RESEARCH ARTICLE

DATASET REPORTS

Detection of slums in Rio de Janeiro through satellite images

Detecção de favelas no Rio de Janeiro por meio de imagens de satélite

Hanna Diniz Cunha [a*], Andrea Diniz da Silva [b], Bernardo Braga Martins [b], Bruno Sá Guedes [b], Ian Monteiro Nunes [c], Marcelo Rodrigues de Albuquerque Maranhão [d], Miguel do Nascimento Faria Conforto [b]

[a] UN Regional Hub for Big Data in Brazil, 20231-050, Rio de Janeiro, RJ, Brazil.
[b] Escola Nacional de Ciências Estatísticas - ENCE/IBGE, 20231-050, Rio de Janeiro, RJ, Brazil.
[c] Diretoria de Pesquisas do Instituto Brasileiro de Geografia e Estatística – IBGE, 20031170, Rio de Janeiro, RJ, Brazil.
[d] Diretoria de Geociências do Instituto Brasileiro de Geografia e Estatística – IBGE, 20031170, Rio de Janeiro, RJ, Brazil.

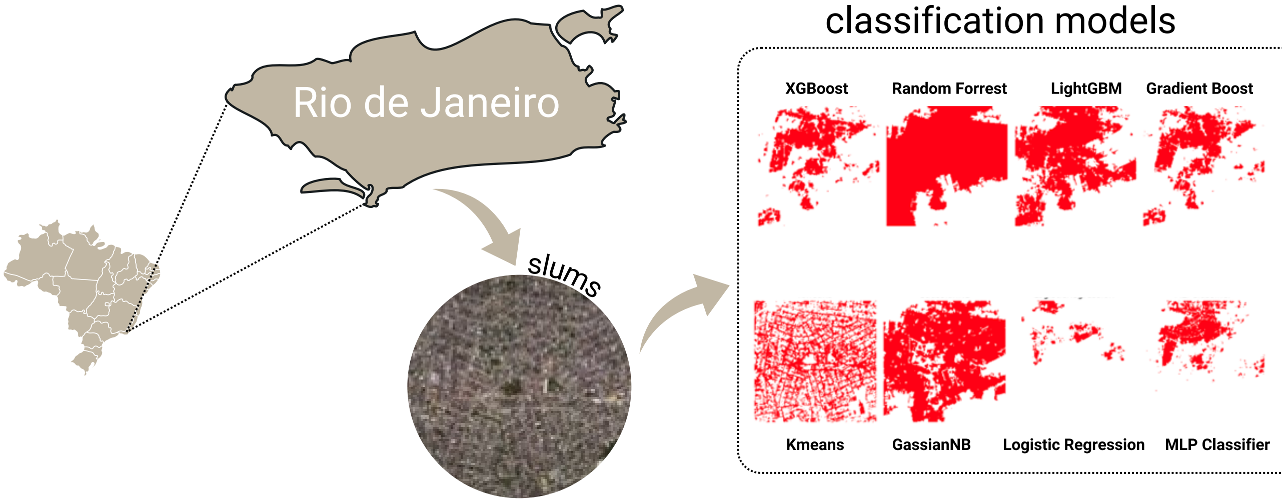
Abstract
According to UN-Habitat, more than one billion people live in informal settlements worldwide, of which 200 million living in Africa and another 100 million in Latin America, mainly in countries such as Brazil, Mexico, Colombia, Peru, and Argentina. Rio de Janeiro has 1,074 favelas, representing 22% of the city's total population, making it the Brazilian municipality with the highest percentage of people living in favelas. Ensuring human rights through access to basic services for the populations living in these settlements, through programs and public policies, depends on timely and reliable data. However, despite spending decades establishing their national statistical systems, usually based on data collection directly from individuals, in most countries, the data produced in traditional ways does not portray the dynamics of these populations promptly. As an alternative, we combined free satellite imagery with machine learning and deep learning to identify the area occupied by favelas in the city of Rio de Janeiro. We compared the results of eight distinct segmentation models using the IoU and F1 as metrics. Among the evaluated methods, two stood out for their performance: GradientBoost and XGBoost.

Keywords: Informal settlements. Human rights. SDG 11. Satellite imagery. Open access data. Machine learning. Supervised learning.

Resumo
Segundo a ONU-Habitat, mais de um bilhão de pessoas vivem em assentamentos informais no mundo, das quais 200 milhões estão localizadas na África e outras 100 milhões na América Latina, principalmente em países como Brasil, México, Colômbia, Peru e Argentina. Na cidade do Rio de Janeiro, há 1.074 favelas, onde vivem 22% da população, tornando-se o município brasileiro com o maior percentual de pessoas vivendo em favelas. A garantia dos direitos humanos na forma de acesso a serviços essenciais das populações vivendo nesses assentamentos, por meio de programas e políticas públicas, depende de dados oportunos e confiáveis. No entanto, apesar de passarem décadas estabelecendo seus sistemas estatísticos nacionais, geralmente baseados em coleta de dados diretamente com indivíduos, na maioria dos países, os dados produzidos de forma tradicional não retratam de forma oportuna a dinâmica dessas populações. Como alternativa, usamos imagens de satélite gratuitas em combinação com aprendizado de máquina e aprendizado profundo para identificar a expansão ou retração de favelas na cidade do Rio de Janeiro. Utilizando as métrica IoU e F1, foram comparados oito modelos de classificação, dentre os quais dois se destacaram por seu desempenho: GradientBoost e XGBoost.

