TEMPORAL ANALYSIS OF THE URBAN SPRAWL IN GUARUJÁ USING MACHINE LEARNING AND SATELLITE IMAGES

Tèhrrie König[1], Hermann J. H. Kux[2], Alessandra C. Corsi[3]

[1]National Institute for Space Research - INPE, Av. Dos Astronautas 1758, Jardim da Granja, São José dos Campos, SP, 12227-010, tehrriekonig@gmail.com/tehrrie.pacheco@inpe.br; [21]National Institute for Space Research - INPE, Av. Dos Astronautas 1758, Jardim da Granja, São José dos Campos, SP, 12227-010, hermann.kux@inpe.br; [3]Technologial Institute Research - IPT, Av. Prof. Almeida Prado, 532, Cidade Universitária, São Paulo, SP, 05508-901, accorsi@ipt.br.

ABSTRACT

Satellite images and remote sensing techniques allow several studies, including the identification and characterization of land use and occupation and urban sprawl. Urban sprawl is usually associated with environmental degradation, and an increase in disasters has been observed. Therefore, this study analyzes the urban sprawl of Guarujá municipality using remote sensing techniques, machine learning, and Data Mining. The NDVI was performed to enhance the identification of the vegetation. A temporal analysis from 1990 to 2020 was performed, using satellite images from the Landsat series. The results indicate an increase in the urban area, and, consequently, a decrease in the vegetation. The CART algorithm correctly distinguished the urban areas, the vegetation, and the water. Moreover, the NDVI provided important information about environmental degradation and loss of biomass.

***Keywords* —** Urban sprawl, Data Mining, Machine learning, Landsat.

1. INTRODUCTION

The use of satellite images associated with remote sensing techniques has been used in several different types of research, such as land use and occupation, environmental degradation, disaster management, detection of burning areas, and urban sprawl [1]–[6]. Considering the advance of remote sensing techniques and the total amount of digital data available, it is necessary to determine the most relevant data and information for each study. Therefore, it is recommended to use machine learning and Data Mining. Data Mining is a process that automatically finds patterns and attributes from large data volumes, clustering them. In remote sensing applications, the Data Mining process is used to extract attributes and characteristics (spatial and spectral information) from pixels or objects (regions) present in digital images [7]. There are several different data mining algorithms: decision- tree, Self-Organizes Maps (SOM), Neural networks, the C4.5 algorithm implemented in WEKA software, and Classification and Regression Trees (CART) in the eCognition platform, among others.

Urban sprawl has been occurring in most cities since the 1950s and is usually associated with environmental degradation. Notwithstanding, the increase in the number of disasters has been documented [8]. That's the scenario of Guarujá municipality (Brazil), where the urban sprawl reduces the natural vegetation, and steep slopes have suffered anthropic changes. This study analyzes the urban sprawl of Guarujá municipality using remote sensing techniques, machine learning, and Data Mining. A temporal analysis was conducted to verify how the urban expansion from 1990 to 2021 affected the area, comparing the four satellite image classifications. To improve understanding of land use and occupation and its influence on disasters, a detailed classification of the Vila Baiana neighborhood, which frequently suffers from landslides, was provided.

2. MATERIAL AND METHODS

The study area of this work is the Guarujá, municipality, and Vila Baiana neighborhood, located within the Brazilian southeastern State of Sao Paulo, as presented in Figure 1.

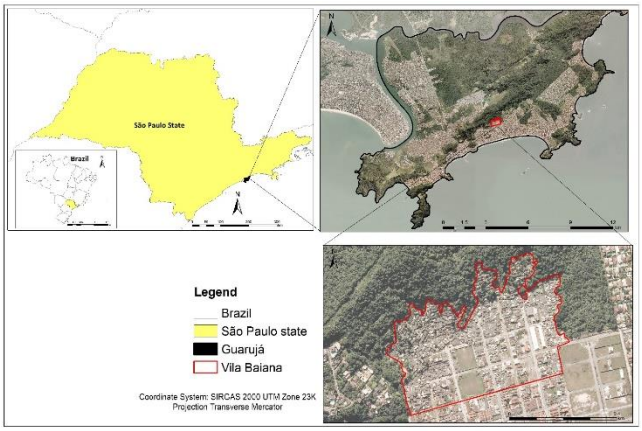

Figure 1. Study area location.

