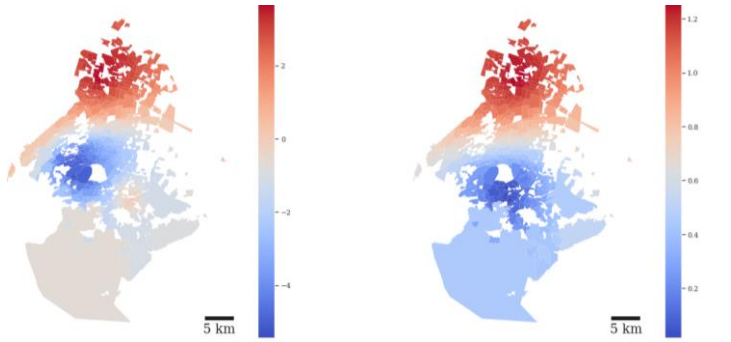
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for service room.



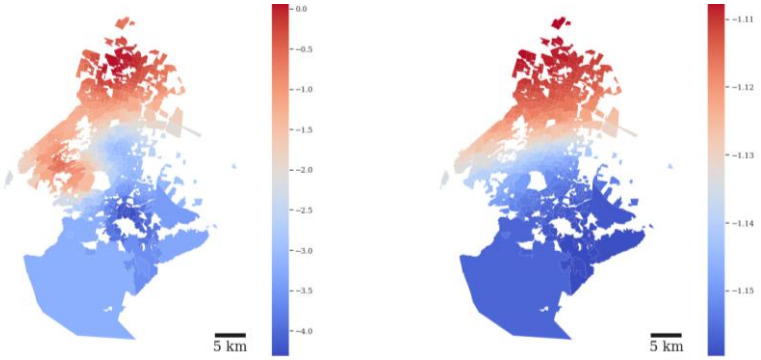
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for gated community.



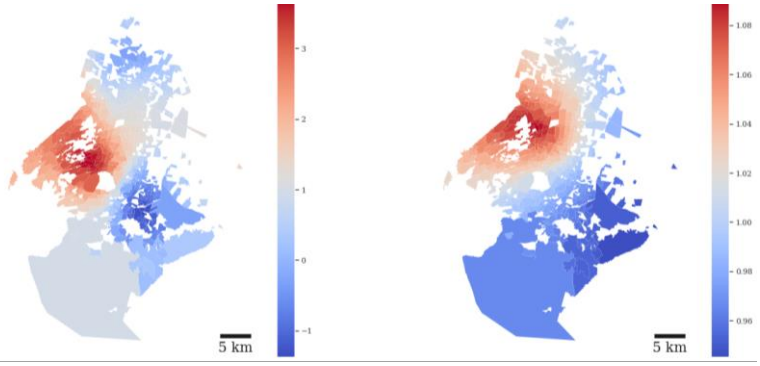
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for high impact crime rates per area.



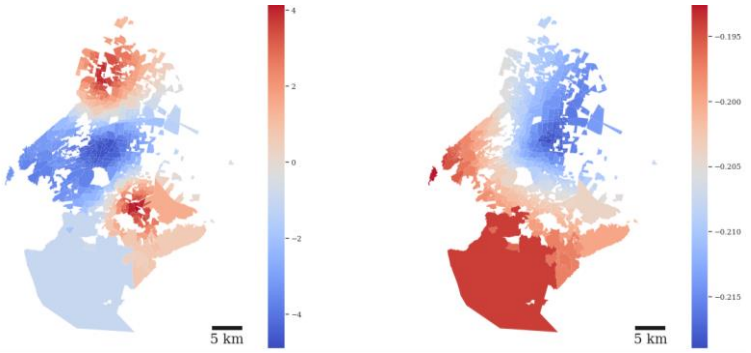
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for low impact crime rates per area.



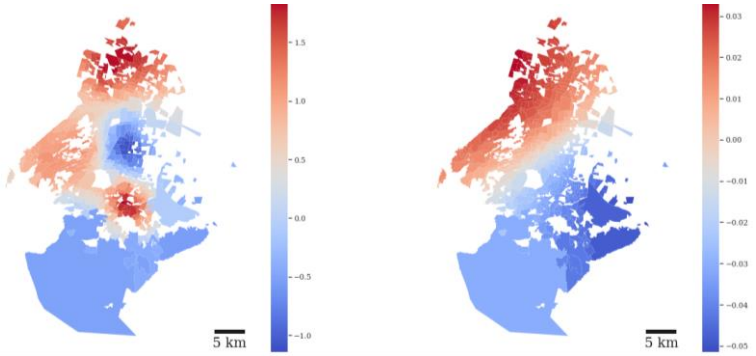
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to nearest park.



