Data-driven Urban Inequality: Infrastructure, Income and Environmental conditions of São Paulo municipalities.

Carlos Simão Struckas Filho
Mestrando em Administração e Sociedade - PPGAdS
Universidade Federal de São Carlos UFSCAR
São Carlos, Brazil
orcid.org/0000-0003-1696-7411

Mário Sacomano Neto
Professor Titular do Mestrado Profissional em Administração e Sociedade - PPGAdS
Universidade Federal de São Carlos UFSCAR
São Carlos, Brazil
orcid.org/0000-0002-2561-1700

Fábio Molina da Silva
Professor Titular do Mestrado Profissional em Engenharia de Produção - PPGPEP
Universidade Federal de São Carlos UFSCAR
São Carlos, Brazil
https://orcid.org/0000-0002-0407-1231

Abstract*— This study examines the relationship between urban infrastructure, income distribution, and environmental conditions across municipalities in São Paulo State, Brazil, within the context of Urban planning and Smart Cities. Using aggregated data from the 2022 Brazilian Census, the research applies statistical and machine learning techniques, including K-means clustering, correlation analysis, linear regression, and a Naive Bayes model. Results identify clusters linking low income, limited infrastructure, and reduced urban arborization, indicating priority areas for public policy. However, no significant linear relationship was found between income and infrastructure variables, possibly due to data aggregation limitations. The Naive Bayes model achieved 73.37% accuracy in predicting arborization levels, demonstrating its usefulness for urban planning. Overall, the study highlights the potential of data-driven approaches and artificial intelligence to support decision-making and promote more efficient, equitable, and climate-responsive urban policies.

Keywords—Urban Planning, Public Infrastructure, Climate Justice, Data Analysis, Inference algorithms.

I. Introduction

Traditionally, it is perceived in the urban planning literature that better serviced lands were concentrated into areas with higher income, leaving significant parts of society into areas which lack basic services such as water, sanitation and electricity. [1], [2], [3]

When we consider the emergence of climate change, the problem is compounded by the need to identify areas where climate threats are most prevalent, which generally again falls into areas with greater social vulnerability.[4], [5], [6], [7], [8]

All those are relevant challenges to the urban administration, whose would be faced by the implementation of Smart Cities technologies, a concept that refers not only to the automation of urban management processes and controls, but mainly to cities that are able to collect data from their territory in order to compile it and provide information that can support government decisions and the evaluation of implemented public policies, ensuring greater efficiency and effectiveness to public management.[9], [10], [11]

The proposal of this research was to apply data analysis technics to identify quantitative evidence of the need to implement policies that provide serviced lands and the concentration of those services into higher income areas in São Paulo State.

II. Method & Procedures

This research was developed as a quantitative case study and as part of the content of the Data Analysis class of the UFSCAR master program.

For the intent of this study, it was used aggregated data at the municipal level for of the São Paulo state from the year of 2022, available at the tables "Agregados_por_municipios _renda_responsavel_BR" and "Agregados_por_municipios _entorno_domicílios_BR", derived from the most recent Brazilian Census.[12]