Palavras-chave: Favelas. Direitos humanos. ODS 11. Imagens de satélite. Dados abertos. Aprendizado de máquina. Aprendizado supervisionado.

Graphical Abstract



*Corresponding author: Hanna D. Cunha. E-mail address: hanna.cunha2008@hotmail.com

Submitted: 16 September 2024; Accepted: 18 September 2024; Published: 19 September 2024.

1. Introduction

According to UN-Habitat (2021), more than one billion people live in informal settlements worldwide, of which 200 million live in Africa and another 100 million in Latin America, mainly in Brazil, Mexico, Colombia, Peru, and Argentina. In the city of Rio de Janeiro, there are 1.074 slums where 22% of the population live. It is the Brazilian municipality with the highest percentage of people living in *favelas* (UN-Habitat, 2023). Given that the vast majority of social programs rely on data to create public policy and allocate funding for providing essential services, delivering human rights to those who most need it is an effective manner. As such, ensuring human rights in the form of access to basic services for the populations living in these settlements, through programs and public policies, depends on timely and reliable data. However, despite spending decades establishing their national statistical systems, usually based on data collection directly from individuals, in most countries, the data produced in traditional ways does not portray the dynamics of those populations promptly.

While censuses are the traditional source of data gathering on the way a population lives, having been performed for centuries, they are very hard to apply by any government, or under any circumstance. In particular, they still face many challenges, the main ones being the time and cost to conduct them, the hesitancy of respondents to cooperate, and the rapid obsolescence of the data. These problems are much greater when surveying informal communities, where people tend to have strong concerns about privacy and the intrusiveness of public officials, developed as a means of self-defense. Added to these issues are the physical difficulty of accessing many of these areas (especially those built on hillsides) and the presence of organized gangs suspicious of intruders. Traditional censuses depend on honest cooperation and access to the people, without which the quality of the data will be questionable.

The first modern record of a census can be attributed to the province of Quebec, Canada, known at the time as La Nouvelle France. While Europeans began registering citizenship in Wurtemburg in 1622, the first known census is attributed to the Babylonians in roughly 3800 BC (ABS, 2006). Similar to its modern counterpart, the Babylonian census was conducted every six to seven years, and it served to tally the number of people and livestock along with the quantities of butter, honey, milk, wool, and vegetables produced, as a measure of wealth at the time (Ibid). In 1958, at the request of the United Nations Statistical Commission, the United Nations released the first set of principles and recommendations for population and housing censuses (UN, 2017). The document describes censuses as vital sources of data for policymaking and planning, including boundary delimitation for administrative and research purposes, as well as having countless uses by businesses and labor organizations. In particular, the data are used for the development of benchmark housing statistics, assessment of housing quality, and formulation of housing policies and programs.

While traditional censuses can be extremely informative and are still the main source of official data, they have shortcomings, such as the dependence on the availability of resources. In addition, given that the census is conducted usually every 10 years, the data gathered by it does become obsolete the farther it is from the reference date. This issue is particularly heightened for policies that deal with highly volatile and fast-shifting situations and rely on the most up-to-date information to allocate funding for services. To circumvent the issue of cooperation and obsolescence, and access and try to bridge the large data gap, many countries have resorted to other types of gathering and analyzing data on various topics, including informal settlements.

According to the Big Data Project Inventory, compiled by the United Nations Global Working Group on Big Data, “34 [National Statistical Systems (NSS)] from around the world have registered 109 separate big data projects” (MacFeely, 2019). According to the data published by the group in 2018, six national statistics offices (NSOs) and other national agencies had reported using satellite imagery nationally, while seven had reported doing so internationally. Furthermore, a total of 14 projects had “Population/migration” as their focus, both nationally and internationally, while another 15 of the projects registered were “Geographical/spatial” in nature, again both nationally and internationally. These projects have very different goals and methods, and some use satellite imagery while others rely on mobile phone data or web scraping. Thus, it is clear that many countries have been using big data to understand how their population relates to land use.

In Latin America and the Caribbean, the use of big data is also a reality. According to the International Consultation on the Use of Big Data for Statistics in Latin America and the Caribbean, conducted annually by the United Nations Regional Hub for Big Data in Brazil, 14 countries in Latin America and the Caribbean reported using big data to produce either experimental or official statistics in the past two years (Hub, 2024; Silva et al., 2023). While each of these countries uses different types of big data for different purposes, a major use of all the respondents of the consultation is satellite imagery, with a total of seven instances reported of use for official statistics and eight for experimental statistics. The use of big data, and specifically satellite imagery, to produce statistics, both official and experimental, is still a novel development; Here we contribute by showcasing the use of satellite imagery in combination with machine learning to identify slums.

