Assessing Land Use and Cover Changes arising from the 2022 water crisis in Southeast China: A comparative analysis of Remote Sensing Imagery classifications and Machine Learning algorithms.

Eduardo Soares Nascimento [1], Fernanda Sayuri Yoshino Watanabe [2], Maria de Lourdes Bueno Trindade Galo [2], Erivaldo Antonio da Silva [2]

[1] Postgraduate Program in Cartographic Sciences (PPGCC), Department of Cartography, School of Technology and Sciences São Paulo State University (FCT-UNESP), 19060-900 Presidente Prudente, São Paulo, Brazil. e.nascimento@unesp.br
[2] Department of Cartography, School of Technology and Sciences, São Paulo State University (UNESP), São Paulo 19060-900, Brazil (fernanda.watanabe, trindade.galo, erivaldo.silva)@unesp.br

Keywords: Land Use and Land Cover, remote sensing imagery, machine learning, OLI/Landsat 8, random forest, and K-Nearest Neighbors Classifier.

Abstract

The water crisis in the southeast region of China in 2022, caused by one of the worst heatwaves on record, was characterized by severe shortages of water resources, leading to challenges for local communities, agriculture, and industry. To analyze changes in land use and land cover (LULC) in the Jialing River region, Chongqing, China, we compared Remote Sensing (RS) imagery classifications before and after the intense heat waves of 2022. We evaluated the performance of two machine learning algorithms, KDTree KNN and Random Forest (RF), in LULC classifications. The classifications were carried out based on the RS images from the OLI/Landsat 8 system, NDWI index, and SRTM data. The model performances were similar, the classification accuracy showed that the RF algorithm was superior to KDTree KNN. The RF LULC classification and area calculation corroborate with the visual analysis, reaffirming the superiority of RF, which shows a decrease in water surface area, unlike DKTree KNN.

1. Introduction

In 2022, China experienced one of the worst heatwaves on record. Starting on June 13, 2022, the Beijing Climate Center classified it as one of the most severe crises ever faced, considering the intensity of the heat, the geographical scale affected, and the prolonged duration. The precipitation deficit exacerbated the situation, as the insufficient amount of rain was not enough to meet the water needs of parts of the region. The decrease in the water surface area and the volume of the main river was one of the primary visual problems encountered.

It is widely recognized that data collection and imagery production through orbital sensors have become crucial for the identification and discrimination of objects on the Earth's surface. This is due to their unique characteristics, such as wide coverage area, short periods between revisits, and the free availability of some orbital system platforms. These factors allow a more comprehensive and detailed view of the Earth's surface, assisting in various applications, from environmental studies to decision-making in areas such as defense, agriculture, and natural resources (Novo, 2010).

Remote sensing is expected to play a role in monitoring environmental phenomena like heatwaves and water scarcity, leveraging its capabilities to provide invaluable insights for mitigating their impacts, monitoring, and enhancing resilience. Satellite imagery from remote sensing will be instrumental in analyzing changes in land use and land cover (Novo, 2010). This analysis will enable the quantification of these changes, essential for understanding environmental dynamics and facilitating informed decision-making toward resource management.

The classification of these images is a viable option for spatial analysis. It can be performed by using classical techniques of Digital Images Processing or through machine learning, with more advanced techniques. Monard and Baranauskas (2003) Conceptualize Machine/Deep Learning as a system that acquires knowledge automatically, capable of making decisions based on the understanding acquired from successful solutions to past problems. It is known that classification algorithms can be divided into two main groups: supervised (i) and unsupervised (ii). In (i), training samples are provided, whose labels correspond to the class to which they belong (Monard and Baranauskas, 2003). This differs from algorithms (ii), which do not require pre-defined labels for the input set during the learning process. In this study, two supervised Machine Learning classifiers were compared to evaluate them and investigate the water crisis in the southeastern region of China.

The land use and land cover (LULC) classification is a process of assigning land occupation classes to pixels and categorizing them, meaning grouping the pixels of the image into land occupation classes (Alshari; Gawali, 2021). Land use refers to the purpose that the land serves, while land cover refers to the surface cover of the land, whether vegetation, soil, water, or other elements (Rajendran et al., 2020). LULC assessment is necessary to sustain, monitor, and plan the utilization of natural resources (Nayak; Mandal, 2019 and Singh et al., 2020). The LULC classification has a direct impact on the atmosphere, soil erosion, and water, while indirectly being linked to global environmental issues. In recent decades, machine learning techniques have dominated conventional classification methods used for LULC classification in remote sensing (Saini; Rawat, 2023).

This paper presents the results obtained from a comparison of two machine learning techniques, KDTree k-nearest neighbors

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-253-2024 | © Author(s) 2024. CC BY 4.0 License.

(KNN), and Random Forest, for the change detection classification map of the interest region. It compares the outcomes of applying machine learning algorithms for the LULC classification from multispectral images and SRTM data. The results from image classification are expected to help quantify the changes in LULC in the region affected by the heatwave.

In this context, our goal was to assess the effectiveness of various machine learning algorithms in LULC classification, aimed at analyzing the alterations witnessed in the Jialing River region, Chongqing, China, stemming from the 2022 water crisis.

