amazonULC: A Data Package with Urban Land Cover Classifications for a Selection of Cities in the Brazilian Amazon

Abstract

Remote sensing has been a primary method for collecting information about Earth's surface for decades and is an even more relevant resource in the case of developing nations. Despite this, cities in the Amazonian region lack databases and cartographic publications. For this data paper, we select a group of four cities and one metropolitan region as study sites to provide land cover classification maps. Using imagery from the CBERS-4A satellite's WPM sensor, we created a classification model that combines the Geographic Object-Based Image Analysis method, data mining strategies, and a random forest machine learning algorithm. The results are valuable to discern different intra-urban land cover classes, with an overall accuracy level in the validation samples of over 85%. An R package provides analysis-ready data and documentation to support researchers who wish to avail themselves of these resources.

Keywords: Urban Land Cover, Amazon, Remote Sensing, Random Forest, Data Mining, Reproducible research

1. INTRODUCTION

Remote sensing has been the primary method for collecting data about the Earth's surface in recent years, particularly for developing countries (Zhu et al., 2022). Numerous opportunities exist to map and monitor cities using remote sensing data which provide accurate quantitative data over time and space, and operate as the basis for physical, climatic, and socioeconomic indicators, sometimes as a complement or even as an alternative to conventional surveys such as the census. In addition, these data can assist in urban planning once they are converted into information and are consistently prepared and updated (Weng and Quattrochi, 2018).

Land use and land cover mapping are the most significant applications of remote sensing in urban studies in Brazilian cities (de Almeida, 2010). Despite this, most methods that use geoprocessing and remote sensing tools to investigate Brazilian cities have concentrated on Southeastern Brazil (dos Santos et al., 2022). Although research about urban areas in the Amazonian region is not completely absent, public managers still need to include such studies in public policies. This is somewhat complicated by the lack of databases and cartographic publications specific to Amazonian cities (Cardoso et al., 2020).

The Amazon region presents some unique characteristics. They include the historical process of human occupation in a rain forest region, with distinctive climatic and environmental characteristics. This is paired with relatively low levels of socioeconomic development, and a strong cultural influence from indigenous peoples, sometimes complicated by land and environmental conflicts. In addition, the growth of small and medium-sized cities is a characteristic of the urbanization process in the Amazon, following the reorganization of the national urban network and a new territorial division of labour (Jr, 1998, 2011).

In this context, the objective of this data paper is to present an Open Data Product (ODP) useful to support research and policy analysis in the cities of Altamira, Cametá, Marabá, Santarém and part of the Metropolitan Area of Belém, all them in the state of Pará. ODPs provide analysis-ready data often with a value-added component (see Arribas-Bel et al., 2021). Presently, we obtain remotely sensed data from

*Corresponding author

public sources; the value added comes from the development of an urban land cover classification system for the cities mentioned above. Using imagery from the CBERS-4A satellite's WPM sensor, we created a classification model based on the Geographic Object-Based Image Analysis (GEOBIA) method, data mining strategies, and a random forest machine learning algorithm.

This paper is an example of open and reproducible research that uses only open software and data sources for imagery processing analysis. We obtained all data from publicly available sources and organized them in the form of the Amazon-ULC R data package. In addition to the Open Data Product, all the code necessary to reproduce, modify, and contribute to the analysis is also shared, in line with best practices in spatial data science (Brunsdon and Comber, 2021; Desjardins et al., 2022).

2. MATERIALS AND METHODS

We summarise our methodology in five main steps.

The definition of the study area is the first one. The WPM image preprocessing comes second, followed by the segmentation, extraction, and selection features. The classification of the segments is the final process. These steps are described in more detail below.