2. Related Work

As technologies advance, new methods become available and new ways to conduct analyses become possible. With the advancement of very high resolution (VHR) satellite imagery and computational efficiency, Graesser et al. (2012) were able to rethink existing theories of how to classify settlements to the still very convoluted territory of urban landscapes. According to them, “There are many existing methods to map urban structures and/or objects, but they all require information extracted from an image to work efficiently.” For them, image feature extraction is an essential step in the classification of settlements. Nonetheless, the urban landscape is extremely complex and requires a feature, or set of features, equally complex to even attempt to understand and process it, let alone classify it. Finally, with the advance of computational technology and processing, the authors were able to attempt a “thorough evaluation of them in a systematic manner for urban classification.”

Subsequently, Duque et al. (2017) explored the potential of machine learning to identify informal settlements with the use of VHR imagery. While previous studies have demonstrated that the physical characteristics of informal settlements are distinguishable from those of regular settlements, classification still requires human interaction. To overcome the need for humans and to classify larger areas, the authors switched from object-based image analysis (OBIA), previously the most commonly used method, although not the only one, to a machine learning (ML) approach. This approach is unsupervised, which means that after the initial supervised training, it can conduct the classification largely by itself, which enables the coverage of larger areas.

Another way of overcoming the limitations of OBIA, as shown by Prabhu and Raja (2018), is to use it in combination with object-oriented classification (OOBIA), which performs much better in "selecting the parameters such as scaling and merging for segmentation of the images." However, this method also has some limitations, which the authors overcame with a gray-level co-occurrence matrix (GLCM)-based feature extraction technique, which can better process texture and better identify rooflines, building outlines, and urban structures. However, while texture can help differentiate between regular and informal settlements, the approach ultimately fails regarding scale. This leads to the use of the space-frequency analytical discrete-wavelet transform (DWT) tool, which "can effectively characterize the images at different scales." Lastly, the final step of the detection approach is the application of a discrete wavelet frame transform (DWFT), which is responsible for overcoming the loss of information related to the down-sampling of frequency elements that occurs during DWT.

Another limitation of the use of satellite images to identify and classify settlements is spatial heterogeneity. According to Wang et al. (2019), "the urban landscape as a complex geographic system is composed of hierarchical patterns and discrete objects in a spatial and temporal continuum with different scales and anisotropy." However, while the use of VHR imagery has been the best to support comprehensive mapping and monitoring of the spatial extent and physical characteristics of geographic locations and settlements, it still has limited processing capabilities with regard to characteristics of scale and anisotropy (Prabhu & Alagu Raja, 2018; Wang et al., 2019). To address this issue, Wang et al. (2019) attempted to fill this gap by "analyzing the impact of scales and anisotropy detected in the scale space and frequency domain for the calculation of texture indices that ultimately govern the detection of slums."

Still on the subject of heterogeneity and morphology, in the case of processing extremely large and/or variable areas, an important obstacle is the highly variable morphology of poverty (Stark et al., 2020). According to those authors, "Mapping these settlements is not a trivial task" and there are certain challenges of variability that need to be addressed. The first challenge is that of interurban variability, "where morphological slum features can change depending on their particular geographical location." In reality, there is no international consensus on the definition of informal settlements, let alone on their morphological features and characteristics. According to the authors, the "morphologic appearances of poverty can be different in every city, ranging from very dense low-rise shacks in Mumbai to three-story buildings in Medellin." The second challenge, or the second aspect of the variability challenge, is that of intraurban variability, where there are large differences between informal settlements even within the same city. "Although the slum areas in Lagos are located within the same city, their morphological appearance is inherently different."

The most recent technological advancement is that of deep learning. While both machine learning and deep learning are subsets of the field of artificial intelligence (AI), machine learning still needs large amounts of human interaction, since it often requires the manual identification of features and classifiers to properly adjust the algorithm, whereas deep learning can "learn from its own errors" (Google Cloud). According to Persello and Kuffer (2020), "The recent introduction of deep learning techniques, such as CNNs and fully convolutional networks (FCNs), has shown great potential for automatically learning the spatial, textural and morphological characteristics of deprived areas and to produce accurate classification maps within an end-to-end learning framework." This approach not only "overcomes the limitations of previous deprivation indices, which relied on weighted indicators that are sensitive to the choice of individual weights," it is largely an unmanned, automated approach. Because of the ability to learn from its own mistakes, the use of deep learning enables the coverage of large amounts of land.

