Analysing Urban Sprawl Using Machine Learning: A Study of Southern Santa Catarina

Sofia Knabben Niveiros [a].

[a] Faculty of Architecture and Urbanism, College of Southern Region, State University of Santa Catarina, Laguna, Brazil, sofianiveiros@gmail.com.

Abstract. This article is part of a broader research project aiming to develop automated models to analyze urban growth in the cities of the southern region of Santa Catarina, focusing on the phenomenon of urban sprawl. Urban sprawl is characterized by the unregulated and horizontal expansion of cities, leading to infrastructure problems and real estate speculation, which displaces low-income populations to the suburbs. The research uses a literature review to identify challenges and methodologies in studies that applied the Random Forest (RF) algorithm to urban sprawl analysis. Sixty-five articles were analyzed, demonstrating that although RF is widely used for its simplicity and robustness, it has limitations in complex urban scenarios, particularly in classifying rapidly growing areas and in data quality. The article also compares RF's performance with other machine learning algorithms, such as Artificial Neural Networks (ANN), Support Vector Machine (SVM), U-Net, and k-Nearest Neighbors (k-NN), highlighting the need for mixed approaches to overcome RF's limitations. Finally, the study concludes that, in the context of the southern region of Santa Catarina, integrating different algorithms can be an effective strategy to improve urban sprawl analyses, provided that the practical resources and limitations are considered, thus contributing to more efficient and sustainable urban planning.

Keywords. Random Forest, Urban Sprawl, Machine Learning, Southern Brazil.

1. Introduction

This article is part of a larger research project whose goal is to develop automated models to analyze urban growth patterns in cities of the southern region of Santa Catarina. The phenomenon known as urban sprawl is characterized by the uncontrolled, low-density, horizontal growth of a city, often accompanied by infrastructure problems. Real estate speculation contributes to this process, making urban centers more expensive and forcing low-income populations to move to the suburbs. Consequently, these areas need to be regularized later with urban planning measures, which end up costing substantial public funds.

Given the relevance of this issue, it becomes essential to understand and anticipate the direction of such growth, enabling more efficient and sustainable urban planning. For this purpose, a bibliographic review was conducted, allowing comparative analyses of the challenges and difficulties faced by researchers using different algorithms. Thus, this article aims to examine the methodologies employed, contributing to the construction of more equitable and sustainable cities.

2. Methodology

To understand how other researchers conducted similar studies, a bibliographic review was carried out. This review was conducted by the research team, which chose the Google Scholar and ScienceDirect databases due to their free access to articles. After testing search terms, the keywords “urban sprawl," "machine learning," and " espraiamento urbano " were defined, which aligned more closely with the study’s objectives. The time frame between 2015 and 2024 was established to ensure the recency of the studies. Additionally, in Google Scholar, the “include citations” option was activated, while ScienceDirect was configured for “all open access.”

The search resulted in 82 articles from Google Scholar and 193 from ScienceDirect. These articles were then equally divided among the researchers, who cataloged information such as title, authors, year of publication, territorial scope, methodology, typology, and a summary of each article. Based on relevance, the articles were classified, resulting in 65 selected articles..

This new list was then distributed for a more detailed reading, focusing on objectives, data sources, databases, machine learning algorithms used, and the methodology employed to create the models. After this second review, it was identified that RF was predominantly used over other algorithms, directing the research toward this method [1]. However, as the research progressed, the need emerged to deepen the understanding of RF's functioning, particularly the challenges faced by other researchers and possible solutions, which defines this article's objective.

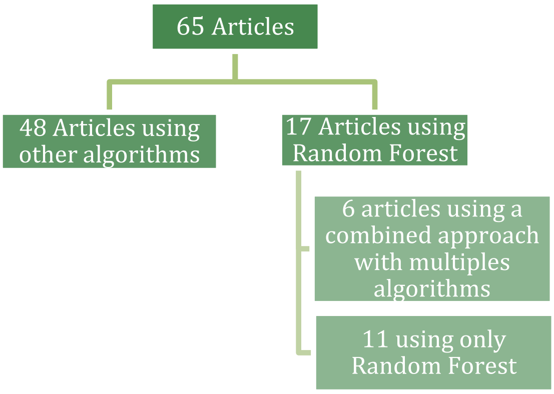
65 Articles
48 Articles using other algorithms
17 Articles using Random Forest
6 articles using a combined approach with multiples algorithms
11 using only Random Forest

Fig. 1 – Algorithm Distribution in Reviewed Articles.