2. Materials and methods

The study area is located at the confluence of the Yangtze and Jialing rivers in southwest China. In mid-August 2022, China was experiencing one of the worst heatwaves ever recorded. According to the Beijing Climate Center monitoring, the phenomenon began on June 13 and is considered one of the most severe, considering the intensity of the heat, the geographical area affected, and the duration. Consequently, one of the hardest-hit regions was Chongqing, in the southwest of the country. The city recorded temperatures of up to 45°C and experienced 11 days with temperatures exceeding 40°C.

2.1 Definition of classes and interpretation key

According to Florenzano (2002) and Novo (2010), the interpretation key has the main objective of characterizing the features of interest, to facilitate the identification of other similar features in the image. Based on this, interpretation elements are established, such as tonality, shape, and texture, among others. In Figure 1, we have the scene of interest captured from Google Earth software, which enabled the identification of the classes present in the scene. Figure 1 shows the different classes present in the scene of interest: Buildings, Vegetation, Water, Roads/Streets, and Bare Soil.

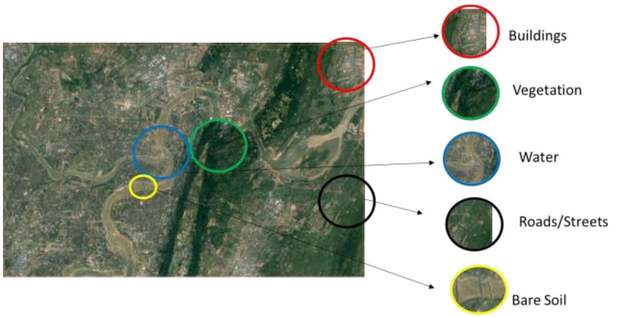


Figure 1. Scene captured from Google Earth of the Study Area.

2.2 Data

The OLI/Landsat 8 images from August 6, 2020, were selected for the period before the water crisis, with path/row at 128/039, processing level L2. The period after the water crisis selected the day August 12, 2022, from the same region, at the same level. According to the classes defined in section 2.1, it is known that to discriminate them, bands from the visible spectrum (Costal Blue/Aerosol "B1", Blue "B2", Green "B3", and Red "B4") are necessary, as shape, color, and texture information are contained in these bands. Additionally, it is important to note that bands from the infrared spectrum (Near Infrared "B5", Short-wave infrared 1 "B6", and Short-wave infrared 2 "B7") effectively discriminate water from other targets. To assist in the discrimination of the water target, the Normalized Difference Water Index (NDWI), proposed by McFeeters (1996) (Equation 1), was used.

$$NDWI = \frac{\rho Green - \rho NIR}{\rho Green + \rho NIR} \quad (1)$$

Given that:
$\rho Green$ = Reflectance in the green spectrum;
ρNIR = Reflectance in the Near Infrared spectrum.

The use of this index is justified as it is intended for the analysis and assessment of water resources and flooded areas to highlight the delineation of water features.

In addition to the data mentioned above, an SRTM (Shuttle Radar Topography Mission) digital elevation model with 30 m spatial resolution was also used, officially distributed for free by the United States Geological Survey (USGS) through the Earth Explorer portal. SRTM images are widely used in geomorphology, in which different textures represent varied relief domains, as well as in the information that can be extracted through computational processing. Thus, this type of image can be essential in discriminating spectrally similar objects but with different textures, such as water, which has a smooth texture, and vegetation, which has a rough texture. While it is true that altitude alone is less sensitive to texture variations compared to its derivatives (such as slope, aspect, and entropy), it still provides valuable information for specific applications, particularly in identifying water bodies. The continuous behavior of the altitude variable offers significant assistance in distinguishing water bodies from other land cover types. Therefore, the set of attributes to be used for the selected periods of 2020 and 2022 was established. The scene of interest in the false-color composite (5R-4G-3B) can be seen in Figure 2.

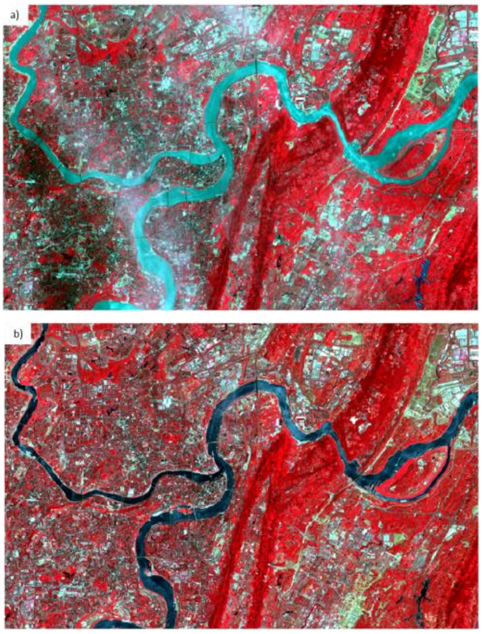


Figure 2. False Color (5R-4G-3B). a) in 2020; b) in 2022.

In Figure 2, the scenes of interest can be visualized. As known, data from SRTM and the NDWI spectral index were used to discriminate the classes and perform the stacking. Figures 3 and 4 show the input data set for the classifiers. All the bands have been stacked (Coastal Blue, Blue, Green, Red, Near Infrared, Medium Infrared 1, Medium Infrared 2, NDWI, and SRTM image). In a), the False-color composite 5R-4G-3B, b) the True-color composite 4R-3G-2B, c) the SRTM digital model, and d) the NDWI index. The a) and b) are just color compositions that represent the stacking of bands.