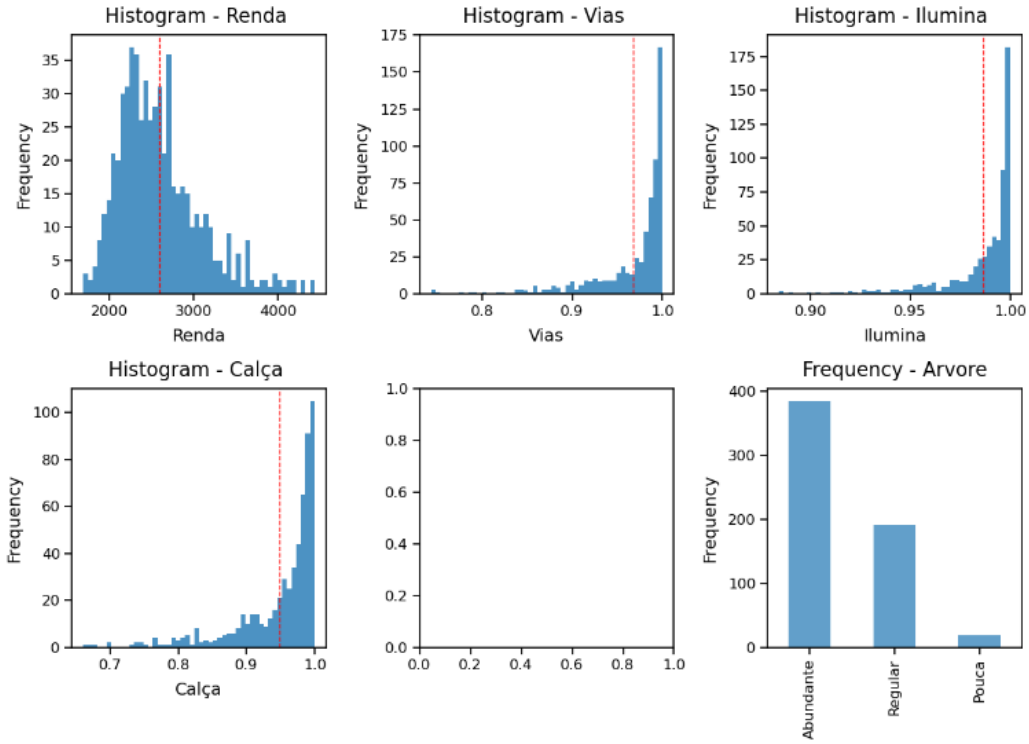
The collected data were analyzed in an iterative manner using Python, Visual Studio built-in extensions, and AI-based tools to apply various statistical analysis methods and data structuring techniques. The objective was to identify relevant patterns related to urban structure and household income, grounded in established data analysis literature. [13], [14]

As proxy for family's income was used the "mean of the income of the responsible with income of permanent occupied houses" and for urban infrastructure was used variables: i. coverage of paved streets; ii. coverage of urban illumination; iii. coverage of sidewalks; and, iv. number of trees on the block face; as summarized in TABLE I.

TABLE I. Original Dataframe Varibles.

Name	***Variable***	***Type (Unit)***
RENDA	mean of the income of the responsible with income of permanent occupied houses	Continuous (R$)
VIA	coverage of paved streets	Continuous (%)
ILUMINA	coverage of urban illumination	Continuous (%)
CALÇA	coverage of sidewalks	Continuous (%)
ARVORE	number of trees on the block face	Categorical ("Few" <= 2 trees "Regular" >2 and <5 "Abundant" >5 trees)

Fig. 1. Histogram and frequency of data variables.



The full code and datasets used in this research are accessible at: https://anonymous.4open.science/r/Urban-infra-x-Trees-x-Income-183C/

III. Data and Discussions

The original database was composed of 645 original registers, each representing a set of data from a specific São Paulo State city, and was explored throughout the research by different approaches, procedures and data treatments, whose are summarized into this section.

A. *Descriptive Statistics*

The first step for any statistical approach is the descriptive statistics, a series of procedures that allow for an initial understanding of the dataset, and the interpretation of the situation represented into the data and possible information that can be obtained through more in-depth analysis.[13]

First procedure inserted in the code was the elaboration of the histogram and frequencies of the variables, which allowed to interpret predominant characteristics of the analyzed municipalities, and their means (red line), as shown in Fig 1.

Another analysis conducted was the execution of the boxplot, this technique group data by quartiles and would allow the researcher to check the dispersion of the results and evaluate the existence of outliers.

For this dataset we would infer a concentration of the data values around the 1º and the 3º Quartile (blue section), but also a high number of outliers register (small circles), as shown in Fig 2.

Fig. 2. Bloxplot of numerical Variables.