For this work, the remote sensing data consist of:

- WPM images from the CBERS-4A satellite: two orthorectified images, one panchromatic and one multispectral, from the year 2020. The WPM sensor provides panchromatic and multispectral images simultaneously. The panchromatic images have 2 meters of spatial resolution, with a spectral range between 0.45 and 0.90 µm. Multispectral images have a spatial resolution of 8 meters, with spectral bands: blue (blue, 0.45 - 0.52 µm), green (green, 0.52 - 0.59 µm), red (red, 0.63 - 0.69 µm), NIR (near infrared, 0.77 - 0.89 µm). The radiometric resolution of the images is 10 bits. The imaged swath width is 92 km and the revisit period is 31 days (INPE, 2019);

The following software applications supported this research:

- QGIS 3.18 (Team, 2021): for the thematic maps, and image segmentation.
- TerraView 5.6.3: for preprocessing the satellite images, with the GeoDMA 2.0.1 add-on (Körting et al., 2013) for extracting attributes.
- Python (vanRossum, 1995) and R Languages (R Core Team, 2022) : preparing and mining the data and classifying the geographic objects with the Random Forest algorithm.

2.1. Study areas

Together, the study areas comprise about 1200 km^2 located above the municipal seats of Altamira (153 km^2), Cametá (44 km^2), Marabá (164 km^2), Santarém (143 km^2) and part of the Metropolitan Area of Belém (614 km²) (Figure 1). All cities are located in the state of Pará, within the Brazilian Legal Amazon. The study site also covers areas classified as rural in the surroundings of the cities. For Altamira and Cametá, we used the delimitation of dos Santos et al. (2022). For the other three study areas, we made the delimitation according to the methodology (Gonçalves et al., 2021) - which uses nighttime light images to identify potentially populated areas. According to the *Fundação Amazônia de Amparo a Estudos e Pesquisas*, it is estimated that the municipalities mentioned above have more than 3.5 million inhabitants, representing more than 38% of the population of the State of Pará.

2.2. Pre-processing image

We started the pre-processing stage by clipping the images with the study area polygons. Then, we fused the images with the Principal Components Analysis (PCA) technique. Afterward, we calculated the following indices on the original images:

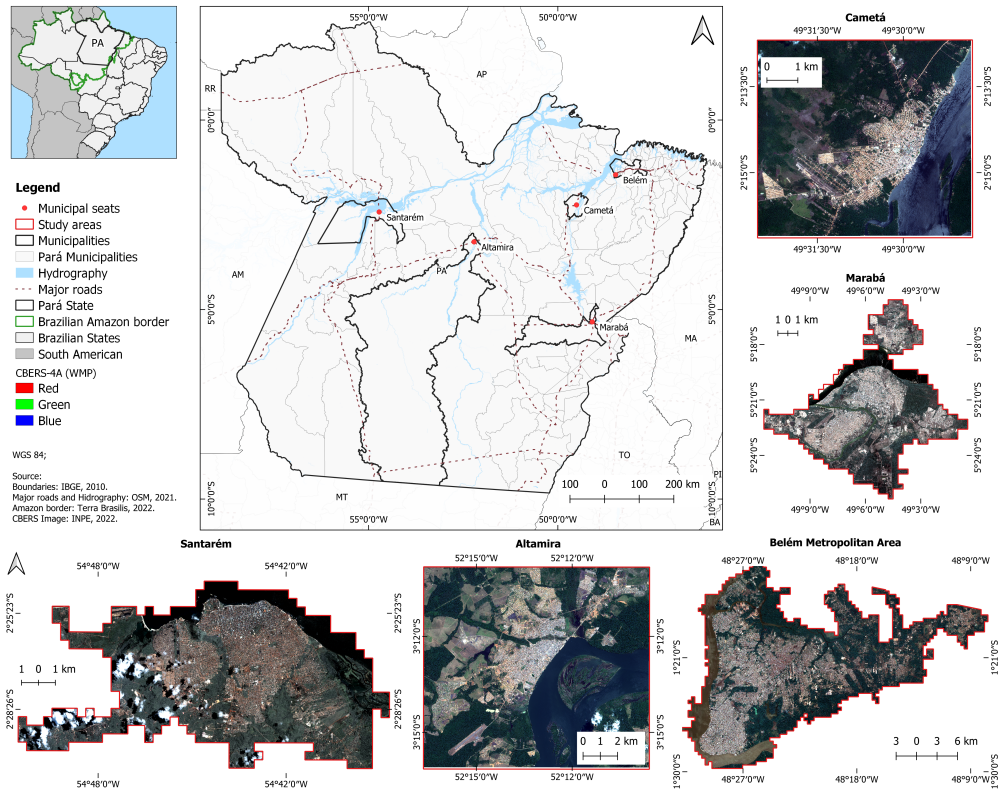

Figure 1: Location of the study areas.

