Metro Impact on the Property Value of Surrounding Areas: A Case Study in São Paulo, Brazil

Guilherme F. Alves[a]

[a]Department of Transport Engineering at the Polytechnical School, University of São Paulo, São Paulo, Brazil.

ABSTRACT

This study investigates the influence of metro networks on property values in São Paulo, Brazil, leveraging extensive data on property tax (IPTU) and real estate valuation from 1995 to the present. The research adopts spatial econometric methods, combining data science and spatial analysis, to explore the relationship between metro accessibility and residential property prices. The results indicate a non-linear, positive correlation between metro proximity and property valuation, aligning with global urban development trends. This paper contributes valuable insights into urban metro impacts on real estate, offering implications for urban planning and policy-making in rapidly urbanizing cities.

CONTACT: Guilherme Fernandes Alves   ✉: gf10.alves@gmail.com

1 Introduction

This document describes a research project from the 'Geospatial Data Science' course at the University of São Paulo, under supervision of the Professor Mariana Giannotti. The course teaches students to combine data science with spatial analysis. After foundational data science instruction and practical labs, students apply their knowledge to a project of their choice. This project examines how metro networks affect nearby property values.

The Brazilian '*Imposto Predial e Territorial Urbano*' (IPTU), also known as the 'Property Tax', is a tax collected every year from property owners in Brazil. This tax data, along with property values, is released annually by the Government of each city, including the city of São Paulo. Records are available from 1995 to the present. This extensive dataset enables studies on how the metro network influences property values in nearby areas.

The paper starts with a literature review, which was not exactly the purpose of the course so it may be a bit shorter than expected. Next, the methods used to conduct the analysis are described in detail, allowing for the replication of the study if desired. Results are then shared and discussed, and the paper ends with the concluding remarks. Notably, there is an appendix section to save valuable information that may not be necessary for understanding the research.

2 Literature Review

The impact of metro accessibility on residential property values is a topic of growing importance in urban planning and economic studies. Given this context, the work of Li (2018) provides valuable insights for the city of Xi'an, China. The study's main finding is a generally positive, non-linear relationship between proximity to metro stations and property values. The greatest increase in property values is observed at distances ranging from 300 to 1,200 meters from metro stations.

In a similar fashion, the research conducted by Sun et al. (2016) investigates the influence of subway line construction on residential property values in Tianjin, China. Utilizing an Hedonic Pricing Model, this study examines the changes in property values within a 1km radius of the newly completed subway line 3 stations. The findings revel a positive impact of subway lines on surrounding property values, with a more noticeable effect in the outskirts than in downtown areas. This suggests that the impact of subway construction on residential prices varies across different areas of the city.

Contrasting with the aforementioned studies, the research by Campos and Almeida (2018) focuses on São Paulo, Brazil. This study aims to decompose the implicit effects, neighborhood, and adjacency in influencing property prices both within and between properties, as well as districts. The results reveal that a significant portion (96.89%) of the property price is determined by intrinsic and locational characteristics, while a smaller fraction (3.11%) is attributed to neighborhood and adjacency effects, including the characteristics of the districts and spatial spillover among them.

Based on these insights, this study seeks to quantify the impact of metro stations on surrounding property

values in São Paulo. An equation will be proposed to estimate property values, incorporating the distance to the metro network and its temporal dynamics.

3 Methods

It is worth to start this section by mentioning the dataset used. The IPTU data from São Paulo is publicly available. However, a significant effort have been made by Fernando Gomes and Salvarani (2022) to condense this data, enhancing its usability. Specifically, this study utilizes the summarized property values per transport zone for each year, without any alterations.

To start conducting the analysis, the official transport zones of the city of São Paulo, provided by the São Paulo Metro Company (de São Paulo 2017), were used. This segmentation of the city is created aiming the analysis of origin-destination matrices of the public transport network, which are build every 10 years for the city of São Paulo. For this study, the data from 2017 was used, which is the most recent one. Along with the spatial data, different variables are also included in the dataset, such as the number of inhabitants, number of jobs, and number of households.