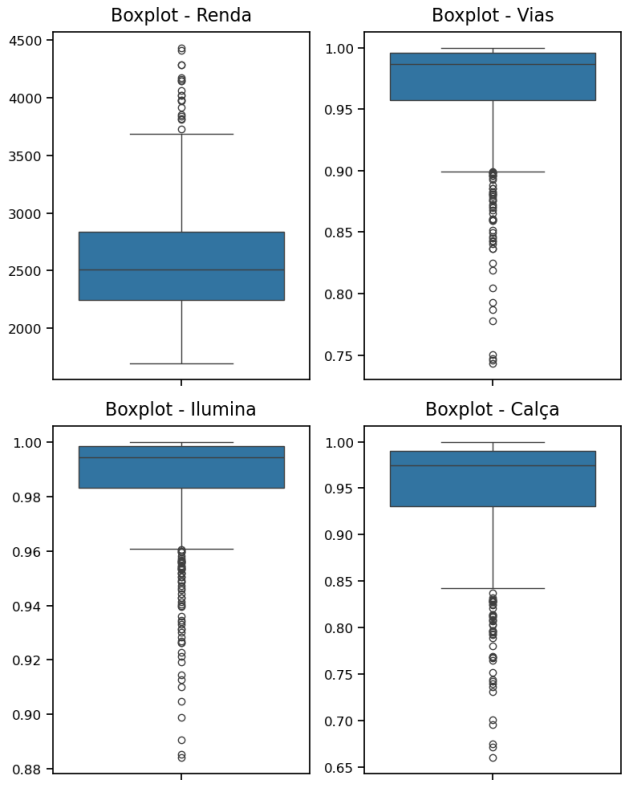


Fig. 3. Clusters of RENDA x INFRA x ARVORE, 3 groups.

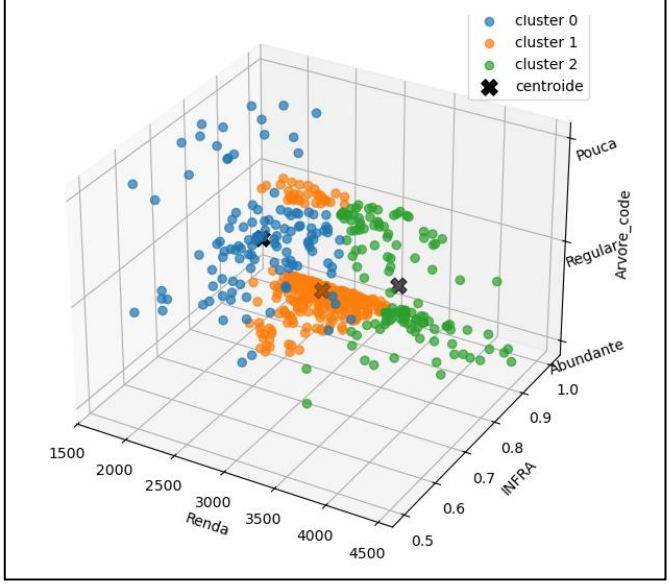
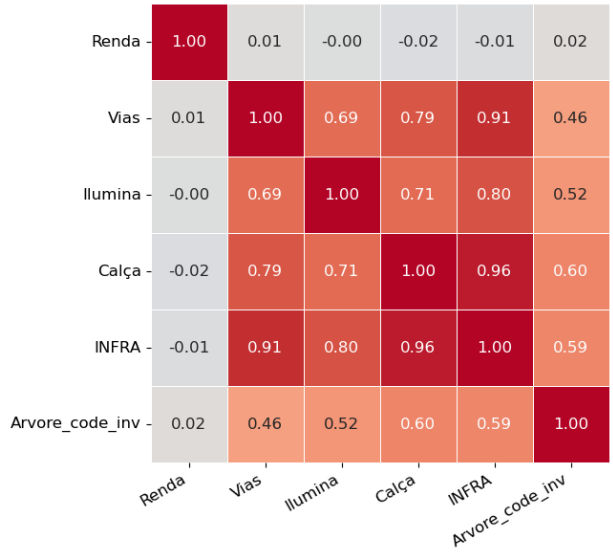


Fig. 4. Correlation matrix of DF variables.



B. *K-means*

K-means analysis has the intent to group similar data registers into an expectation to identify possible classification for those registers.

For this purpose, the dataframe (DF) was evaluated the grouping data into 2D and 3D matrices. After some iteration was observed the necessity to remove outliers from the DF. which were made through the z-score procedure, resulting in the elimination of 50 registers.

Even after the removal of the outliers from the DF the analysis returned unsuccessful, at first glance, resulting in irrelevant results. However, after scrutinizing data, it was proposed to aggregate data from VIA, ILUMINA and CALÇA into a unique variable called INFRA, measured by the result of the multiplication of the prior 3 variables.

After this procedure, was re-run the k-mean algorithm and generated a 3D graphics of the aggregated data, show at Fig. 3, where we would observe that the municipalities would be classified into 3 different clusters: i. Cluster 0: cities whose have families with lower income, regular public arborization and lower urban infrastructure; ii. Cluster 1: cities with medium family income (around R$2.200), more abundant public arborization and higher urban infrastructure; and, iii. Cluster 2: cities with families of higher income, a regular public arborization and a mostly high urban infrastructure.