- Normalized Difference Vegetation Index (NDVI): division of the near-infrared (NIR) band by the red band, normalized, to analyze the presence and condition of vegetation (Equation 1):

$$NDVI = \frac{NIR - RED}{NIR + RED}. \quad (1)$$

- Normalized Difference Roof Index (NDRI) (De Pinho et al., 2011): division of the red band by the blue band, normalized, to identify the presence of ceramic roofs and areas with exposed soil (Equation 2):

$$NDRI = \frac{RED - BLUE}{RED + BLUE}. \quad (2)$$

- Bare Soil Area Index (BAI): the normalized division of the blue band by the near-infrared band, to identify exposed soil (Equation 3):

$$BAI = \frac{BLUE - NIR}{BLUE + NIR}. \quad (3)$$

- Normalized Difference Water Index (NDWI): the normalized division of the green band by the near-infrared band (Equation 4). This formula highlights the amount of water in water bodies:

$$NDWI = \frac{GREEN - NIR}{GREEN + NIR}. \quad (4)$$

We obtained texture metrics from the Gray Level Co-occurrence Matrix – (GLCM) (Haralick et al., 1973), considering all bands of the original images. We computed all the GLCM metrics using 3 x 3 pixels as the window size. We also added fraction images produced by a linear spectral mixture model (LSME) to perform the classification. The LSME technique assumes that given the spectral response of pure targets, it is possible to extract highlighted features of the desired targets in a synthetic image format - which would facilitate the identification of such targets. The LSME uses a linear relationship to symbolize the spectral mixture of each pixel target, and the spectral response is defined as a combination of each component of the mixture, obeying a proportion between the components whose summation equals (Shimabukuro and Ponzoni, 2019).

In this study, we chose the pure pixels that served as the input to the LSME model directly from the WPM multispectral images (with an 8-meter spatial resolution), adopting the classes of vegetation, soil, and water.

2.3. Image segmentation and feature extraction

The fused WPM image was segmented using the Mean-shift algorithm. Mean-shift is a region-based segmentation algorithm that uses local homogenization, where each pixel is replaced by the average of pixels in a search window whose value is within a predefined distance interval (Comaniciu and Meer, 1997). Dos Santos et al. (2022) performed LULC maps for Altamira, Cametá and Marabá based on WPM imagery, using the mean-shift algorithm implemented in the Orfeo Toolbox (OTB) (Grizonnet et al., 2017). The values for the spatial radius, range radius, maximum number of interactions, and minimum region size determine the size and shape of the segments. Therefore, we adopt the following values: spatial radius of 5 pixels, range radius of 100 pixels, maximum number of interactions of 100 times, and minimum region size of 15 pixels.

We defined classes of objects by size and shape aiming to identify the land cover classes: “Shrub Vegetation” (SV), “Herbaceous Vegetation” (HV), “Water” (Wa), “Exposed Ground” (EG), “High Gloss Cover” (HG), “Ceramic Cover” (Ce), “Fiber Cement Cover” (FC), “Asphalt Road” (As), “Terrain Road” (Te), “Cloud” (Cl) and “Shadow” (Sh).

After segmenting the WPM images, we extracted features for the geographic objects. We extracted spatial features and the mean, maximum, minimum, standard deviation, and range values of the biophysical index layers, GLCMs, multispectral, and fraction images.

2.4. Feature extraction

We employed a stratified random sampling following the prevalent land cover classes. We set 70% of samples for training and the remaining 30% for validation. After sampling, the variables underwent a pre-classification treatment process, where potential `null` values were filled in. We next ordered the variables based on their ability to distinguish between different interest classes. We used the R^2 from the Analysis of Variance (ANOVA) for the ordering. Due to their poor ability to discriminate between classes, we eliminated the attributes with R^2 lower than 0.1.