According to IBGE (2019) [9], the municipality has 320,459 inhabitants and a territorial extension of 144,794 km². The mean annual precipitation is 3,000 mm, and the mean annual temperature is 22ºC. As for its geology, the area is on a crystalline plateau, with gneiss and granite from the Pre-Cambrian period. Tropical forests cover the area, and the coastal plain has quaternary coastal sediments of fluvial-marine origin. In some areas, the crystalline basement becomes apparent [10].

The urban occupation started in plain and mangrove areas. However, the city has experienced considerable population growth since the 1950s, intensified in the 1970s, with the economic development due to industries, port-related activities, civil constructions, and tourism. Consequently, the price of land increased sharply, and people with low income started to build their houses in steep areas, on cheap but improper terrain [10].

To analyze the urban sprawl of Guarujá four images of Landsat satellites from the years 1990 (Landsat 5), 2013 (Landsat 8), 2020 (Landsat 8), and 2021 (Landsat 8) were acquired and preprocessed. Using an orthophoto with 1-meter of spatial resolution and eCognition software, the Vila Baiana neighborhood was classified accordingly to the types of soil covers (ceramic roof, concrete roof, vegetation).

The satellite images were pre-processed, which consisted of two steps: pansharpening and orthorectification. Both processes were developed using ArcGIS and ENVI software. The pansharpening operation provided an image with the best spatial resolution from the panchromatic band while retaining the spectral content from the multispectral bands. Landsat 5 images do not have a panchromatic band, therefore, the pansharpening processes were performed exclusively for Landsat 8 images. The Gram-Schmidt method was chosen due to its improvement for the best distinction of objects (vegetation, urban area, sand/bare soil, water) in the scene [11]–[14] The image segmentation and the sample acquisition were performed using the eCognition software. The multiresolution segmentation was performed using the following parameters: shape 0.1 and compactness 0.5.

The algorithm CART, implemented in the eCognition software, was used to extract the most relevant attributes, generating a decision tree. Based on this decision tree, the images were classified using the Object-based Image Analysis (OBIA) paradigm. The OBIA paradigm extracts semi-supervised information from satellite images. It clusters similar objects, considering the pixel information and its neighbors [15], [16].

To improve the distinction between the urban area and the vegetation, the Normalized Difference Vegetation Index (NDVI) was calculated. The NDVI is used to differentiate the vegetation areas from the non-vegetation areas. The leaves have a strong reflectance in the near-infrared band and a weak reflectance of chlorophyll and other leaf pigments in the visible wave band red [17], [18]. The NDVI formula is presented in equation 1, and the temporal analysis of vegetation changes using NDVI is shown in Figure 3.

$$\boldsymbol{NDVI} = \frac{(\boldsymbol{NIR}-\boldsymbol{Red})}{\boldsymbol{NIR}+\boldsymbol{Red}} \quad (1)$$

The Vila Baiana classification characterized the different covers, such as types of roofs, roads, and vegetation. An orthophoto with 1 m of spatial resolution was used. Due to the different sizes of the objects in the scene, two segmentation levels were applied: level 1 to discriminate larger objects, such as blocks and streets, and level 2 to identify the types of roofs and vegetation cover. The first level consisted of the distinction between blocks and streets.

A multiresolution segmentation was performed using a thematic layer and the following parameters: scale 500, shape 0.9, and compactness 0.5. Following, the “elliptic fit” attribute was performed to identify if an object fits in an elliptic with similar proportions, in which 0 means that the object does not fit, and 1 means it fits. The block objects range from 0.6 to 0.8. Afterward, using the “assign class” algorithm, two threshold conditions were defined: objects >= 0.6 are assigned as blocks, and objects <= 0.1 are assigned as roads.

The second segmentation level is performed to identify smaller objects, such as types of roofs and vegetation cover. In segmentation Level 2, the class Blocks were used as a filter, meaning that the segmentation procedure occurs only within the blocks. The multiresolution segmentation algorithm was applied, using the following parameters: scale 30, shape 0.1, and compactness 0.5. The last step before the classification was the sample acquisition. It consists of the selection of image features that correctly represent the class objects. Five classes were described: ceramic roofs, concrete roofs, roofs with different materials (named “other roofs”), arboreal vegetation, and grass vegetation. Data mining is important to determine the most relevant attributes to classify the image. This study used the CART algorithm, available on the eCognition platform. It results in a decision tree containing the variables and determining thresholds for the identification and separation of each class. And the last step consists of the application of thresholds and variables to classify the image.

3. RESULTS AND DISCUSSION

The urban sprawl is a direct consequence of population growth and the development of Guarujá municipality. During the past 31 years (1990-2021), deforestation of native vegetation increased to open space for urbanization. Figure 2 presents the temporal analysis of urban sprawl for 1990, 2013, 2020, and 2021.