The analysis also provides a perception that the existence of urban infrastructure is an important driver for the existence of a dense urban arborization.

C. *Correlation and Linear regression*

The first procedure to evaluate the possibility of a linear regression is to observe the existence of correlation between the observable and dependent variables, done by the calculation of the correlation matrix. As shown in Fig. 4, it was perceived into the DF the existence of correlation between the observable variables (usually stated as correlation ≥0,6), but not with the dependent variable RENDA, which were almost 0.

As expected by the low correlation, although all alternative treatments as: i. change of data type (continuous, categorical, log, exponential, aggregated), ii. elimination of outliers, iii. rearrangement of dependents and independent variables; linear regression won't resulted into any significant relations (measured by R^2), as exemplified into Fig 5.

D. *Naive Bayes prediction model*

One last procedure implemented into the algorithm was the proposal of a machine learning procedure that would classify the city expected arborization, measured by ARVORE, based on the data of VIA, ILUMINA and CALÇA.

This approach would be of great importance to urban planners when testing their urban development models, as it may predict the expected results and even measure the impact by the implementation of the proposed interventions or to evaluate with new identified areas, normally irregular settlements.

For this effort, the code was elaborated to use half of the dataset, randomly picked, to train the prediction model, and

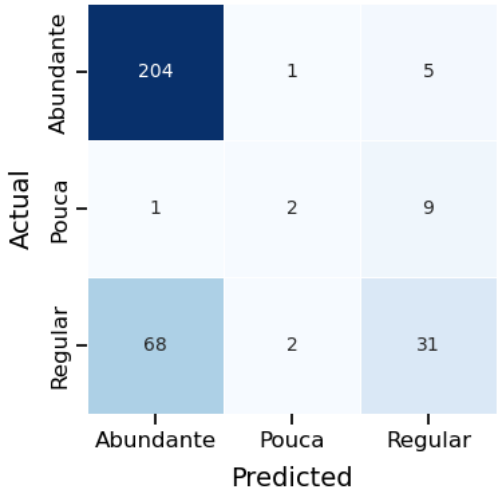
Fig. 5. Exemple of Linear Regression analysis with backward-elimination.

```
Início do processo de eliminação recursiva (backward-elimination) sem RendaC

Modelo com 5 features: ['Vias', 'Ilumina', 'Calça', 'Arvore_Pouca', 'Arvore_Regular']
R2 test: 0.0127 | MSE test: 205564.7795
Menor |coef|: Arvore_Regular = 45.012732

Modelo com 4 features: ['Vias', 'Ilumina', 'Calça', 'Arvore_Pouca']
R2 test: 0.0098 | MSE test: 206173.6784
Menor |coef|: Ilumina = 80.298388

Modelo com 3 features: ['Vias', 'Calça', 'Arvore_Pouca']
R2 test: 0.0101 | MSE test: 206102.4158
Menor |coef|: Arvore_Pouca = 573.059520

Modelo com 2 features: ['Vias', 'Calça']
R2 test: -0.0041 | MSE test: 209059.5261
Menor |coef|: Calça = 510.901032

Ordem de remoção: ['Arvore_Regular', 'Ilumina', 'Arvore_Pouca', 'Calça']

Resumo do processo:
01) remove Arvore_Regular  | R2 0.0127 | MSE 205564.7795 | restantes 4
02) remove Ilumina         | R2 0.0098 | MSE 206173.6784 | restantes 3
03) remove Arvore_Pouca    | R2 0.0101 | MSE 206102.4158 | restantes 2
04) remove Calça           | R2 -0.0041 | MSE 209059.5261 | restantes 1
```

the other half was used to test the accuracy of the model. This approach is recommended to eliminate eventual auto affirmative bias.

The Confusion Matrix (Fig. 6) would give a better insight into the quality of the prediction model, as it confronts the predicted values against the actual DF values. As can be seen, the model was better at identifying if the arborization is "Abundante" (higher than 5 trees per block face), having correctly classified 204 registers of a total of 210. On the other hand, the model has wrongly classified most of the register with actual "Pouca" (Few) and "Regular" values, therefore should not be used to predict those classifications, resulting in an overall accuracy of 73,37%.

Fig. 6.Confusion Matrix for ARVORE classification.