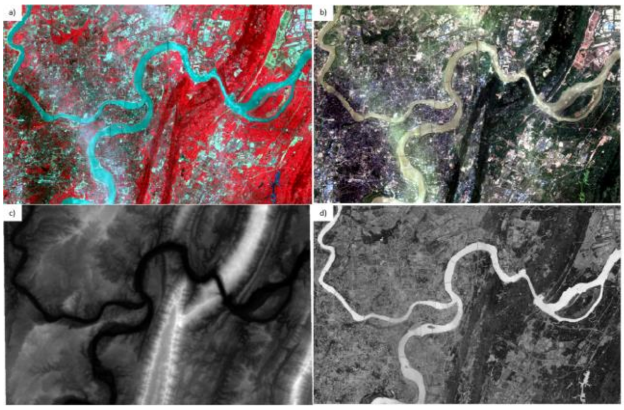


Figure 3. Stacked data for 2020. a) False-color composite 5R-4G-3B, b) True-color composite 4R-3G-2B, c) SRTM digital model, and d) NDWI index.

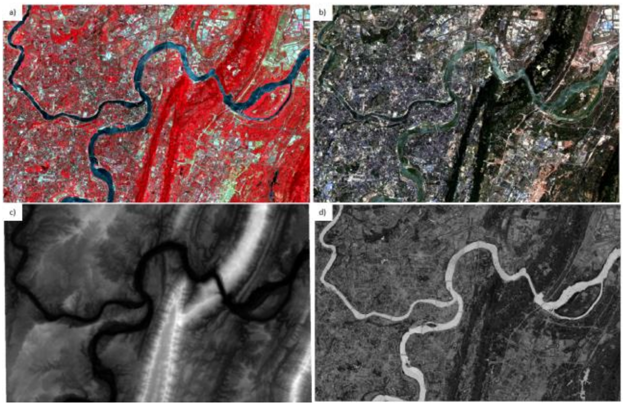


Figure 4. Stacked data for 2022. a) False-color composite 5R-4G-3B, b) True-color composite 4R-3G-2B, c) SRTM digital model, and d) NDWI index.

It is worth noting that the stacked data for 2020 and 2022 includes the following bands: Coastal Blue, Blue, Green, Red, Near Infrared, Medium Infrared 1, Medium Infrared 2, NDWI, and SRTM image. As observed in Figure 3, a decrease in the river's watercourse is visible. In Figures 3 and 4, item d, such change becomes more evident between the two periods.

2.3 Machine Learning Algorithms

2.3.1 Random Forest: The Random Forest algorithm, as described by Breiman (2001), is a classification method that utilizes multiple Decision Trees. Each tree is built from a random sample of the training data and a subset of attributes, and predictions are made by aggregating the results of all trees. This method is effective for handling large datasets and attributes, showing quick learning compared to other algorithms (Gao et al. 2009). The algorithm consists of five main steps:

1. Inputting training samples and labels, along with defining hyperparameters such as the number of trees (Ntree) and the number of attributes to evaluate (Mtry).

2. Random selection of samples to ensure diversity in tree construction, with the number of attributes considered in each tree defined by the parameter Mtry.

3. Selection of relevant variables based on the Gini Index or Entropy criteria, in which variables with the least variation or confusion are chosen for each node.

4. Creation of a forest composed of the defined number of trees (Ntree), with 1/3 of samples reserved for validation (Out-Of-Bag) and the remaining 2/3 used for training.

5. Classification of pixels based on the majority class vote, adjusting tree parameters based on predefined results and generating accuracy parameters after training.

2.3.2 KDtree KNN Classifier: The K-Nearest Neighbors Classifier (KNN) is a supervised learning algorithm that relies on the principle that similar points tend to belong to the same class. It utilizes the KDTree data structure to accelerate the search for nearest neighbors in large datasets. KDTree organizes points in a multidimensional space through a binary decision tree, with each node representing a split based on one dimension. The KNN Classifier with KDTree begins by selecting the number K of nearest neighbors for classification. Distances between the point of interest and all other points are calculated and stored in the KDTree. The K nearest points are identified through the KDTree search, and the most common class among them is assigned to the point of interest. KDTree enables faster neighbor search, enhancing KNN efficiency in large datasets. The algorithm's output is a class association, determined by the majority vote among the K nearest neighbors. Normalizing training data can improve accuracy, especially when features have different scales. Weighting neighbor contributions based on distance is also common, with closer neighbors having more influence. KNN is sensitive to the local data structure, emphasizing the importance of training data representation. KNN algorithm's sensitivity to local data structure arises from the distance metric used to identify the nearest neighbors. This dependency on local distances makes the algorithm highly responsive to how data points are arranged in the feature space, influencing its performance based on the local patterns present in the training data. For this algorithm, a division of 2/3 of the samples for training (training test split) and 1/3 for validation was utilized (test size).

2.3.3 Homogenization of classes: The Random Forest and KDtree KNN classifiers are pixel-based classification algorithms, which generally exhibit noises, including isolated pixels, misclassified pixels, or inconsistently classified pixels. The application of the Crivo function (filtering, based on the removal of small isolated areas within a neighboring area, using the convolution of a 3x3 window, where neighbors are analyzed to the central pixel, and the most frequent value is assigned to the central pixel, allowing for the homogenization of the thematic map), allow the correction of errors that were encountered. This function is implemented as a plugin in QGIS software.