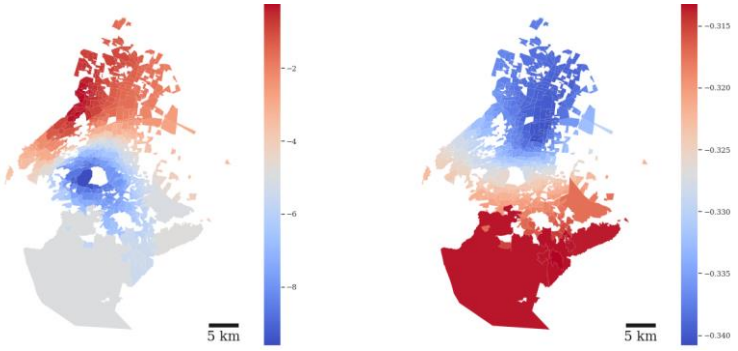
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to nearest school.



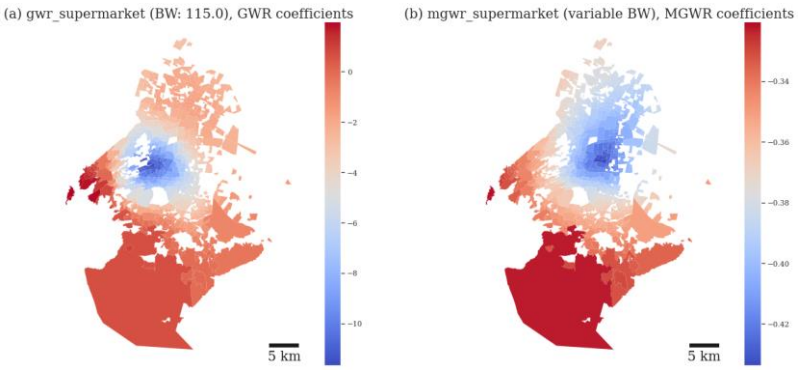
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to nearest university.



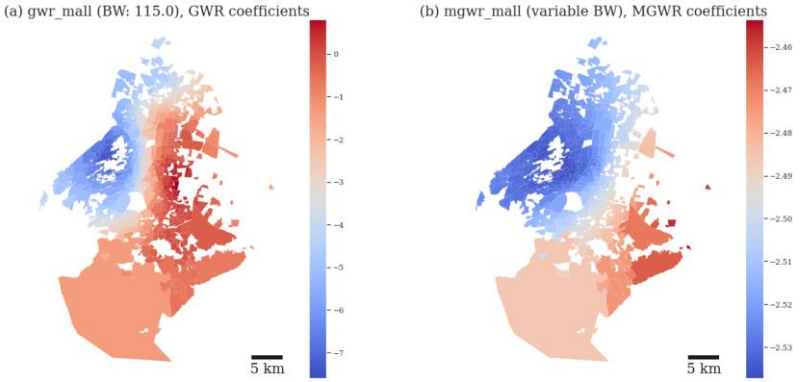
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to nearest hospital.



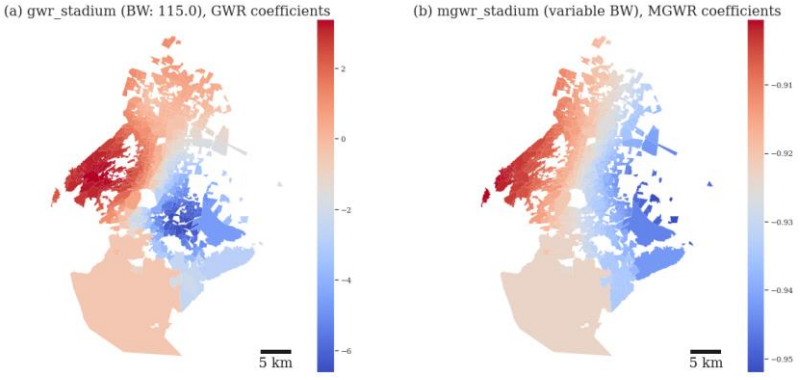
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to nearest sport facility.



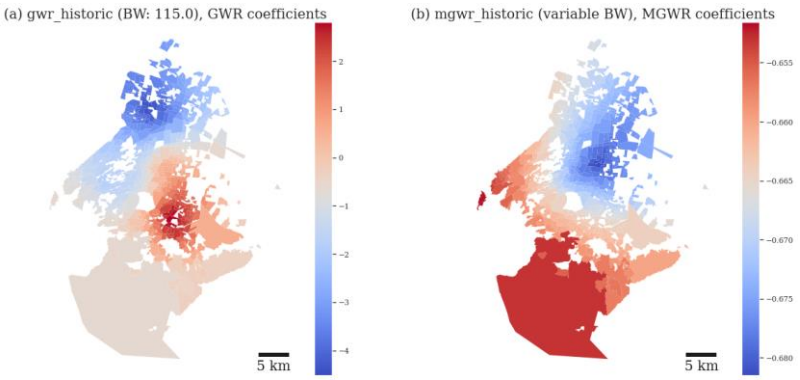
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to nearest supermarket.