IV. Conclusions

Although the data collected during this research reinforce the need and possibilities of the public managers to use data driven information to provide ways and public policies that promote social and climate justice., it could not evidence, as affirmed by the literature, that the cities infrastructure has been captured by the highest income social classes, which may be the result of the use of aggregated city data or the chosen variables.

The research also emphasizes the potential of usage of coding and AI to fast implement tables and methods of analysis to support decision making into the public management.

Another important contribution of this research is to provide a compilation of techniques that can be used to the understanding of urban dynamics by municipal managers, urban planners, academics, and society.

A suggested pathway for new works at this field is to improve the database with data about: 1. Economic variables (cities: PIB, income, rate of employment *etc*); 2. urban equipment (location and quality of: i. parks & recreation; ii, educational; iii. healthcare; iv. social); and, 3. sustainability data (flooding, heat islands, social vulnerability *etc*), joining all those data to implement a geocomputation system that would allow to interpret the spatialized data and the evolution of data throughout the time.

References

[1] T. Strickland, "The financialisation of urban development: Tax Increment Financing in Newcastle upon Tyne," *Local Econ.*, vol. 28, no. 4, pp. 384–398, Jun. 2013, doi: 10.1177/0269094213475857.

[2] M. Serageldin, D. Jones, F. Vigier, and E. Solloso, "Municipal financing and urban development". Nairobi: United Nations Human Settlements Programme, 2008.

[3] R. Rolnik and J. Klink, "Crescimento econômico e desenvolvimento urbano: Por que nossas cidades continuam tão precárias?" *Novos Estudos CEBRAP*, no. 89, pp. 89–109, 2011, doi: 10.1590/S0101-33002011000100006.

[4] L. A. Salinas and M. Janoschka, "The Biopolitics of Housing Financialization: An Approach to the Epistemic Violence of Financial Capitalism," *Revista de Estudios Sociales*, vol. 2023, no. 85, pp. 59–75, 2023, doi: 10.7440/res85.2023.04.

[5] C. V. Diezmartínez and A. G. Short Gianotti, "US cities increasingly integrate justice into climate planning and create policy tools for climate justice," *Nat. Commun.*, vol. 13, no. 1, Dec. 2022, doi: 10.1038/s41467-022-33392-9.

[6] M. Fobi Kontor, A. Brown, and J. R. Núñez Collado, "Climate Justice and Heat Inequity in Poor Urban Communities: The Lens of Transitional Justice, Green Climate Gentrification, and Adaptation Praxis," Jun. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/urbansci9060226.

[7] L. Porter *et al.*, "Climate Justice in a Climate Changed World," *Planning Theory & Practice*, vol. 21, no. 2, pp. 293–321, Mar. 2020, doi: 10.1080/14649357.2020.1748959.

[8] F. M. Sousa *et al.*, "Justiça climática urbana: uma revisão sistemática das iniquidades na distribuição de soluções baseadas na natureza," in *Contribuições Multidisciplinares Para o Conhecimento Atual*, Aurum Editora Ltda, 2025, pp. 689–702. doi: 10.63330/aurumpub.012-054.

[9] G. Esposito, A. Terlizzi, M. Guarino, and N. Crutzen, "Interpreting digital governance at the municipal level: Evidence from smart city projects in Belgium," *International Review of Administrative Sciences*, vol. 90, no. 2, pp. 301–317, Jun. 2024, doi: 10.1177/00208523231167538.

[10] T. Yigitcanlar, R. Mehmood, and J. M. Corchado, "Green artificial intelligence: towards an efficient, sustainable and equitable technology for smart cities and futures," *Sustainability (Switzerland)*, vol. 13, no. 16, Aug. 2021, doi: 10.3390/su13168952.

[11] E. G. R. L. Domingues, "Mudanças climáticas e planejamento urbano:," *Revista Brasileira de Direito Urbanístico | RBDU*, vol. 11, no. 20, pp. 511–536, Sep. 2025, doi: 10.55663/RBDU.v11.i20.ART19.RJ.

[12] C. T. do C. D. IBGE, "Censo Demográfico 2022 : população e domicílios : primeiros resultados". Rio de Janeiro: IBGE, 2023.

[13] L. Paulo. Fávero and P. Prado. Belfiore, *Data science for business and decision making*. Academic Press, an imprint of Elsevier, 2019.

[14] J. T. Vanderplas, "Python data science handbook: essential tools for working with data". O'Reilly Media, Incorporated, 2023.