2.5. Classification Model

We developed a supervised classification model using the training base with the random forest (RF) algorithm for each city. We used the Randomized Search technique to define the RF model's hyperparameters. In this technique, we input the algorithm values of hyperparameters that will be randomly selected and combined, returning a combination that results in the best possible classification. The method selects the optimal set of hyperparameters using a performance metric and a fixed number of iterations. Randomized Search uses cross-validation, dividing the training base into k parts (folds), and the model is trained and evaluated k times. The algorithm chooses a component (fold) for each iteration to use as an evaluation before training the model on the other $k-1$ parts.

We examined the hyperparameter values listed in Table 1 using Randomized Search. We used a 5-fold cross-validation, an F1-Score performance indicator, and a maximum of 100 parameter combinations (iterations). After the Random-search selecting the best hyperparameters for the RF model, we classified the entire database.

Table 1: Hyperparameters tested in the RF model.

Hyperparameter	Values
Number of trees	[1, 20, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 700, 800, 900, 1000, 1500, 2000]
Criterion	['gini','entropy']
Maximum depth	[5, 10, 20, None]
Minimum number of samples to split an inner node	[2, 5, 10]
Minimum number of samples to be in a leaf node	[1, 2, 4]
Boostrap	['True', 'False']

3. RESULTS

We developed a land cover classification model for each city. The database contains the set of model hyperparameters that produced the best classification. The metrics F1-Score Macro (average of the F1-Score of the different classes), F1-Score Weighted (average of the F1-Score of the different classes weighted by the number of samples), and Global Accuracy of the classification produced by the RF model over the validation samples are presented in Table 2 for each study area.

Table 2: Metrics for evaluating the land cover classifications for each study area.

Study area	**F1-Score Macro**	**F1-Score Weighted**	**Global Accuracy**
Altamira	0.90	0.90	0.90
Belém Metropolitan Area	0.75	0.95	0.95
Cametá	0.85	0.85	0.85

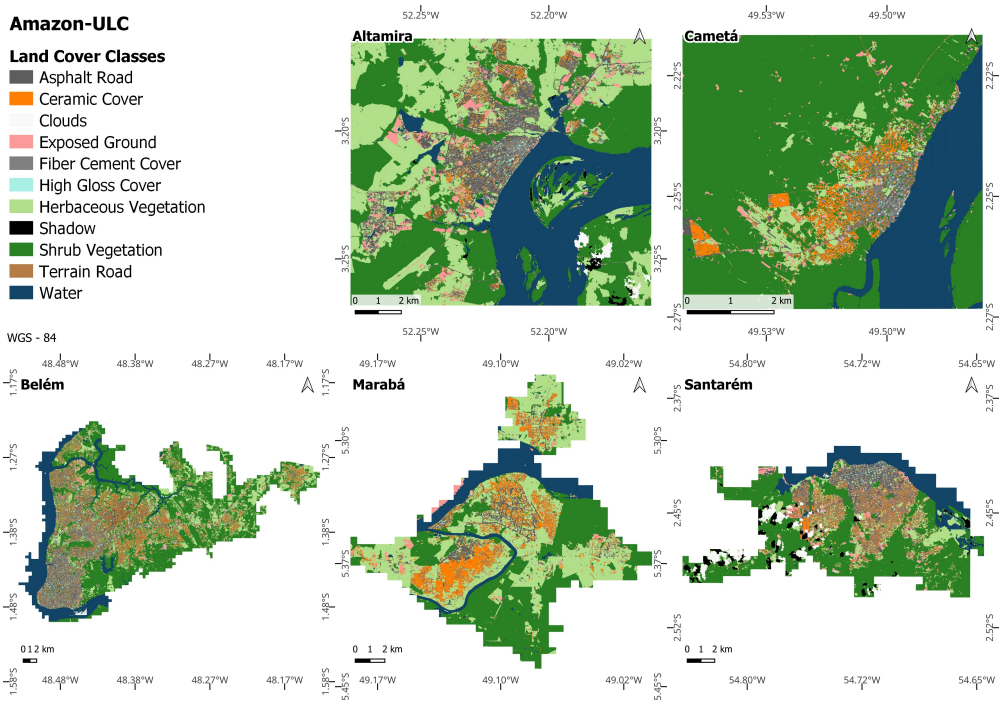

Figure 2: Urban land cover classification maps of Altamira, Cametá, Belém Metropolitan Area, Marabá and Santarém.