2.3.4 Acquisition of Training and Validation Samples: To collect training samples for machine learning algorithms in remote sensing, one must define and characterize the information classes (land cover types) using interpretation keys.

These keys help identify features of interest in images based on elements such as color, shape, texture, and structure. Additionally, associations through deduction, induction, and analogy are used to relate the properties of objects to their surroundings. Interpretation keys are constructed to organize information and aid in the visual interpretation of images, leveraging tools like Google Earth Pro and updated satellite imagery to identify and characterize features accurately, thereby facilitating the acquisition of training data for classification.

The samples were acquired in shapefile format in QGIS software. The polygons (Region of Interest - ROI) were selected according to the stacked bands. Two sets of samples were selected, one for the period in 2020 and one in 2022. Approximately 300 samples were acquired, with approximately 50 for each class. Given that urban areas present a wide diversity of characteristics and variability in the class, it is important to collect a significant amount of samples for training purposes to ensure that the entire class variety is adequately represented.

For the Random Forest algorithm, the hyperparameter values used for model validation involved splitting the dataset into two parts: 1/3 of the dataset, consisting of 100 samples, was reserved for validation (Ntree parameter), while the remaining 2/3, comprising 200 samples, were used for model training (Mtry parameter). For the KDTree KNN algorithm, the same approach was used, with a split of 200 samples for training and 100 for testing. Below is an example of the samples taken to train and validate the algorithm.

In Figures 5 and 6, the features of interest collected as training and validation samples were identified and highlighted, aiming to capture the full variability of the classes and represent them when the rest is classified. The training and validation samples in shapefile format were exported for opening in the Snap software, where supervised classifications were performed.

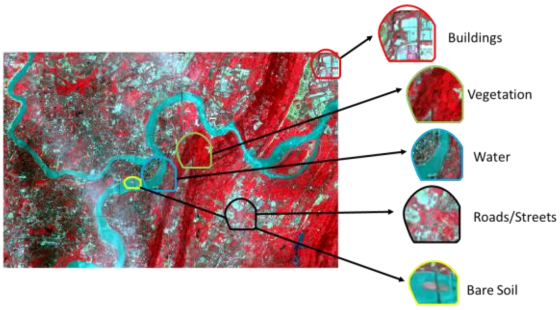


Figure 5. Example of samples collected for training and validation for each class in 2020.

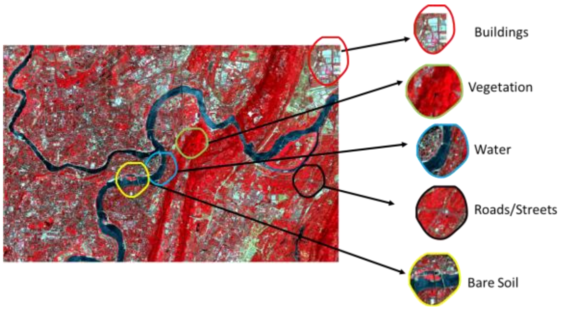


Figure 6. Example of samples collected for training and validation for each class in 2022.

2.3.5 Model performance and Classification Accuracy: For the two algorithms used in this paper, the comparative metric used was the Root Mean Square Error (RMSE), which is already implemented in the Snap software. It is calculated from the square root of the average of the squared difference (error) between the classification result and the ground truth, ranging from zero to infinity, and the closer to 0, the better the dataset for the model in question. For both algorithms, a total of 100 samples were utilized, as previously mentioned, for model validation.

For classification accuracy, the sample size was calculated using Equation 2, proposed by Fitzpatrick-Lins (1981), which is based on binomial statistics, allowing a number of samples to depend on a specified level of significance.

$$n_0 = \frac{(Z_{\frac{\alpha}{2}})^2 p(1-p)}{d^2} \tag{2}$$

where n_0 is the initial sample size, $Z_{\frac{\alpha}{2}}$ is the confidence level (normal distribution), p is the minimum desired accuracy level for the cartographic product (binomial distribution), and d is the maximum allowable error. For this work, a confidence level of 95% was defined $(Z_{\frac{\alpha}{2}} = 1{,}96)$, accuracy level of 0.15, and a maximum error of 0.05, resulting in 196 sample elements.

Additionally, Fitzpatrick-Lins (1981) emphasizes that the choice of samples should not be biased. Therefore, the ACATAMA plugin implemented in the QGIS software ensures that the samples are non-aligned stratified, combining a random scheme (with low trends) with priority for greater geographical coverage (systematic and stratified). This sample design is the most suitable for estimating classification accuracy.

The confusion matrix can be defined as a square matrix, where the columns represent the ground truth and the rows indicate the generated classification, serving as an important tool for representing the accuracy of each category. When a pixel is correctly classified, 1 is added to a particular position on the main diagonal of the matrix; otherwise, 1 is added to the position defined by the classified category (row) x ground truth (column) (Congalton, 1991). We observe the confusion matrix in Table 1.

		j (columns) Ground truth		
		I	j	Total X_{i+}
i (rows) **classificaton**	i	X_{ii}	X_{ij}	X_{i+}
	J	X_{ji}	X_{jj}	X_{j+}
	Total X_{+i}	$X_{+\,i}$	$X_{+\,j}$	X

Table 1. Confusion Matrix.