2.1 Random Forest

Random Forest is a supervised machine-learning algorithm developed by Breiman in 2001 [2], known for being simple and robust. According to Biau et al. (2016), RF operates based on the "divide and conquer" principle: data samples are randomly selected, with a tree predictor developed for each fraction, which is then aggregated to form a consolidated prediction [3]. This method stands out for applying to a wide range of predictive problems with few parameters to adjust, and it performs better with fewer samples compared to other algorithms. To achieve this, all articles using RF as a methodology were carefully read, noting the difficulties encountered by the researchers, and, in cases where auxiliary methodologies were used, how these methodologies contributed to the research results.

3. Results

After analyzing the reviewed articles, categories were created to classify the articles based on the main types of challenges faced by researchers regarding using the RF algorithm.

3.1 Classification Problems

Errors in correctly classifying data applied to more complex areas were observed in almost all studies [4-17]. These failures occurred for various reasons; however, one of the most mentioned across different articles was RF's difficulty in classifying complex urban networks, such as heterogeneous urban fabrics and rapidly growing areas. These challenges occur especially in scenarios where land use changes rapidly, resulting in low classification accuracy, as seen in the study [4], where the low resolution of satellite images, combined with urban complexity, resulted in significant errors.

3.2 Data Problems

The challenges regarding data were classified into two categories to facilitate understanding: the lack of data and the quality of data. The relationship between these problems is close, with minor variations between articles. The quality of the images was identified as a challenge in 11 studies [4, 6, 8-12, 14-17], while the insufficient quantity of data was cited in 9 articles [4-8, 10, 12, 16, 18]. Studies that collected satellite images with resolutions ranging from 10 to 30 meters, such as Sentinel-2 or Landsat, experienced greater difficulties in classification, associated with lower image quality. In terms of data scarcity, these issues appeared to have a greater impact on studies in regions with low international visibility [8].

3.3 Construction Problems

Errors related to the structuring of the RF algorithm, such as the presence of imbalanced data or the inclusion of irrelevant variables, were cited in six studies. These problems often arose from previously discussed issues, such as data and classification problems. Additionally, spatial autocorrelation in densely built areas was identified as a significant limitation of RF, resulting in inaccurate predictions and compromising the quality of the results.

3.4 Others

The "Other" category groups specific problems identified individually in some studies that do not fit into any of the previous categories. These challenges are characterized by their unique nature and are specific to each study, suggesting a lower likelihood of recurrence in different contexts.

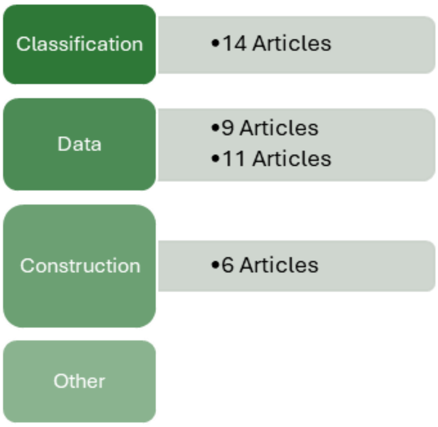

Fig. 2 – Distribution of Articles by Topic.

Additionally, an analysis was conducted on the articles that used other machine learning algorithms besides RF, identifying how these alternative methodologies contributed to the research. The types used included ANN, SVM, U-Net, and k-NN. Each of these algorithms faced unique challenges, but in many cases, they helped mate RF errors, especially when applied to complex urban scenarios.

3.5 Challenges and Solutions of Other Algorithms

Artificial Neural Networks (ANN): ANN demonstrated a greater ability to handle complex data nonlinearities, particularly in areas with abrupt transitions between land-use classes where RF frequently failed. However, neural networks require large volumes of data and significant computational resources, making their use more challenging in studies with data or processing limitations. Moreover, ANN is more prone to overfitting, which may limit its generalization capability [8].

Support Vector Machine (SVM): SVM is effective for classifying data with well-defined margins, making it useful in contexts where transitions between different land-use classes are clear. However, SVM struggled in high-dimensional scenarios and class overlap, making it less effective in dense urban environments. In urban sprawl studies, RF handled accuracy better, especially when there was greater complexity in transitions between impermeable surfaces and exposed soils [13].