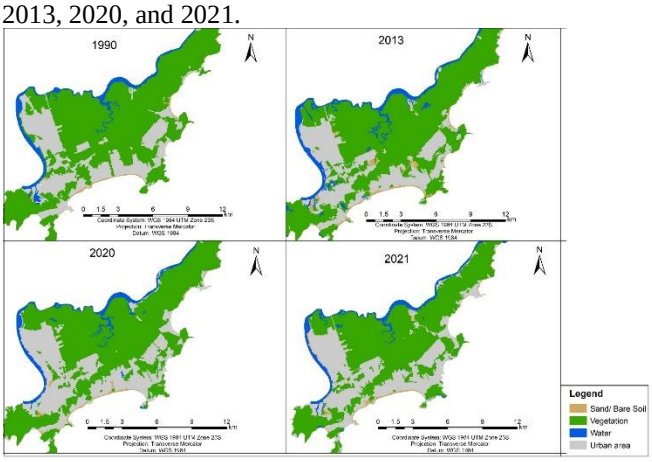

Figure 2. Urban sprawl from 1990, 2013, 2020, and 2021.

Analyzing Figure 2, it is possible to realize that was an intense increase in the urban area from 1990 to 2021, represented in gray. Moreover, the urban sprawl continues to

intensify rapidly since, in 2021, the increase was 7,9% above the amount in 2020. Furthermore, the removal of natural vegetation continues, giving space to the city's expansion. Table 1 shows the built-up area occupied by the town and the vegetation cover for the years 1990, 20213, 2020, and 2021.

Year	Urban area (km²)	Vegetation (km²)
1990	36.25	96.00
2013	37.61	89.91
2020	43.76	88.73
2021	47.25	84.25

Table 1. Variation of urban and vegetation area (km²) from 1990 to 2021.

Analyzing Table 1 it is possible to determine a correlation between the increase in the urban areas and the decrease in vegetation-covered areas. From 1990 to 2021, the urbanization process increased by 30%, while the vegetation area suffered a 12% reduction. The development of urban areas destroys vegetation. To improve the distinction between the urban area and the vegetation, the Normalized Difference Vegetation Index (NDVI) was calculated.

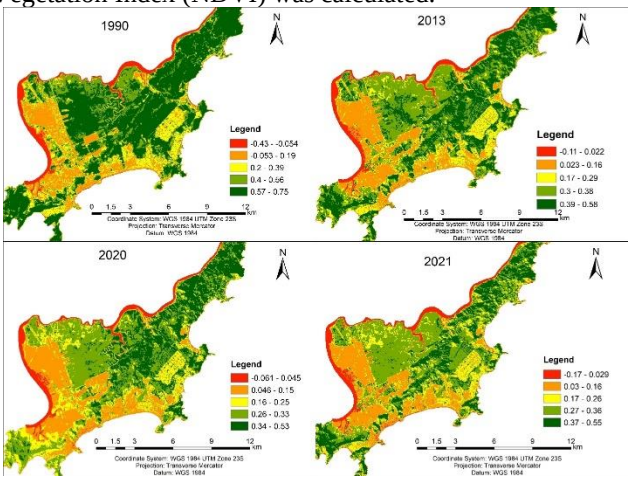

Figure 3. Temporal analysis of variation in NDVI for 1990, 2013, 2020, and 2021.

The analysis of Figure 3 shows the vegetation changes in the past 31 years. The water, represented in red, has low reflectance and, consequently, lower values of NDVI. The colors orange and yellow characterized the urban areas according to the degree of urbanization (high level of urbanization and medium level, respectively). The vegetation is represented by green: the light green areas have lower biomass than those in dark green.

In the 1990 classification, it was observed that vegetation cover is denser and spreads over most of Guarujá municipality. Dark green is the predominant color, meaning that most forest areas were preserved. However, in 2021, a reduction of the vegetation cover areas and the green-leaf density is perceptible. Few forest areas are maintained, and the leaf-area density has decreased.

Those people who cannot afford a house or land in the central part of Guarujá start to build their houses on the slopes, favoring deforestation [19], [20]. The weight of several constructions on steep slope areas, associated with improper water drainage and deforestation, decreases the slope stability and increases the risk of accidents to the population [20]–[22]. Vila Baiana neighborhood is one of these areas, where several houses were improperly constructed in declivity areas, which are commonly affected by landslides. The identification of different constructions, soil, and covers of the Vila Baiana neighborhood is presented in Figure 4.