Where:
i and *j* are classes;
X_{ii} is the number of rows and columns in the matrix;
X_{jj} is the number of rows and columns in the matrix;
X_{i+} e $X_{+\,i}$ are the marginal totals of row *i* and column *i*;
X_{j+} e $X_{+\,j}$ are the marginal totals of row *j* and column *j*;
X is the total number of observations.

From the confusion matrix, the User Accuracy (UA), Producer Accuracy (PA), Overall Accuracy (OA), and the Kappa Coefficient (K) can be calculated (Equations 3, 4, 5, and 6, respectively). The first one refers to the estimates of the fractions of mapped pixels for each class. The second one refers to the sample fractions of pixels of each class correctly assigned to their classes by the classifier. The third one is the ratio of the number of points classified correctly to the total number of sample points used. Finally, the fourth one is the measurement of concordance, i.e., the measure of the difference between the actual agreement of the classification, represented by the values on the diagonal of the confusion matrix, and the chance agreement, which is given by the product of the marginal values of the rows and columns. Below are the equations described.

$$UA_i = \frac{x_{ii}}{x_{+\,i}} \tag{3}$$

$$PA_i = \frac{x_{ii}}{x_{i+}} \tag{4}$$

$$OA = \sum_{i=1}^{k} \frac{x_{ii}}{\mathrm{X}} \tag{5}$$

$$K = \frac{x * \sum_{i=1}^{k} X_{ii} - x * \sum_{i=1}^{k} (x_{i+} * x_{+i})}{x^2 - \ (x_{i+} * x_{+i})} \tag{6}$$

2.3.6 Methodological Flowchart: Based on the items defined and highlighted in the previous sections, Figure 7 presents the general workflow of this work. It began with the download of images from the Earth Explorer website for the two selected periods. Level L2 images were chosen, and the NDWI index was generated, along with the download of the SRTM digital model. The best attributes for representing the classes present in the scene of interest were selected, considering the best attributes for representation and discrimination of the classes of interest. Stacking was performed for each period. Then, the interpretation key was generated to assist in the process of collecting training samples. These samples were input into the classifier along with the stacked bands, resulting in thematic maps that underwent accuracy analysis and change analysis considering the two periods. The RMSE values obtained from the classification of the two algorithms were compared, and a confusion matrix was generated to evaluate the accuracy of the thematic map. After the thematic maps were created for both periods, the areas were calculated in square kilometers using the QGIS software. Subsequently, the areas were subtracted, comparing the two periods in each classifier. These additional steps provided a more detailed analysis of the changes over time and contributed to a comprehensive understanding of the results obtained. Figure 7 provides a summarized overview of the general methodology presented so far.

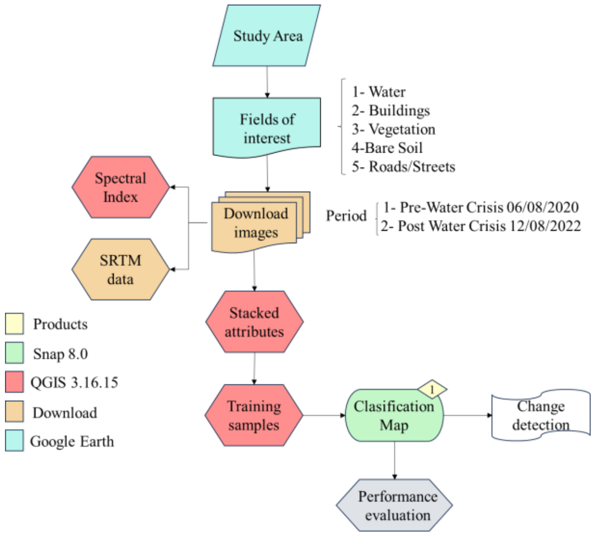


Figure 7. Flowchart

3. Results and discussions

3.1 Land use and Land Cover Classification

The results of the LULC classification by the Random Forest and KDtree KNN algorithms can be seen in Figures 8, 9, 10, and 11. All of them have the homogenization function applied, as mentioned in section 2.3.3.

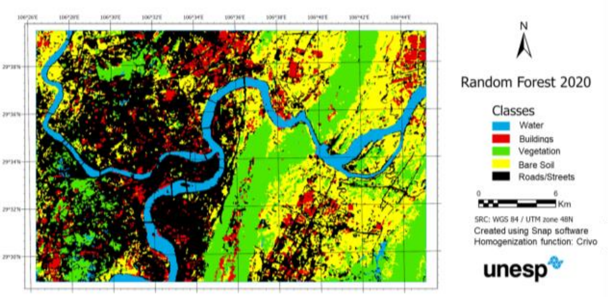

N
Random Forest 2020
Classes
Water
Buildings
Vegetation
Bare Soil
Roads/Streets
0
6
Km
SRC: WGS 84 / UTM zone 48N
Created using Snap software
Homogenization function: Crivo
unesp

Figure 8. Thematic map of land use and land cover, using Random Forest in the year 2020.