Socio-economic variables relevant were chosen for clustering the transport zones. These variables include population, school enrollments, employment, area (in hectares), and per capita income. To run the clustering operation, the k-means method (MacQueen et al. 1967) was selected due to its simplicity and ease of interpretation.

Determining the optimal number of clusters is not a trivial task, and there are many methods to do so. Although different optimization schemas were considered [e.g. the kneed algorithm (Arvai 2020)], the final number of clusters was selected based on the ease of interpretation of the results. In this case, the number of clusters was constrained up to 5, which is a reasonable number of groups to be analyzed and interpreted. Sensitivity analysis was conducted to understand the best number of clusters within the range of 2 to 5.

The next step involves validating the hypothesis that property values are influenced by geographic location. This was examined using the Moran's I test (Moran 1950) for spatial autocorrelation of property values.

After validating the autocorrelation of property values within the city, it was necessary to construct the variables for the final model. The first important variable is the distance to the metro network. To accomplish this, a dataset representing all the metro stations built up to each year in the analysis was constructed. Determining the inauguration date of each metro station was crucial, made possible by the dataset provided by Davi (2019).

The transport zones remain unchanged over the years. For each zone, its centroid was first calculated. Then, varying the years, the distance from the centroid of each zone to the 10th closest metro station was calculated for each year. This approach was applied to each transport zone and for each year. Considering the 10th closest metro station was important to ensure significant variability in the constructed variable, thereby improving the model's quality.

Additionally, it was necessary to convert the property values, available for each year, to their present

value. This was done using the national inflation index, IPCA, obtained from the Brazilian Institute of Geography and Statistics (IBGE), the official index provider.

First, the IPCA for each month, ranging from 1981 to the present, was provided. However, it was necessary to convert this to an annual IPCA. This conversion was achieved by calculating the equivalent annual interest rate, as expressed by equation (1).

$$IPCA_{year} = \prod_{i=1}^{12}(1 + IPCA_{month_i}) - 1 \tag{1}$$

where $IPCA_{year}$ is the annual IPCA, and $IPCA_{month_i}$ is the IPCA for the ith month of the year.

The property values were converted by applying the annual interest rate for each year up to the present, as expressed by equation (2).

$$PropertyValue_{present} = PropertyValue_{year} \times (1 + IPCA_{year})^{(year_{present} - year)} \tag{2}$$

Visualizing the data is crucial for understanding the distribution of variables and identifying possible outliers. This step is challenging, as the outcome variables may be excessively noisy, affecting the visualization. To address this issue, a Savitzky-Golay filter (Savitzky and Golay 1964) to smooth the data and make it easier to visualize without losing the main characteristics of the data.

Finally, to build the model, an exponential regression model was required, as expressed by equation (3).

$$z = \beta_0 \times e^{\beta_1 \times x + \beta_2 \times y} \tag{3}$$

where z is the property value, x is the distance to the metro network, and y is the time.

Before moving to the next section, it important to notice that all the code used to conduct this analysis is available on GitHub, thus allowing for future replication or enhancements of the study.

4 Results and discussion

The parameters and coefficients obtained from the model are as follow. The first coefficient for distance to the metro network is -0.00018. The second coefficient, related to time, is 0.020941040. Intercept: The model's intercept is found to be -34.390.

These finding reveal that 0.00018 is the rate of change of the property value as the distance to the metro network increases. In other words, the property value decreases by 0.00018 for each meter away from the metro network. Similarly, the property value increases by 0.020941040 for each year that passes by.

Figure 1 allows one to visualize these results, showing the actual property values (post-filtering) as blue dots and the exponential fit as a surface, with values associated with a color scale.