U-Net: The integration of RF with U-Net, a deep neural network, proved particularly efficient in improving the classification of densely built urban areas. U-Net was highly efficient in segmenting built surfaces where RF alone had lower accuracy. However, U-Net requires a much larger volume of training data, which becomes a limitation in studies that rely on low-resolution images or cannot obtain many representative samples [4].

k-Nearest Neighbors (k-NN): k-NN showed greater accuracy than RF in scenarios requiring the updating of urban densities, particularly when predicting the density of urban points of interest. However, k-NN is more sensitive to areas with limited or less correlated data, which can affect its performance in urban sprawl studies in regions with inconsistent or sparse data [5].

Overcoming the Limitations of Random Forest: The challenges faced by RF, such as its limitations in capturing spatial relationships and handling highly autocorrelated data, were largely mitigated by these other algorithms. Deep neural networks, such as U-Net, improved the extraction of built surfaces, while k-NN performed better in regions with more dispersed data. Although SVM was useful in classifications with well-defined margins, it had less applicability in urban transition scenarios. These combined approaches suggest that integrating alternative methods with RF can improve urban sprawl analysis, offering greater accuracy and flexibility in capturing the complex dynamics of land use.

4. Discussion

As previously mentioned, RF encountered difficulties in dealing with complex urban areas, particularly where land use changed abruptly and non-linearly [4-6, 13]. This issue was exacerbated using low-resolution images, such as those from Sentinel-2 or Landsat (10 to 30 meters), which did not allow for the detection of small buildings or the distinction between impermeable surfaces and exposed soils, resulting in omission errors [4, 14-16]. Collecting representative samples also posed challenges, particularly in regions with limited data coverage. Spatial autocorrelation in densely built areas generated inaccurate predictions. The use of auxiliary methodologies, such as neural networks and SVM, proved beneficial [13], opening the possibility of adopting a mixed approach. However, the inherent challenges of these other algorithms may outweigh the benefits in the current research.

In the case of this study applied to the southern region of Santa Catarina, issues related to data quality and availability are particularly relevant, as the studied cities are small and medium-sized and lack a robust data collection infrastructure compared to large metropolitan centers. A smaller data volume may force the study to rely on lower-resolution images, affecting analysis precision, especially in areas with more diversified land use. On the other hand, the slower growth of these cities may mitigate data scarcity issues between periods of rapid urban growth, as observed in [4].

Moreover, the research team's inexperience with advanced machine learning algorithms like RF may introduce challenges during the model construction phase, requiring rigorous testing to adjust parameters and avoid overfitting. Although the literature indicated the possibility of mitigating these problems through combined approaches [4, 8, 13], some of these algorithms are even more challenging and require greater expertise to implement, necessitating an assessment of their feasibility in this case.

Finally, adopting a mixed approach using RF in conjunction with more flexible algorithms should be evaluated based on available resources and the precision needed for the analyses. The complementarity between these techniques can improve the accuracy of the results but must be applied in a way that respects the practical limitations of the study.

5. Conclusion

The literature review allowed for the identification and understanding of the main challenges faced by the RF algorithm and other auxiliary methodologies in land-use analysis. Although RF is widely used due to its simplicity and effectiveness, the study demonstrated that it faces significant limitations in complex urban scenarios, particularly about data classification and the quality of available information. The comparison with other algorithms such as ANN, SVM, U-Net, and k-NN showed that they can overcome some of these limitations but also present their challenges, such as the need for large data volumes and greater computational capacity.

Applying this knowledge to the southern region of Santa Catarina, the research suggests that adopting mixed approaches combining RF with other algorithms can enhance the effectiveness of urban sprawl analyses. Such an integrated approach can contribute to more adaptive and effective analyses, especially in regions with limited data.

For future studies, it is recommended to explore deep learning algorithms such as convolutional and generative adversarial networks, as well as integrating multisensory data that can increase the precision of urban analyses. Additionally, the development of machine learning-based tools and their practical application by urban planners represent a promising path for the formulation of more effective and sustainable public policies.