Figure 4. Classification of Vila Baiana

Figure 4 identifies several constructions built on the edge of the arboreal vegetation, especially in steep slope areas.

The houses localized in slope areas, meaning on the edge of arboreal vegetation, are mostly of the concrete roof (totalizing 74 m²) and just a few with a different type of roof (“other roofs” in the classification). In these slope areas, it is not possible to identify roads, but many houses are observed in a small space. The roads are well delimited in the flat land but not on the upper slopes. It indicates any pavemented street in high declivity areas, becoming difficult to access the area. Notwithstanding, the ceramic roof class is the less representative roof type, with only 8.254 m² of constructed area. They occur in places where it is still possible to determine the blocks and the roads, meaning that these areas are part of the urban planning of Guarujá municipality.

The grass vegetation predominates in two blocks, and both are soccer fields for the community to play. The arboreal vegetation occurs mostly in steep slope areas, with only a few polygons mixed with the constructions.

The error matrix was calculated to assess the classification accuracy, as presented in Figure 5.

	Reference polygon							
	Roads	Ceramic roof	Concrete roof	Other roof	Arboreal Vegetation	Grass Vegetation	Total	User's accuracy
Roads	2	0	0	0	0	0	2	1.00
Ceramic roof	0	64	1	1	0	0	66	0.95
Concrete roof	0	0	369	8	2	1	380	0.97
Other roof	0	1	11	325	3	2	342	0.95
Arboreal vegetation	0	0	2	1	728	15	746	0.98
Grass vegetation	0	0	7	1	3	76	87	0.87
Total	2	64	386	337	740	94	1623	
Producer's accuracy	1.00	0.98	0.96	0.96	0.98	0.81		
Global accuracy	0.96							

Figure 5. Matrix error of Vila Baiana classification.

The error matrix shows a spectral confusion between the classes "Other roof" and "Concrete roof." Analyzing the orthophoto, it is observed that some concrete roofs are lighter than others. The material used in these constructions probably has a similar composition to those used by the class "Other roof," justifying the confusion in the classification process. Moreover, spectral confusion between the grass vegetation and the concrete roof is observed. Most polygons wrongly classified as grass vegetation are in areas with high declivity because these sections were covered with forest, and to build houses, the trees were removed, and the grass is regrowth. Therefore, a grass pixel in a polygon can confuse the algorithm, classifying it incorrectly. Due to the spectral similarity, the "Arboreal vegetation" and "Grass vegetation" also presented some confusion. Despite some incorrectly classified polygons, the global accuracy is 0.96, meaning a good accuracy of the classification processes.

5. CONCLUSIONS

The temporal analysis of Guarujá from 1990 to 2021 indicates an increase in urban areas. Consequently, a decrease in vegetation is observed. The CART algorithm correctly distinguished the urban areas, the vegetation, and the water. Moreover, the NDVI provided important information about environmental degradation and loss of biomass. The advance of urban sprawl and land prices induce people to build houses in steep slope areas. The classification of Vila Baiana, using the orthophoto, allows us to discriminate the different covers in the neighborhood. Its identified low build standards, several constructions near each other, and no paved road. These areas are improper for construction, and anthropic changes induce disasters, mostly related to landslides.

8. REFERENCES

[1] C. M. D. Pinho, L. M. G. Fonseca, T. S. Korting, C. M. de Almeida, and H. J. H. Kux, "Land-cover classification of an intra-urban environment using high-resolution images and object-based image analysis," *Int. J. Remote Sens.*, vol. 33, no. 19, pp. 5973–5995, 2012.

[2] F. Guzzetti, S. Peruccacci, M. Rossi, and C. P. Stark, "Rainfall thresholds for the initiation of landslides in central and southern Europe," Meteorol. Atmos. Phys., vol. 98, no. 3–4, pp. 239–267, 2007.

[3] T. Novack, "Classificação da cobertura da terra e do uso do solo urbano utilizando o sistema InterIMAGE e imagens do sensor QuickBird.," p. 214, 2009.

[4] T. C. S. Rodrigues, "Classificação da cobertura e do uso da terra com imagens WorldView-2 de setores norte da Ilha do Maranhão por meio do aplicativo interIMAGE e de mineração de dados," *Diss. do Curso Pós-Graduação em Sensoriamento Remoto - INPE*, p. 108, 2014.

[5] M. G. H. Pechincha and R. T. Zaidan, "Zoneamento de Risco à ocorrência de escorregamentos: uma aplicação na bacia do Córrego Matirumbide, Juiz de Fora, MG," *Rev. Espinhaço*, vol. 4, no. 2, pp. 45–57, 2015.