Study area	**F1-Score Macro**	**F1-Score Weighted**	**Global Accuracy**
Marabá	0.90	0.95	0.95
Santarém	0.92	0.99	0.99

According to statistics, the developed models obtained Macro F1-Scores above 75%, Weighted F1-Scores and Global Accuracy above 85%. These high rates show that the RF model's classification successfully classified most of the validation samples, demonstrating a good capacity for classifying data not used in the model's development. Visually, it is noticeable that the result is satisfactory and in line with the evaluation metrics (Figure 2).

When evaluated using statistical assessment metrics, Santarém had the best classification. We can explain the better performance of the classification model for Santarém because this study site had a more significant number of samples collected, facilitating the identification of the coverage classes by the algorithm.

In general, the Vegetation classes (Shrub and Herbaceous) and the Water class presented the highest F1-Score. It might be explained by the spectral nature of these classes, which are easily distinguished. On the other hand, the Fiber cement and Ceramic classes had the worst F1-Score - both present spectral characteristics similar to the Asphalt Road and Earth Road classes, respectively, which would justify the higher confusion on the part of the classifier algorithm.

4. STUDY LIMITATIONS

This work seeks to provide land cover bases for five locations in the Brazilian Legal Amazon. However, although it has met its objective, some limitations must be pointed out.

The methodology we adopted for the satellite image classification was the GEOBIA approach. This technique requires a segmentation process over the image, grouping the pixels of similar spectral performance. However, we noticed that some targets were not completely separated during the segmentation process,

compromising the classification result by the classifier algorithm. This segmentation error was more noticeable in the city of Marabá. Consequently, "not pure" segments were identified, like segments that could be divided into two or more classes. This limitation is due to our decision to use only algorithms available in open-access software.

Another point to be raised is that even if the classification accuracy levels are considered satisfactory by the research team, it is still possible to find segments misclassified by the Random Forest algorithm.

5. CONCLUSION

This work aimed to provide land cover classification maps for Brazilian Amazonian cities. Land cover maps are valueable to urban planning in Amazonian cities. They are useful for different issues, such as monitoring urban sprawl, restricting construction in environmental protection areas, assisting in urban zoning, and identifying areas of high density, among others. Our classification model used images from the WPM sensor of the CBERS-4A satellite and combined the GEOBIA approach, data mining techniques and the random gorest machine learning algorithm. The results obtained are promising regarding the ability to discriminate intra-urban cover classes, achieving an overall accuracy level above 85% in the validation samples for all study areas. The integration between the GEOBIA approach and a machine learning algorithm was the reason for this success. In addition, this work also emphasizes the application of the CBERS-4A satellite, recently launched by the Brazilian government in partnership with the Chinese government, for Amazon urban studies.

References

Dani Arribas-Bel, Mark Green, Francisco Rowe, and Alex Singleton. Open data products-a framework for creating valuable analysis ready data. *Journal of Geographical Systems*, 23(4):497–514, 2021.

Chris Brunsdon and Alexis Comber. Opening practice: supporting reproducibility and critical spatial data science. *Journal of Geographical Systems*, 23(4):477–496, 2021.

Ana Claudia Duarte Cardoso, José Júlio Ferreira Lima, Juliano Pamplona Ximenes Ponte, Raul da Silva Ventura, and Roberta Menezes Rodrigues. Morfologia urbana das cidades amazônicas: a experiência do grupo de pesquisa cidades na amazônia da universidade federal do pará. *urbe. Revista Brasileira de Gestão Urbana*, 12, 2020. ISSN 2175-3369.