There have also been many experimental studies testing previous ideas. Expanding on the idea of using satellite imagery as an alternative data source for monitoring urban areas, Assarkhaniki et al. (2021) used it for the remote sensing of informal settlements in Jakarta, the capital of Indonesia. The authors used open-source data to paint a medium-quality picture of the settlements in the area. Next, when it came to the method of classification between formal and informal settlements, they relied on pixel-based classification (PBC), with the use of machine learning techniques, to run a combination of classifiers, such as KNN, neural networks, and random forests. Finally, the authors divided training by first mainly using already recognized informal settlements, as described by the World Bank in 2015, and second overlaying OpenStreetMap (OSM) data for roads or built-up areas. While the results of the first step left little to be desired, it used data that at the time were already six years old, so the addition of the OSM data significantly enhanced the final results, increasing the precision range from 0.58-0.93 to 0.79-0.97, and the accuracy range from 0.88-0.97 to 0.93-0.99.

Another study conducted not long thereafter was that of Alrasheedi et al. (2023) in Riyadh City, Saudi Arabia. Unlike the previous study, the authors used "WorldView-3 panchromatic and multispectral images with spatial resolutions of 0.31m and 1.24 m, respectively" for this study, which were VHR images obtained for King Abdulaziz City for Science and Technology (KACST), a government organization located in Riyadh. Next, they used the local community to create a list of the best possible indicators by application of a survey. They then ran this list of possible criteria through various analytical hierarchy processes to select the priority criteria of each possible indicator to select the ones that best described "'informal settlements' or 'old residential and historical neighborhoods'". Lastly, the estimation of scale parameter (ESP) was used to transform "multi-resolution image segmentation into informal settlements, formal settlements, road networks, shadows, vacant areas, and vegetation," and then performed object-based image analysis. While this approach was not very successful for all classifications (e.g., the vegetation accuracy decreased from 0.97 to 0.95), the accuracy of formal and informal settlements and roads, went from 0.59 to 0.98, 0.53. to 0.93, and 0.67 to 0.83 respectively, a resounding success.

A research by Oliveira et al. (2023) also utilized the K-means classifier in São Paulo, Brazil. The authors used multiple models to uncover the most significant features of interest, to determine which public policies are better suited to promote the envisioned change in each location. Much like us, they began by identifying the relevant features using GIS software. However, unlike our study, for the second model they used remote sensing data to "extract spatial, spectral and textural features from satellite imagery." Lastly, to validate their results they used the information provided by the Brazilian Institute of Geography and Statistics (IBGE) on informal settlements. Concerning the selection of classifiers, the authors developed a model that could handle "continuous numerical data such as the spatial features here used," as well as "[perform] well for high volume datasets with small processing time for regular machine capacities." Hence, they chose the unsupervised classifier Kmeans. The authors ended up with four distinct cluster types "with different deprivation aspects, such as higher and lower accessibility to services and infrastructure, sparser and denser occupation; regular and complex morphology; flat and steep terrain."

The final study covered in this section is by Cinnamon and Noth (2023), which is by far the largest and most complex, spanning 20 years. The research focused on Cape Town, South Africa, and the authors monitored the spatiotemporal development of the informal settlements in the city from 2000 to 2020, identifying their locations at the beginning of the study, and tracking "their growth and decline over the 20-year period", until 2020. As the second most populous city in South Africa, and with a history of apartheid until 1994, Cape Town served as the perfect setting for a long-term study of informal settlements and how they grow and change over time. The authors relied on open-source data gathered from local and international web-based data repositories. Additionally, the Cape Town city government itself has an open portal where it provides access to various "key datasets", which "[contain] a variety of spatial datasets, however, no informal settlement location datasets are available." Several of these datasets from the city's open portal were used to identify the locations of informal settlements, including "a dataset on city wards and building parcels," which included "building footprints and information for all formal land tenure throughout the city."

For their study, two types of image data were used, namely "high resolution (8 cm) aerial photographs from 2020", and "multi-temporal medium resolution (30 m) satellite imagery from the Landsat 5 and 8 satellites." To identify informal settlements, they opted for a "combined visual image interpretation and ancillary data overlay approach." Using 2020 high-resolution photos of the city, the authors conducted an "informal land use detection analysis," visually and manually classifying what they believed to be the complex and organic morphology of informal settlements within the limits of Cape Town. Next, they overlaid the formal residential parcel dataset on the aerial photos to better differentiate the irregular from the regular settlements. Last, using the matrix dataset from the OpenUp data portal, they populated their settlement representation, ending up with a total of 254 informal settlement areas of interest throughout the city, which were then monitored over the span of 20 years.

To track the development of these informal settlements, an object-based approach was adopted to monitor and record the "change in built-up vs. non-built-up land." Using this supervised classification model, the authors produced new layers of land classification for every 5-year interval (2000, 2005, 2010, 2015, and 2020), which were then patched together into one. Utilizing the QuantumGis application, they combined two separate raster layers into one output layer through the use of the r. cross-function, which created new raster values "for each unique combination of values from the input raster layers." With the use of this function, four unique raster values were "assigned to each type of land cover change between the 5-year intervals and between the overall 2000 to 2020 study period." In the end, they were able to accomplish a "settlement detection analysis [that] identified that informal settlement areas comprised 1720 ha of Cape Town in 2020, representing 0.7% of the city (246,100 ha)." Furthermore, they found that during 2020, 4.2% of all Cape Town residents (over 190 thousand people) were living in informal settlements.