Overall, as expected, the property values increases as the distance to the metro network decreases. Also, the property value increases as the time passes by, which is a good indicator of the inflation of the property value.
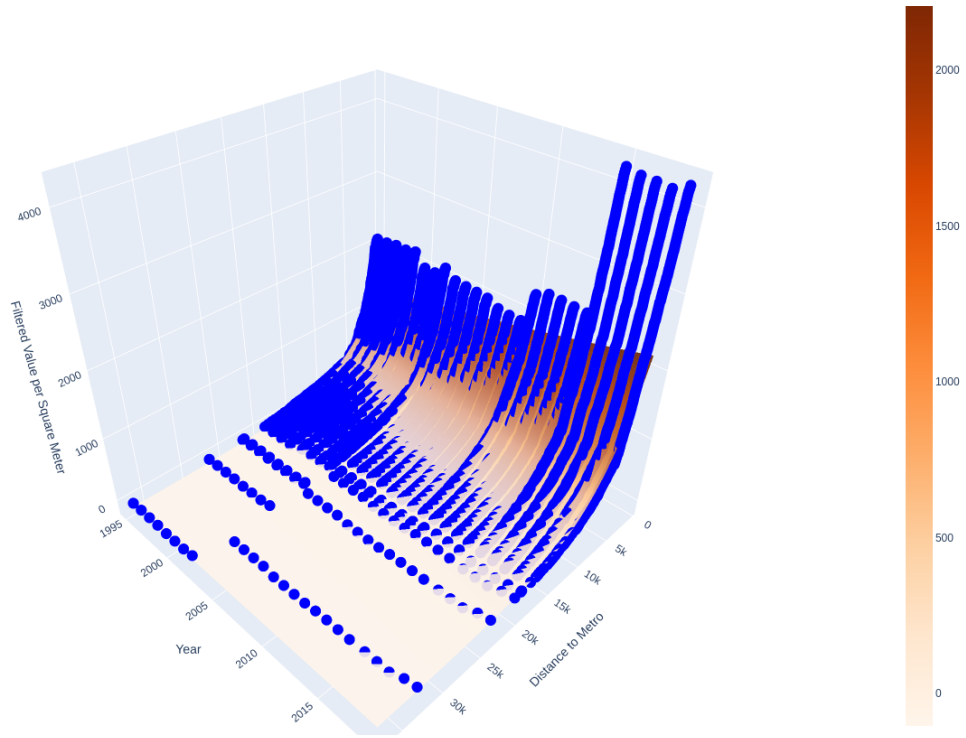
Figure 1 - Property value as a function of the distance to the metro network and the time. The blue dots are the actual values after filtering, and the surface is the exponential fit with its values associated to the color scale.

The goodness of fit of the model was found as 78.12% when comparing the predicted values against the real values.

It is noticeable that the exponential fit is a good option to model the property value based on the distance to the metro network and the time. Nevertheless, It is important to note that the metro network is not the only variable that may impact the property value. This constitutes a limitation of the model, as variables intrinsic to the property itself, such as the number of rooms, may also impact the property value but were not considered in this study since they were not available in the dataset.

As a cavear, it is valuable commenting that the influence of the metro network on the property value decreases as the distance to the metro network increases, achieving a point where this influence is negligible. This could be used to estimate the maximum distance to the metro network where the property value is still impacted by the metro network, thus providing a valuable information for the city planners and future researches.

5 Concluding remarks

By the end of this study, it was possible to conclude that the metro network has a positive impact on the property value of the surrounding areas. This is in line with the literature, and the model was able to explain significant part of the property value.

To future research on top of this study, it is suggested to consider other variables that may impact the

property value. Alternatively, investigating the change of land cover over time and its relationship with the metro network may also be an interesting research topic. LiDAR data could be used to extract the land cover information, and the metro network could be used as a proxy for the urbanization of the city. As a last suggestion, a predictor function could be built to estimate the property value for residences near to the planned future metro stations.

Finally, it is worth to mention that this article has been written for a specific course at the University of São Paulo, and the author is open to suggestions and corrections. Sending the content to an international journal or conference was not the initial intention, but it is not discarded. If that is the case, a second author may be needed.

References

K. Arvai. kneed, Aug. 2020. URL https://github.com/arvkevi/kneed.

R. B. A. Campos and E. S. d. Almeida. Decomposição espacial nos preços de imóveis residenciais no município de são paulo. *Estudos Econômicos (São Paulo)*, 48(1):5–38, Jan 2018. ISSN 0101-4161. URL https://doi.org/10.1590/0101-416148115rce.