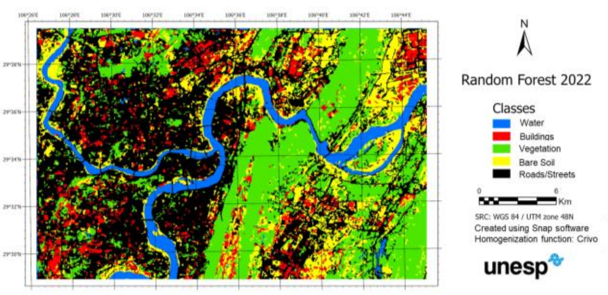

N
Random Forest 2022
Classes
Water
Buildings
Vegetation
Bare Soil
Roads/Streets
0
6
Km
SRC: WGS 84 / UTM zone 48N
Created using Snap software
Homogenization function: Crivo
unesp

Figure 9. Thematic map of land use and land cover, using Random Forest in the year 2022.

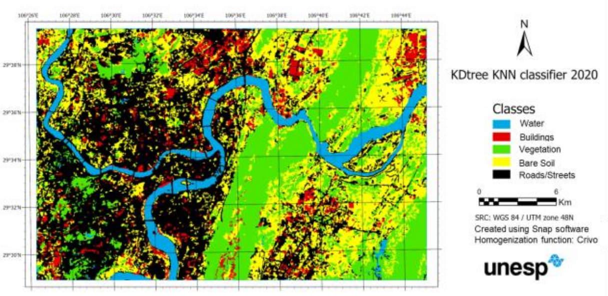

N
KDtree KNN classifier 2020
Classes
Water
Buildings
Vegetation
Bare Soil
Roads/Streets
0
6
Km
SRC: WGS 84 / UTM zone 48N
Created using Snap software
Homogenization function: Crivo
unesp

Figure 10. Thematic map of land use and land cover, using KDtree KNN Classifier in the year 2020.

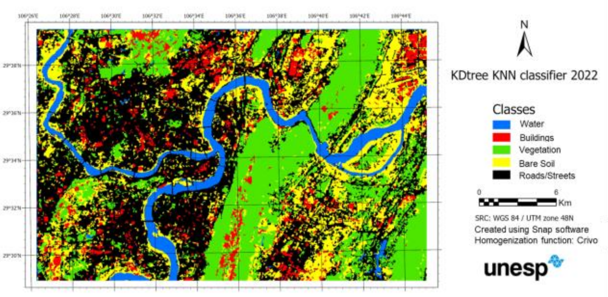

N
KDtree KNN classifier 2022
Classes
Water
Buildings
Vegetation
Bare Soil
Roads/Streets
0
6
Km
SRC: WGS 84 / UTM zone 48N
Created using Snap software
Homogenization function: Crivo
unesp

Figure 11. Thematic map of land use and land cover, using KDtree KNN Classifier in the year 2022.

Visually analyzing the data, it is evident that there is significant confusion among the road class and other classes, both in 2020 and 2022. This confusion results in irregular and inconsistent representations of roads/streets, depicted as black areas, which contradicts the regular nature of the road/streets class. Furthermore, the Building class exhibits random sizes and discontinuities, while the Vegetation and Bare Soil classes cover a large portion of the image. The class with the clearest visual definition is Water, which is accurately delineated in both periods, showing some discontinuities caused by drought when compared between the time frames.

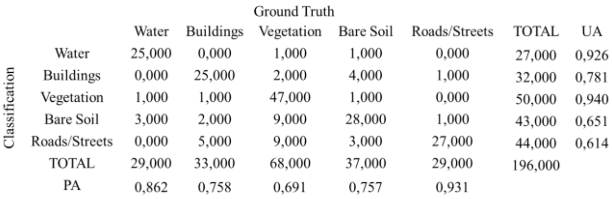
3.2 Model Performance and Classification Accuracy

Table 2 shows the root mean square error values for each classifier in their respective periods. It can be observed that the training samples for the year 2020 provided the best model definition compared to the samples from 2022, which yielded higher values for both classifiers. Additionally, the algorithm determines the relevance of the attributes in distinguishing the classes. For the Random Forest algorithm, the NDWI and SRTM features were the most important, whereas the Coastal Blue band was evaluated as the least important. On the other hand, in the KDtree KNN algorithm, the Near Infrared and SRTM features were classified as the most significant, while the Coastal Blue band also was considered the least influential in the process.

Algorithm	RMSE 2020	RMSE 2022
Random Forest	0,311	0,455
KDtree KNN	0,385	0,546

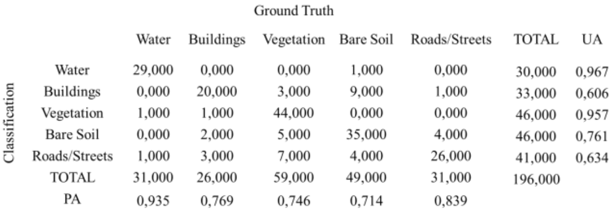
Table 2. Models performance.

The estimation of accuracy for the generated products was calculated to highlight which classifier yielded better results regarding classification. Initially, a total of 196 samples were defined, as shown in section 2.3.5. Thus, it was possible to generate the confusion matrix for classification. In Tables 3 and 4, the confusion matrix for the random forest algorithm in 2020 and 2022, respectively, is presented. In Tables 5 and 6, the confusion matrix for the KDTree KNN algorithm in 2020 and 2022, respectively, is provided.