These examples illustrate the successful application of satellite imagery in analyzing informal settlements. A quick search on Google Scholar reveals nearly 86,000 studies on this topic. Whether through large-scale projects or smaller studies, we aim to further demonstrate the promising potential of this approach.

3. Methodology

Rio de Janeiro, also known as the 'Marvelous City', has a coastline that extends 246 km. Besides the many beaches and other stunning coastal features, the city boasts one of the largest urban forests in the world, along with mountains that reach a peak of 1025 meters above sea level. Rio de Janeiro thus has a skyline filled with the highs of mountains such as the famous Sugar Loaf, and Corcovado (site of the Christ the Redeemer statue), alongside the beautiful low of its beaches, lagoons, and bays, such as Copacabana Beach, Guanabara Bay, and Rodrigo de Freitas Lagoon (Guitarrara, n.d.). Located in the Tropic of Capricorn, it has a pleasant average temperature of about 24 degrees Celsius. Rio de Janeiro is hence not only a marvelous place to visit, but it is also a perfect setting to study the different ways people form communities given its diverse geography and socioeconomic inequalities. Therefore, we decided to use the beautiful city of Rio de Janeiro as the setting for this study.

The study began by downloading a satellite image of Rio de Janeiro. Since it can be hard for most classifiers to process and differentiate data from images with a resolution lower than 10 m (Gram-Hansen et al., 2019), we used a high-quality open-source 50-centimeter resolution satellite image from Google covering the entire city of Rio de Janeiro. However, the processing power required to process the downloaded dataset was out of our research group's available computational power. As such, we preprocessed the dataset merging the 50-centimeter pixels producing a 1-meter spatial resolution image, with only RGB frequencies to detect the red, green, and blue layers.

Next, we decided on the sample size for our testing data, as well as how we would classify our training data. Regardless of whether the code uses machine learning or deep learning, all supervised classifiers can only be as good as the data used to train them. For the train-test split, we settled on 70% of the sample used for training, while the remaining 30% of the sample was used to test how well our classifiers had learned from the training set, and if they were then able to successfully classify areas on their own.

Building the training sample was somewhat laborious. We decided to rely on the enumeration areas used by the IBGE to conduct the population census as units. We then selected the units that IBGE identified as informal settlements to assemble our dataset, and, then randomly selected 30% of the units in our dataset for the training sample. However, although these areas had been designated by the IBGE as informal settlements, there is much more than just structures in those areas, such as greenery, and even regular settlements side-by-side with informal ones. We then visually identified everything that seemed like an informal settlement.

To visually select our areas of interest within the training sample, we used the Quantum GIS application, a free open-source geographic information system software (QGIS, 2024). We uploaded our 1-meter resolution RGB-only image of the city of Rio de Janeiro into the software and then overlaid the training sample on it. We used a polygon-creating tool to create polygons that encompassed everything that looked like informal settlements, and nothing more. A not insignificant amount of the original selection made by the IBGE was greenery, which was skewing the results of our classifiers and needed to be removed. We then combined all polygons into a single mask, which we then used to train our classifiers.

In this regard, the only thing missing to classify informal settlements with the use of satellite imagery and machine and deep learning was the code. The code used in this study was developed as an ongoing effort of the research group Big Data and SDG of the United Nations Regional Hub for Big Data in Brazil, hosted by the National School of Statistical Sciences (RGBDPS, 2024). It is a Python code that consumes the same satellite image of the city of Rio de Janeiro alongside our training mask and uses it to train various models.

Currently, the code requires pixel coordinates and the size of a small test area, which must contain at least a part of one of the training units somewhere. Once this was done, the program ran all of the classifiers and returned a table of scores, which we used to gauge how well each classifier was learning and replicating the patterns for the identification and classification of informal settlements.

This research focuses on a selected area with the ultimate goal of classifying every pixel in the city of Rio de Janeiro. For this study, the test ran in different regions of the city and will be adjusted to best fit all different types of informal settlements in the city. To maximize our chances of success, we are currently running eight models of classifiers, namely XGBoost, Random Forrest Classifier, LightGBM Classifier, Gradient Boost, Kmeans, GassianNB, Logistic Regression and MLP Classifier.

The study has shown some promising results so far. Once all pre-processing was done and we were ready to start running our Python code, we did so by selecting a small area for testing. To adapt the code to the power available in each researcher's machine, the code was adjusted to run test squares, the size of which was chosen by the person running the code. As such, the only thing to be entered into the code before running it was the pixel coordinates for the left-most pixel in our desired area, along with the desired size. In this paper, we are showcasing the pixel coordinates 15733,13834, with a size of 1000×1000 pixels, or 1 km^2. This referred to the wonderful district of Campo Grande.