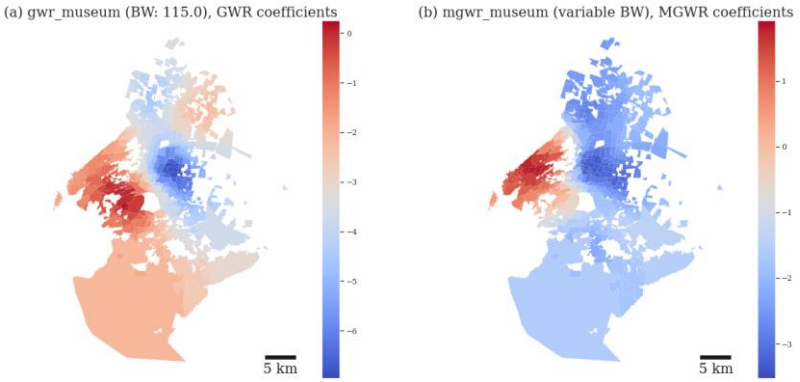
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to nearest mall.



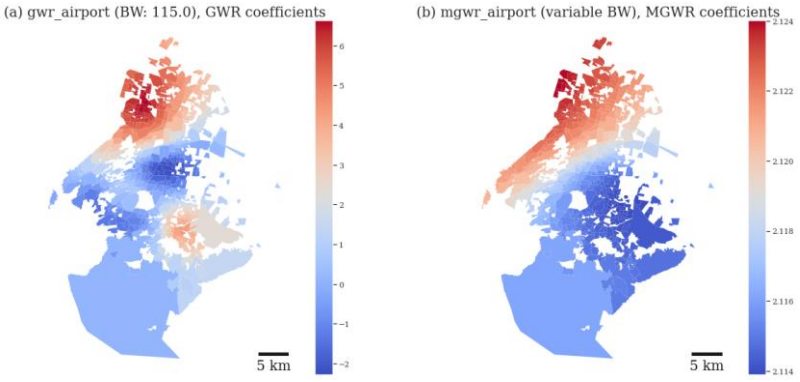
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to nearest stadium.



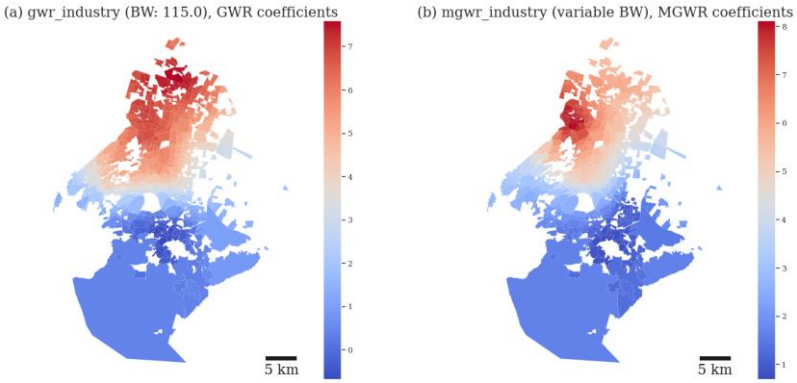
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to nearest historical building.



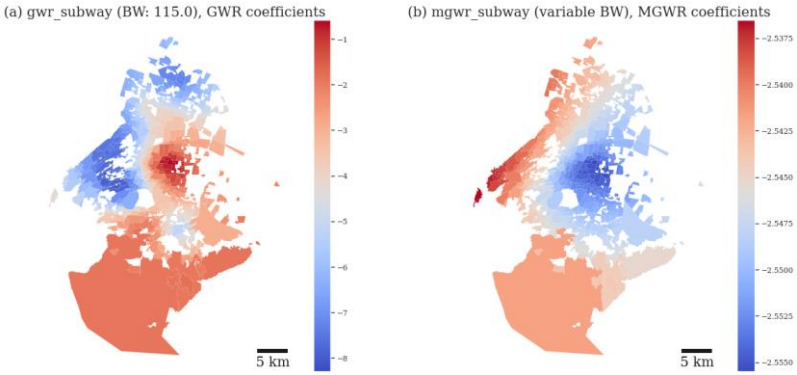
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to nearest museum.



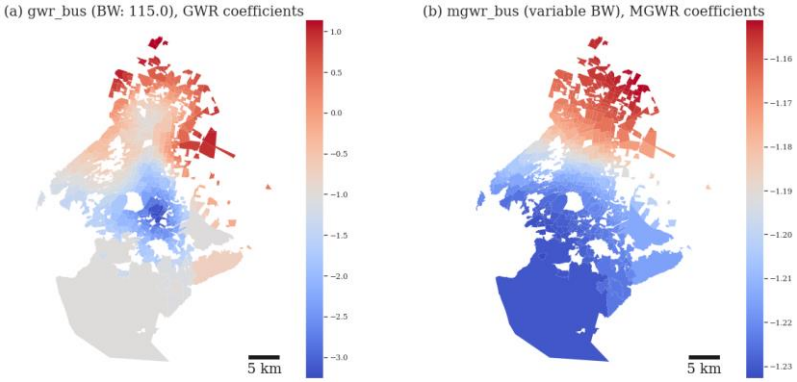
GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to the airport.



GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to industry area.



GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to nearest subway station.



GWR with a bandwidth of 115 neighbours (left) and MGWR with variable bandwidth (right) and regression coefficients for distance to nearest bus stop.