6. Acknowledgement

I would like to thank, first and foremost, the INCBAC Institute for creating this course and providing such a unique learning experience, especially to the professors who dedicated their time to teach and offer the best experience possible. I would also like to thank my research advisor, Professor Fernando dos Santos Calvetti, and my fellow research scholars Maria Beatriz Souza de Oliveira, Júlia Stedile, Sofia Furtado, and Maria Julia dos Santos Yamada. Finally, I am grateful to my mother, father, and boyfriend for all the help and support they gave me throughout this process, which was one of my first experiences with academic writing and my first time writing in a foreign language.

7. References

[1] Stedile J, Souza MB, Furtado S, Niveiros S. Aplicação de machine learning no estudo do crescimento urbano desordenado em Santa Catarina. [in press].

[2] Breiman L. Random Forests. Vol. 45. 2001.

[3] Biau G, Scornet E. A random forest guided tour. Test. 2016 Jun 1;25(2):197–227.

[4] Mugiraneza T, Hafner S, Haas J, Ban Y. Monitoring urbanization and environmental impact in Kigali, Rwanda using Sentinel-2 MSI data and ecosystem service bundles. *Journal of Environmental Management*. 2022; 1(2): 101-120.

[5] Haas J, Ban Y. An extended patch-based cellular automaton to simulate horizontal and vertical urban growth under the shared socioeconomic pathways. *Computers, Environment and Urban Systems*. 2022; 89: 134-150.

[6] Chen Z, Zhou Y, Peng S, Zhang Y, Xiao Z. Toward accurate mapping of 30-m time-series global impervious surface area (GISA). *Remote Sensing of Environment*. 2023; 259: 112357.

[7] Wood J, Barr S, Jones P. A machine learning methodology to quantify the potential of urban densification in the Oxford-Cambridge Arc, United Kingdom. *Land Use Policy*. 2023; 127: 106456.

[8] Khedr M, Kumar L. Land use land cover change detection and urban sprawl prediction for Kuwait metropolitan region, using multi-layer perceptron neural networks (MLPNN). *Journal of Urban Planning and Development*. 2023; 149(1): 05022005.

[9] Johnson B, Carver S, Birch C. Landscape metrics regularly outperform other traditionally-used ancillary datasets in dasymetric mapping of population. *Applied Geography*. 2022; 146: 102669.

[10] Rahman M, Ghani H, Sultan A, Islam M. Exploring the nexus between land cover change dynamics and spatial heterogeneity of demographic trajectories in rapidly growing ecosystems of south Asian cities. *Environmental Research*. 2024; 215: 114312.

[11] Pereira S, Silva T, Oliveira M. Mapeamento pretérito e prognóstico da expansão urbana de Montes Claros/MG usando machine learning. *Revista Brasileira de Cartografia*. 2021; 73(2): 415-432.

[12] Yang Y, Feng Z. Multi-scenario simulation of urban growth boundaries with an ESP-FLUS model: A case study of the Min Delta region, China. *Land Use Policy*. 2023; 126: 106345.

[13] Ghosh S, Debnath P, Kumar R, Chatterjee A. Comparison of accuracy and reliability of random forest, support vector machine, artificial neural network and maximum likelihood method in land use/cover classification of urban setting. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2022; 187: 194-210.

[14] Alam M, Afrin T, Uddin M. Mapping and monitoring land use land cover dynamics employing Google Earth Engine and machine learning algorithms on Chattogram, Bangladesh. *Remote Sensing*. 2023; 15(5): 1156.

[15] Phung T, Wang J, Li Z. Characterizing urban growth in Vientiane from 2000 to 2019 using time-series optical and SAR-based estimates of urban land. *International Journal of Applied Earth Observation and Geoinformation*. 2023; 117: 103058.

[16] Mustafa A, Akhter R. Synergizing Google Earth Engine and Earth Observations for Potential Impact of Land Use/Land Cover on Air Quality. *Environmental Research Letters*. 2024; 19(3): 034015.

[17] Bari M, Islam N, Mahmud R. Spatial indices and SDG indicator-based urban environmental change detection of the major cities in Bangladesh. *Sustainable Cities and Society*. 2022; 85: 104035.

[18] Bashir M, Liu Y, Fan S. Predicting intra-urban well-being from space with nonlinear machine learning. *Computers, Environment and Urban Systems*. 2024; 97: 101900.