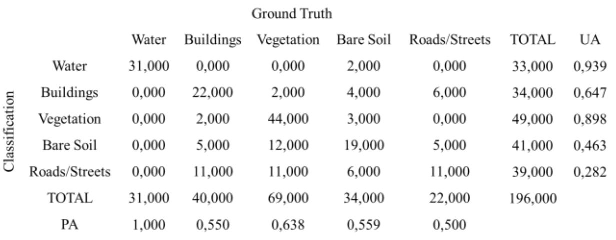
		Ground Truth						
		Water	Buildings	Vegetation	Bare Soil	Roads/Streets	TOTAL	UA
Classification	Water	25,000	0,000	1,000	1,000	0,000	27,000	0,926
	Buildings	0,000	25,000	2,000	4,000	1,000	32,000	0,781
	Vegetation	1,000	1,000	47,000	1,000	0,000	50,000	0,940
	Bare Soil	3,000	2,000	9,000	28,000	1,000	43,000	0,651
	Roads/Streets	0,000	5,000	9,000	3,000	27,000	44,000	0,614
	TOTAL	29,000	33,000	68,000	37,000	29,000	196,000	
	PA	0,862	0,758	0,691	0,757	0,931		

Table 3. Confusion Matrix for the Random Forest classifier in the year 2020.



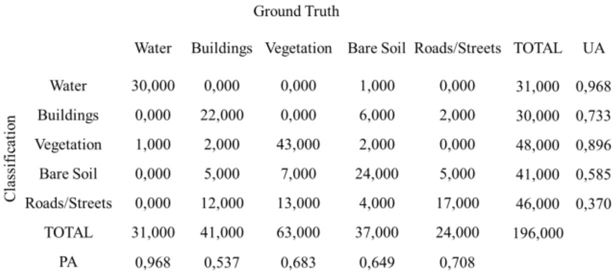
		Ground Truth						
		Water	Buildings	Vegetation	Bare Soil	Roads/Streets	TOTAL	UA
Classification	Water	29,000	0,000	0,000	1,000	0,000	30,000	0,967
	Buildings	0,000	20,000	3,000	9,000	1,000	33,000	0,606
	Vegetation	1,000	1,000	44,000	0,000	0,000	46,000	0,957
	Bare Soil	0,000	2,000	5,000	35,000	4,000	46,000	0,761
	Roads/Streets	1,000	3,000	7,000	4,000	26,000	41,000	0,634
	TOTAL	31,000	26,000	59,000	49,000	31,000	196,000	
	PA	0,935	0,769	0,746	0,714	0,839		

Table 4. Confusion Matrix for the Random Forest classifier in the year 2022.



		Ground Truth						
		Water	Buildings	Vegetation	Bare Soil	Roads/Streets	TOTAL	UA
Classification	Water	31,000	0,000	0,000	2,000	0,000	33,000	0,939
	Buildings	0,000	22,000	2,000	4,000	6,000	34,000	0,647
	Vegetation	0,000	2,000	44,000	3,000	0,000	49,000	0,898
	Bare Soil	0,000	5,000	12,000	19,000	5,000	41,000	0,463
	Roads/Streets	0,000	11,000	11,000	6,000	11,000	39,000	0,282
	TOTAL	31,000	40,000	69,000	34,000	22,000	196,000	
	PA	1,000	0,550	0,638	0,559	0,500		

Table 5. Confusion Matrix for the KDtree KNN classifier in the year 2020.



		Ground Truth						
		Water	Buildings	Vegetation	Bare Soil	Roads/Streets	TOTAL	UA
Classification	Water	30,000	0,000	0,000	1,000	0,000	31,000	0,968
	Buildings	0,000	22,000	0,000	6,000	2,000	30,000	0,733
	Vegetation	1,000	2,000	43,000	2,000	0,000	48,000	0,896
	Bare Soil	0,000	5,000	7,000	24,000	5,000	41,000	0,585
	Roads/Streets	0,000	12,000	13,000	4,000	17,000	46,000	0,370
	TOTAL	31,000	41,000	63,000	37,000	24,000	196,000	
	PA	0,968	0,537	0,683	0,649	0,708		

Table 6. Confusion Matrix for the KDtree KNN classifier in the year 2022.

In Table 3, a kappa index of 0.71 and an OA value of 0.77 were observed. It was noted that vegetation had the highest number of confusions, with approximately 69.1% correct predictions (PA), with only 47 pixels classified correctly out of the total 68 for this class. Confusion occurred with the exposed soil and road classes (9 pixels each), and 2 pixels for buildings and 1 for water bodies.

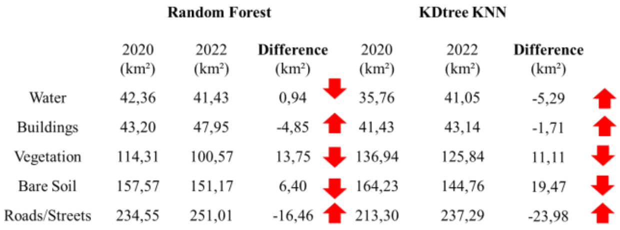
For Table 4, a kappa index value of 0.72 and an OA of 0.78 were obtained. It was also observed that the Bare Soil had only 71.4% correct predictions, followed by vegetation with 74.6%, and the buildings class with approximately 76.9%.