4. Results and Discussion

Before running any classifiers or even doing any pre-processing of our image, we adjusted the Python code to provide a preview of the data for the test square, including the image, the mask of informal settlements, and the images overlaid on each other. In this respect, once the desired pixel coordinates were uploaded, the code returned the selected satellite image, the mask created by the visual classification indicating the presence of an informal settlement, and the overlapping of both (**Fig. 1**).

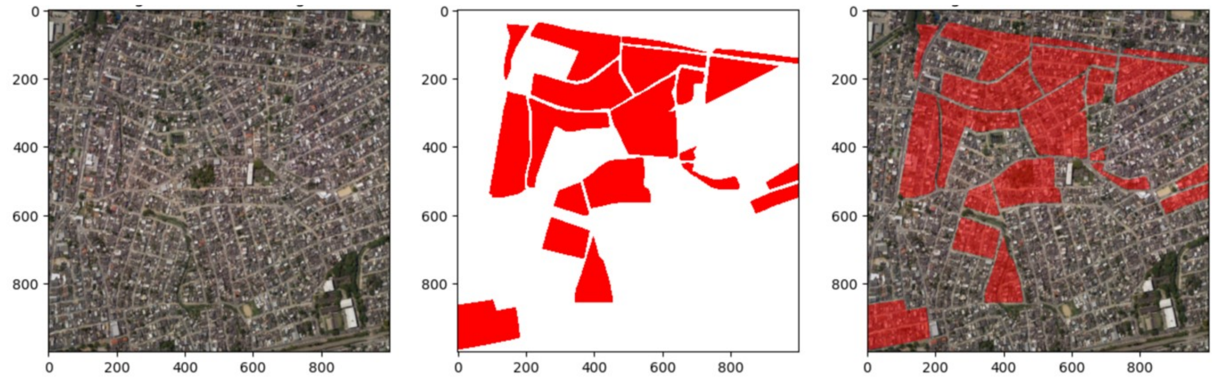


Fig. 1 Satellite image, mask and mask over image

Next, we ran the classifiers. They analyzed the different features of the image, and the areas marked as informal settlements and tried to replicate the classification. This was by far the heaviest and most time-consuming part of the entire code exercise. Once the classifiers had all run to completion, the code once again returned a visual representation of the classifiers' predictions (**Fig. 2**).

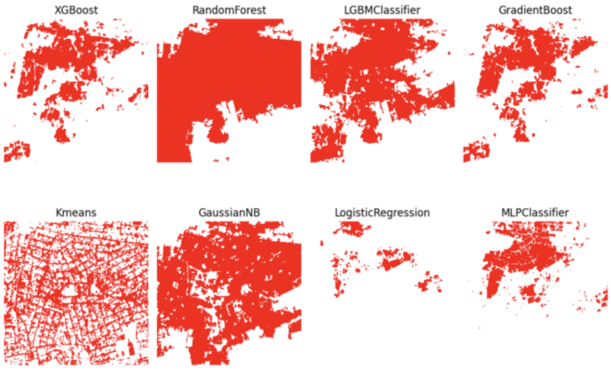


Fig. 2 Prediction of the presence of informal settlements by classifier

Some classifiers performed better at making predictions than others. Out of our eight classifiers, Kmeans, LogisticRegression, and GaussianNB did very poorly when considering the harmonic mean of the precision and recall (F1 score), while XGBoost and GradientBoost performed the best.

Table 1 Performance of classifiers by various metrics

Models	Accuracy	Precision	Recall	F1 Score	IoU
XGBoost	89.3%	82.8%	74.6%	78.5%	**75.6%**
GradientBoost	**90.3%**	**85.0%**	76.2%	**80.4%**	59.7%
LGBMClassifier	73.3%	49.4%	87.6%	63.2%	54.9%
RandomForest	55.6%	36.6%	**94.8%**	52.8%	54.6%
Kmeans	53.0%	25.3%	40.9%	31.3%	53.4%
MLPClassifier	80.9%	72.8%	43.1%	54.1%	50.5%
GaussianNB	48.7%	30.1%	72.6%	42.6%	49.4%
LogisticRegression	75.3%	64.4%	12.7%	21.2%	49.4%

To better understand the different scores and where the classifiers stood in relation to each other, we separated the classifiers into three groups based on their Intersection over Union (IoU) performance, namely:

- High performance (70% or more): XGBoost
- Moderate performance (50% to less than 70%): RandomForest, LGBMClassifier, GradientBoost, Kmeans, and MLPClassifier
- Low performance (less than 50%): GaussianNB and LogisticRegression

While some classifiers might seem exceptional according to one metric or another, it is important to consider all of them. An exceptionally high recall, such as the RandomForest and the LGMBClassifiers, can mean these models are memorizing the original mask rather than learning from it, which would lead to low accuracy and precision. This was exactly the case with our classifiers. On the other hand, even though GradientBoost had an overall accuracy of 90.3%, it still only had 59.73% in IoU, also known as the Jaccard Index, which is calculated by dividing the overlapping area by the union area. In fact, even though the

GradientBoost model had better accuracy, recall, and precision than XGBoost, their IoU scores were inverted, with the XGBoost ahead by over 15 percentage points. Although there were two close contenders for best classifier among most of the different metrics, with the GradientBoost and XGBoost being very close together, considering the IoU score, a far better metric to judge segmentation tasks, the XGBoost had a significant advantage, making it our best classifier currently.