[6] T. König, H. J. H. Kux, and R. M. Mendes, "Identificação De Áreas De Suscetibilidade a Escorregamentos De Encosta Utilizando O Modelo Matemático Shalstab," *Bol. Geogr.*, vol. 37, no. 3, pp. 228–243, 2020.

[7] T. S. Körting, L. M. Garcia Fonseca, and G. Câmara, "GeoDMA-Geographic Data Mining Analyst," *Comput. Geosci.*, vol. 57, pp. 133–145, 2013.

[8] E. V. Marcelino, L. H. Nunes, and M. Kobiyama, "Banco De Dados De Desastres Naturais: Análise De Dados Globais E Regionais," *Caminhos Geogr.*, vol. 7, no. 19, pp. 130–149, 2006.

[9] IBGE, "Cidades e estados do Brasil," 2019. .

[10] R. Araki, F. Sergio, and B. Ladeira, "Vulnerabilidade Associada a Precipitações E Fatores Antropogênicos No Município De Guarujá (Sp) – Período De 1965 a 2001," 2001.

[11] V. A. Pesck and A. A. Disperati, "Comparação de técnicas de fusão aplicadas à imagem Quickbird-2," *Floresta e Ambient.*, vol. 18, no. 2, pp. 127–134, 2011.

[12] S. P. Polizel, M. L. Marques, N. R. Costa, E. Rossi, and M. V. Ferreira, "Aplicação e avaliação de técnicas de fusão em imagens Ikonos e GeoEye," *An. do XV Simpósio Bras. Sensoriamento Remoto - SBSR, Curitiba - 2011*, no. 1988, pp. 447–451, 2011.

[13] R. Pu and S. Landry, "A comparative analysis of high spatial resolution IKONOS and WorldView-2 imagery for mapping urban tree species," *Remote Sens. Environ.*, vol. 124, pp. 516–533, 2012.

[14] G. T. Meneghetti and H. J. H. Kux, "Mapeamento Da Cobertura Da Terra Do Município De Raposa (Ma) Utilizando Imagens Worldview-Ii , O Aplicativo Interimage E Mineração De Dados Land Cover Mapping of Raposa (MA) Municipality Using WorldView-II Images , the InterIMAGE System and Data Minin," pp. 365–377, 2014.

[15] G. J. Hay and G. Castilla, "Object-based image analysis: Strengths, weaknesses, opportunities, and threats (SWOT)," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, pp. 4–5, 2006.

[16] I. Dronova, "Object-based image analysis in wetland research: A review," *Remote Sens.*, vol. 7, no. 5, pp. 6380–6413, 2015.

[17] M. K. B. Lüdeke, P. H. Ramge, and G. H. Kohlmaier, "The use of satellite NDVI data for the validation of global vegetation phenology models: Application to the frankfurt biosphere model," *Ecol. Modell.*, vol. 91, no. 1–3, pp. 255–270, 1996.

[18] J. Meng, X. Du, and B. Wu, "Generation of high spatial and temporal resolution NDVI and its application in crop biomass estimation," *Int. J. Digit. Earth*, vol. 6, no. 3, pp. 203–218, 2013.

[19] M. C. Modenesi-gauttieri and S. T. Hiruma, "A expansão urbana no planalto de campos do jordão. Diagnóstico geomorfológico para fins de planejamento," *Rev. do Inst. Geológico*, vol. 25, pp. 1–28, 2004.

[20] T. König, H. J. H. Kux, and R. M. Mendes, "Shalstab mathematical model and WorldView-2 satellite images to the identification of landslide-susceptible areas," *Nat. Hazards*, vol. 97, no. 3, 2019.

[21] R. M. Mendes, M. R. M. de Andrade, C. A. Graminha, C. C. Prieto, F. F. de Ávila, and P. I. M. Camarinha, "Stability Analysis on Urban Slopes: Case Study of an Anthropogenic-Induced Landslide in São José dos Campos, Brazil," *Geotech. Geol. Eng.*, vol. 36, no. 1, pp. 599–610, 2018.

[22] R. M. Mendes, M. R. M. D. Andrade, J. Tomasella, M. A. E. D. Moraes, and G. B. Scofield, "Understanding shallow landslides in Campos do Jordaõ municipality - Brazil: Disentangling the anthropic effects from natural causes in the disaster of 2000," *Nat. Hazards Earth Syst. Sci.*, vol. 18, no. 1, pp. 15–30, 2018.