Davi. Inauguração de estações do metrô-sp e da cptm, 2019. URL https://www.kaggle.com/datasets/dsrdavi/inauguracao-de-estacoes-do-metrosp-e-da-cptm/.

M. de São Paulo. Pesquisa origem e destino. https://transparencia.metrosp.com.br/dataset/pesquisa-origem-e-destino, 2017. Accessed: 2021-06-01.

M. A. G. Fernando Gomes and L. Salvarani. Painel cadastral da cidade de são paulo, sep 2022. URL https://doi.org/10.55881/cem.db.dash01.

Z. Li. The impact of metro accessibility on residential property values: An empirical analysis. *Research in Transportation Economics*, 70:52–56, 2018. ISSN 0739-8859. URL https://doi.org/10.1016/j.retrec.2018.07.006.

J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950. ISSN 00063444. URL http://www.jstor.org/stable/2332142.

A. Savitzky and M. J. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

H. Sun, Y. Wang, Q. Li, and F. R. Villatoro. The impact of subway lines on residential property values in tianjin: An empirical study based on hedonic pricing model. *Discrete Dynamics in Nature and Society*, 2016:1478413, 2016. ISSN 1026-0226. URL https://doi.org/10.1155/2016/1478413.

6 Appendix

Consider this section as a place to save valuable information that may not be included in the main text. Reading this section is not mandatory for completely understanding the research.

Finally, we emphasize that the work has been made available on-line and any contributions are welcome. The smallest, but still valuable, contribution is to star the repository on GitHub.
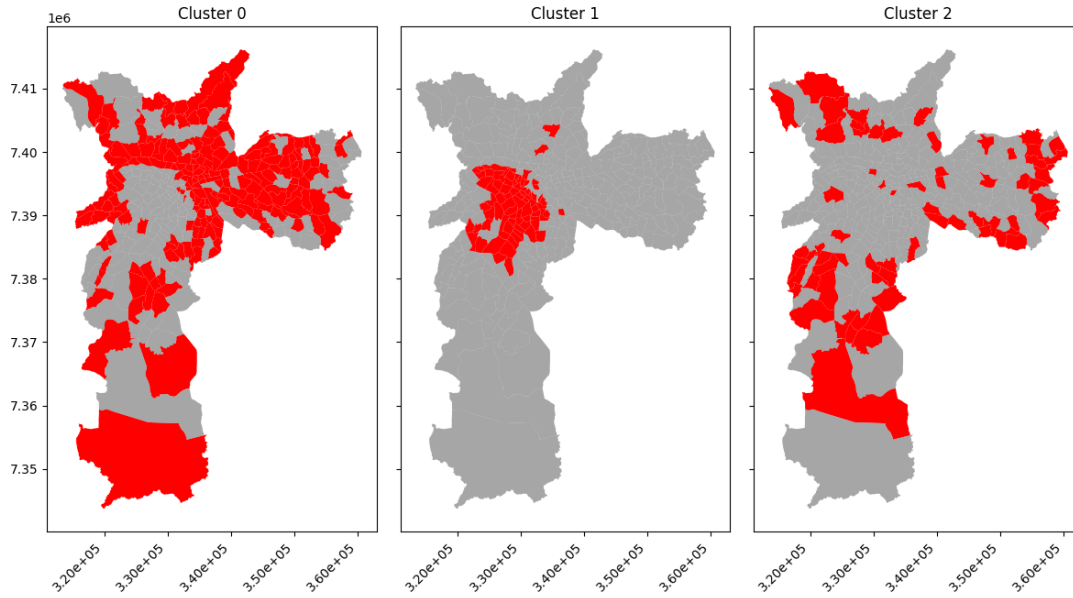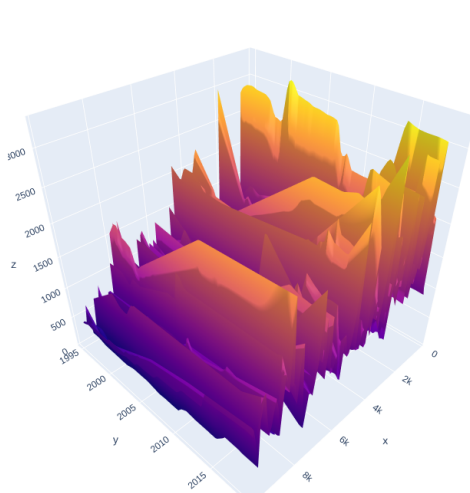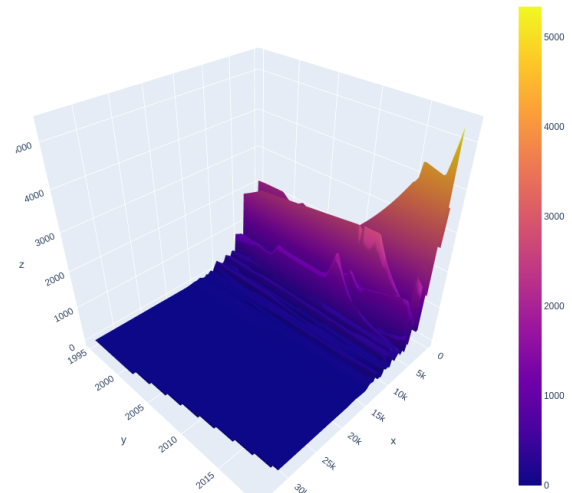