These numbers are, however, only true for this unique 1 km square. The results do not apply to other areas. Since the different areas of the city have different geological features and informal settlements largely follow and adapt to these, they can look very different from one another. We are currently working on generalizing from a single square to the entire city. Rio de Janeiro is a large and hugely varied city, not only when it comes to its geography, but also its settlements and buildings. Informal settlements come in many different shapes and sizes. While some are small and go up hillsides, others are in large and flat areas. Therefore, we are in the process of assessing which of the models are best at recognizing the many different types of slums in the entire city, and how to best take them into account for a final study.

5. Conclusion

The census is the single most important tool the government has when it comes to urban planning. In particular, it is the best and most accurate way of properly establishing any public policies, including education, health, social security, pensions, sanitation, housing, and urban infrastructure, along with essential services to homes and businesses, to name a few. However, hidden societies are dependent on staying under the radar, particularly of the government. This is facilitated by the fact there are many areas where government agents rarely go. With the safety of these populations and surveyors in mind, we are trying to complement the data from the census takers by using satellite imaging and machine learning.

Our machine learning classifiers have proven efficient in recognizing and classifying slums using remote sensing data. While not all classifiers reached the original proposal of the study, two of our models produced excellent results while others have shown promise. While the XGBoost model is the best classifier in this study, given the significant distance from its closest contender by the IoU score, the GradientBoost model had the best overall accuracy, precision, recall, and F1 score, albeit not by much. As such, it is the combination of these many factors that put XGBoost ahead of GradientBoost.

While the results have been extremely promising, as always there are limitations and weaknesses. For now, we are still working on applying the classifiers to a larger area, trying to overcome the heterogeneity among different informal settlements. Even though only a sample of areas is needed to train the classifiers, a sample of a large area is still large, which must be visually identified, potentially a very time-consuming process. Similarly, while this study is a collaborative effort, in the case of a large-scale or country-wide application, this will be a significant effort to consider, since any model can only be as good as the data used to train it.

On the road to the full study, we have already begun our efforts to generalize our findings to the entire city. We have extended our test area to 30% of recognized slums of the city of Rio de Janeiro, and we have begun our visual recognition efforts. Since the area is large and maintaining quality is imperative, we are working hard to go through all the test data manually and classify the informal settlements. Additionally, we are in the midst of trying to extend our classifier selection with more complex algorithms, such as Convolutional Neural Networks (CNN), which would significantly improve our generalization ability. Furthermore, additional preprocessing of remote sensing data could also be beneficial to further improve the performance of our existing classifiers. As such, we look forward to bringing the full version of this study to readers once it is finalized.

Acknowledgements

The authors are grateful to the anonymous reviewers for their constructive comments. The third author thanks the Office to Coordinate Improvements of Higher Education Personal (CAPES) for the scholarship granted.

Authors’ Contributions

H.D.C: Writing - Original Draft; A.D.S.: Writing - Review & Editing, Supervision, Project administration; B.B.M.: Writing - Original Draft, Software; B.S.G.: Software; I.M.N.: Writing - Review & Editing, Supervision, Software; M.R.A.M.: Supervision; M.N.F.C.: Software. All authors read and approved the final manuscript.

Availability of data and materials

On demand from the Corresponding Author

Ethics approval and consent to participate

Not Applicable

Consent for publication

Not Applicable

Competing interests

The authors declare that they have no competing interests.

References

ABS - Australian Bureau of Statistics. (2006). *2006 census: Census through the ages*. Accessed on September 17, 2024. Available at: <https://www.abs.gov.au/websitedbs/D3310114.nsf/4a256353001af3ed4b2562bb00121564/eadaffffb171cab6ca257161000a78d7>.