D Comaniciu and P Meer. Robust analysis of feature spaces: color image segmentation. pages 750–755, 1997. ISBN 1063-6919. doi: 10.1109/CVPR.1997.609410.

Cláudia Maria de Almeida. Aplicação dos sistemas de sensoriamento remoto por imagens e o planejamento urbano regional. *arq. urb*, pages 98–123, 2010. ISSN 1984-5766.

Carolina Moutinho Duque De Pinho, Marta Eichemberger Ummus, and Tessio Novack. Extração de feições urbanas em imagens de alta resolução espacial a partir do estudo do comportamento espectral dos alvos. *Rev. Bras. De Cartogr*, 63:439–448, 2011.

Elise Desjardins, Christopher D. Higgins, and Antonio Páez. Examining equity in accessibility to bike share: A balanced floating catchment area approach. *Transportation Research Part D: Transport and Environment*, 102:103091, 2022. ISSN 1361-9209. doi: https://doi.org/10.1016/j.trd.2021.103091. URL `https://www.sciencedirect.com/science/article/pii/S1361920921003874`.

Bruno Dias dos Santos, Carolina Moutinho Duque de Pinho, Gilberto Eidi Teramoto Oliveira, Thales Sehn Korting, Maria Isabel Sobral Escada, and Silvana Amaral. Identifying precarious settlements and urban fabric typologies based on geobia and data mining in brazilian amazon cities. *Remote Sensing*, 14:704, 2022.

Gabriel Crivellaro Gonçalves, Lucas Maia de Oliveira, Ana Paula D'Asta, and Silvana Amaral. Geoinformação para a visbilidade das Áreas urbanas de cidades amazônicas. *Revista Geoaraguaia*, 11:149–165, 2021. ISSN 2236-9716.

Manuel Grizonnet, Julien Michel, Victor Poughon, Jordi Inglada, Mickaël Savinaud, and Rémi Cresson. Orfeo toolbox: open source processing of remote sensing images. *Open Geospatial Data, Software and Standards*, 2:15, 2017. ISSN 2363-7501. doi: 10.1186/s40965-017-0031-6. URL `https://doi.org/10.1186/s40965-017-0031-6`.

Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, pages 610–621, 1973. ISSN 0018-9472.

INPE. Câmeras imageadoras cbers-4a, 12 2019. URL `http://www.cbers.inpe.br/sobre/cameras/cbers04a.php`.

Saint-Clair Cordeiro Da Trindade Jr. A cidade dispersa: os novos espaços de assentamentos em belém e a reestruturação metropolitana., 1998.

Saint-Clair Cordeiro Da Trindade Jr. Cidades médias na amazônia oriental: das novas centralidades à fragmentação do território. *Revista Brasileira de Estudos Urbanos e Regionais*, 13:135, 2011. ISSN 2317-1529.

Thales Sehn Körting, Leila Maria Garcia Fonseca, and Gilberto Câmara. Geodma—geographic data mining analyst. *Computers and Geosciences*, 57:133–145, 2013. ISSN 0098-3004. doi: https://doi.org/10.1016/j.cageo.2013.02.007. URL `https://www.sciencedirect.com/science/article/pii/S0098300413000538`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

Yosio Edemir Shimabukuro and Flávio Jorge Ponzoni. The linear spectral mixture model, 2019.

QGIS Development Team. Qgis geographic information system, 2021. URL `https://www.qgis.org`.

Guido vanRossum. Python reference manual. *Department of Computer Science [CS]*, 1995.

Qihao Weng and Dale A Quattrochi. *Urban remote sensing*. CRC press, 2018. ISBN 1315166615.

Xiao Xiang Zhu, Chunping Qiu, Jingliang Hu, Yilei Shi, Yuanyuan Wang, Michael Schmitt, and Hannes Taubenböck. The urban morphology on our planet–global perspectives from space. *Remote Sensing of Environment*, 269:112794, 2022. ISSN 0034-4257.