Figure 2 - Cluster analysis of the transport zones of the city of São Paulo based on socio-economic variables. The number of clusters was selected based on the ease of interpretation of the results.

(a) Actual Data

(b) Filtered Data

Figure 3 - Property value as a function of the distance to the metro network and the time.
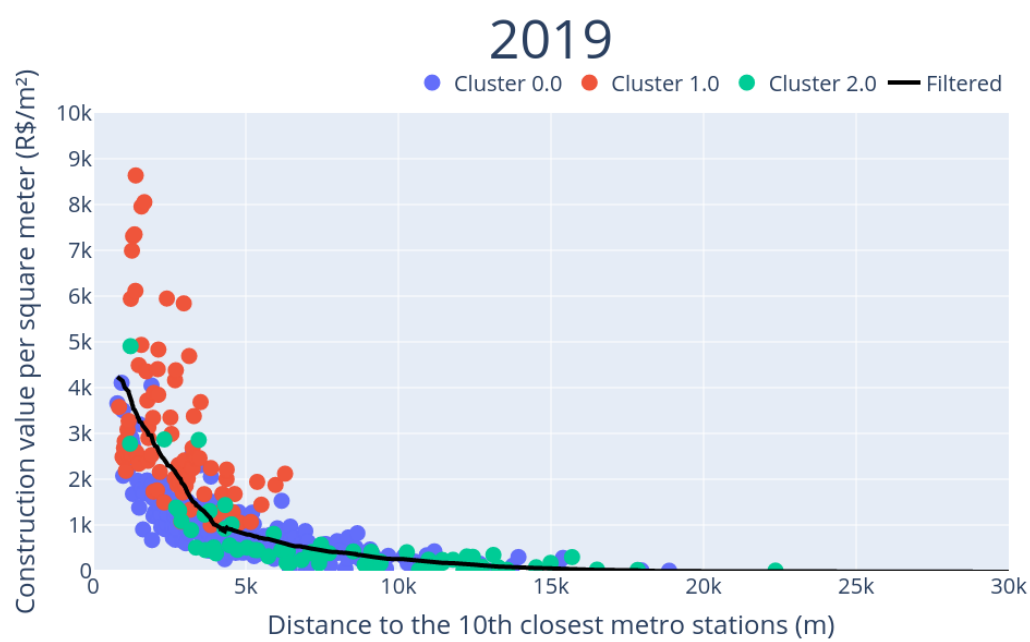
7

Figure 4 - 2D plot of the constructed area for each transport zone in the city of São Paulo. The color of the points represents the cluster to which the transport zone belongs. Cluster 1 is the central area of the city, typically with higher property values.