Alrasheedi, K. G., Dewan, A., & El-Mowafy, A. (2023). Using local knowledge and remote sensing in the identification of informal settlements in Riyadh City, Saudi Arabia. *Remote Sensing*, *15*(15), 3895. https://doi.org/10.3390/rs15153895

Assarkhaniki, Z., Sabri, S., & Rajabifard, A. (2021). Using open data to detect the structure and pattern of informal settlements: an outset to support inclusive SDGs’ achievement. *Big Earth Data*, *5*(4), 497–526. https://doi.org/10.1080/20964471.2021.1948178

Cinnamon, J., & Noth, T. (2023). Spatiotemporal development of informal settlements in Cape Town, 2000 to 2020: An open data approach. *Habitat International*, *133*, 102753. https://doi.org/10.1016/j.habitatint.2023.102753

Duque, J., Patino, J., & Betancourt, A. (2017). Exploring the potential of machine learning for automatic slum identification from VHR imagery. *Remote Sensing*, *9*(9), 895. https://doi.org/10.3390/rs9090895

Graesser, J., Cheriyadat, A., Vatsavai, R. R., Chandola, V., Long, J., & Bright, E. (2012). Image based characterization of formal and informal neighborhoods in an urban landscape. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *5*(4), 1164–1176. https://doi.org/10.1109/jstars.2012.2190383

Gram-Hansen, B. J., Helber, P., Varatharajan, I., Azam, F., Coca-Castro, A., Kopackova, V., & Bilinski, P. (2019). Mapping Informal Settlements in Developing Countries using Machine Learning and Low Resolution Multi-spectral Data. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 361–368. https://doi.org/10.1145/3306618.3314253

Guitarrara, P. (n.d.). *Cidade do rio de janeiro: Mapa, Bandeira, População*. Brasil Escola. https://brasilescola.uol.com.br/brasil/cidade-do-rio-de-janeiro.htm

Hub, UN Regional Hub for Big Data in Brazil. (2024). Consultation on the Use of Big Data in Latin America and the Caribbean. https://hub.ibge.gov.br/consulta.htm.

MacFeely S. (2019). The Big (data) Bang: Opportunities and Challenges for Compiling SDG Indicators. United Nations Conference on Trade and Development. DOI: 10.1111/1758-5899.12595.

Oliveira, L. T., Kuffer, M., Schwarz, N., & Pedrassoli, J. C. (2023). Capturing deprived areas using unsupervised machine learning and open data: a case study in São Paulo, Brazil. *European Journal of Remote Sensing*, *56*(1). https://doi.org/10.1080/22797254.2023.2214690

Persello, C., & Kuffer, M. (2020). Towards uncovering socio-economic inequalities using VHR satellite images and deep learning. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 3747–3750. https://doi.org/10.1109/IGARSS39084.2020.9324399

Prabhu, R., & Alagu Raja, R. A. (2018). Urban Slum Detection Approaches from High-Resolution Satellite Data Using Statistical and Spectral Based Approaches. *Journal of the Indian Society of Remote Sensing*, *46*(12), 2033–2044. https://doi.org/10.1007/s12524-018-0869-9

QGIS. (2024). Accessed on September 17, 2024. Available at: <https://qgis.org/project/overview/>.

RGBDPS, Research Group on Big Data for Public Statistics. (2024). Accessed on September 17, 2024. Available at: <https://dgp.cnpq.br/dgp/espelhogrupo/787479>.

Silva, A. D. da, Oliveira, B. M. M. de, Peixoto, Í. G., & Souza, L. B. S. de. (2023). Overview of the use of big data for official statistics in Latin America and the Caribbean. *Statistical Journal of the IAOS, 39*(1), 171–177. https://doi.org/10.3233/SJI-220092

Stark, T., Wurm, M., Zhu, X. X., & Taubenbock, H. (2020). Satellite-Based Mapping of Urban Poverty With Transfer-Learned Slum Morphologies. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 5251–5263. https://doi.org/10.1109/JSTARS.2020.3018862

Stark, T., Wurm, M., Zhu, X. X., Taubenbock, H. (2020). Satellite-based mapping of urban poverty with transfer-learned slum morphologies. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 5251–5263. https://doi.org/10.1109/jstars.2020.3018862

UN, United Nations. (2017). Principles and Recommendations for Population and Housing Censuses. Accessed on September 17, 2024. Available at: <https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Principles_and_Recommendations/Population-and-Housing-Censuses/Series_M67rev3-E.pdf>.

UN-Habitat. 2021. Relatório Anual Brasil 2020. Accessed on September 17, 2024. Available at: <https://brasil.un.org/pt-br/137253-onu-habitat-brasil-re%C3%BAne-desafios-e-conquistas-de-2020-em-relat%C3%B3rio-anual>.

UN-Habitat. 2023. Relatório Anual 2022 do ONU-Habitat. Accessed on September 17, 2024. Available at: <https://relatorio-anual-2022.netlify.app/ >.

Wang, J., Kuffer, M., & Pfeffer, K. (2019). The role of spatial heterogeneity in detecting urban slums. *Computers, Environment and Urban Systems*, *73*, 95–107. https://doi.org/10.1016/j.compenvurbsys.2018.08.007

Wang, J., Kuffer, M., & Pfeffer, K. (2019). The role of spatial heterogeneity in detecting urban slums. *Computers, Environment and Urban Systems*, *73*, 95–107. https://doi.org/10.1016/j.compenvurbsys.2018.08.007

DATASET
REPORTS

journals.royaldataset.com/dr