In Table 5, a kappa index value of 0.55 and an OA of 0.64 were obtained. This classifier, when applied with the sieve, resulted in many confusions compared to others, with almost all classes except water obtaining prediction values below 63.8%. This is attributed to the fact that KDtree KNN and the sieve function both work with neighbors, leading to erroneous classification of objects of interest when applying a homogenization function, as the algorithm already implicitly applies homogenization.

Similarly, in Table 6, a kappa index of 0.61 and an OA of 0.69 were obtained. Comparing the two algorithms, it is evident that Random Forest presented better metrics in both periods compared to the KDtree KNN algorithm. Thus, Random Forest more accurately represents the features of interest for the given area.

3.3 Change detection

In Table 7, the areas in km² of changes between the periods of 2020 and 2022 are presented. The areas were calculated using the QGIS software.



	Random Forest				KDtree KNN			
	2020 (km²)	2022 (km²)	**Difference** (km²)		2020 (km²)	2022 (km²)	**Difference** (km²)	
Water	42,36	41,43	0,94	⬇	35,76	41,05	-5,29	⬆
Buildings	43,20	47,95	-4,85	⬆	41,43	43,14	-1,71	⬆
Vegetation	114,31	100,57	13,75	⬇	136,94	125,84	11,11	⬇
Bare Soil	157,57	151,17	6,40	⬇	164,23	144,76	19,47	⬇
Roads/Streets	234,55	251,01	-16,46	⬆	213,30	237,29	-23,98	⬆

Table 7. Area in km² for the studied period, considering both machine learning algorithms.

It is notable that through visual analysis alone, it is possible to observe a decrease in water volume as the severity of the water crisis unfolds, as confirmed by the Random Forest algorithm, illustrated with the upward-pointing red arrow in the "Difference" column. Conversely, the KDtree KNN algorithm yielded values indicating an increase in water area across during the period, which contradicts the reality of the water crisis experienced in the region. Similar to the accuracy metrics of the classification pointing to the Random Forest classifier as superior for this problem, the calculation of the area reaffirms this result, as it is evident that the water volume has decreased.

Remote sensing and Land Use and Land Cover (LULC) classification are effective tools for monitoring the impacts of drought. The study demonstrates that the Random Forest algorithm outperforms the KDtree KNN in terms of accuracy and consistency, particularly in detecting changes in water volume, a crucial indicator of drought. Visual analysis and accuracy metrics, such as the Kappa Index and Overall Accuracy, confirm the superior performance of the Random Forest algorithm in identifying land cover changes. The significance of features like NDWI and SRTM for enhancing classification accuracy is also noted. This study emphasizes the importance of employing advanced remote sensing techniques and robust algorithms for environmental monitoring and managing the impacts of climate change.

4. Conclusion

LULC mapping was achieved successfully through Radom Forest algorithm. Visual analysis alone indicated that the Random Forest algorithm yielded superior results. This was further confirmed by the Kappa Index and OA metrics obtained from the confusion matrix, as well as by comparing each class classified to the ground truth. However, the KNN algorithm did not perform as well as expected and did not achieve success in accurately representing the intended outcomes.

This aspect was further verified by calculating the area of each class and comparing them between epochs. The Random Forest algorithm proved to be more accurate in detecting changes in the area of interest from one epoch to another.

The KDtree KNN algorithm classifies using information from neighbors, which may lead to excessive generalization when subjected to the thematic homogenization function (Crivo), resulting in incorrect and inconsistent associations between classes. In contrast, Random Forest exhibits superior performance as it does not rely on neighborhood information, thus avoiding such issues.

Regarding the model accuracy, it can be affirmed that the samples collected in 2020 were the most effective in distinguishing between classes, as evidenced by their lower RMSE values compared to those from 2022.

The situation regarding the Jialing River is alarming, as demonstrated by the reduction in water volume observed through visual analysis and quantified by the difference in river area depicted in Table 7, utilizing the Random Forest algorithm. The diminishing river area serves as a poignant reminder of the environmental challenges facing the region, underscoring the imperative need to safeguard water resources and address the impacts of climate change.

REFERENCES

Alshari, E. A.; Gawali, B. W. Development of classification system for LULC using remote sensing and GIS. *Global Transitions Proceedings*, v. 2, n. 1, p. 8–17, 2021. Available in: <https://www.sciencedirect.com/science/article/pii/S2666285X21000029>.

Breiman, L. Random Forests., v. 45, p. 5–32, 2001. Access on: 16/5/2022.

Congalton, R. G. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, v. 37, n. 1, p. 35–46, 1991. Elsevier. Access on: 28/8/2022.

Fitzpatrick-Lins, K. Comparison of Sampling Procedures and Data Analysis for a Land-Use and Land-Cover Map. 1981. Access on: 28/8/2022.

Florenzano, T. G. Imagens de Satélite para Estudos Ambientais. São Paulo, 2002.

Gao, D.; Zhang, Y.-X.; Zhao, Y.-H. Random forest algorithm for classification of multiwavelength data. *Astron. Astrophys*, v. 9, n. 2, p. 220–226, 2009. Available in: <http://www.raa-journal.orghttp://www.iop.org/journals/raa>. Acesso em: 16/5/2022.

Monard, M. C.; Baranauskas, J. A. Conceitos sobre aprendizado de máquina. Sistemas inteligentes-Fundamentos e aplicações, v. 1, n. 1, p. 32, 2003.

Nayak, S.; Mandal, M. Impact of land use and land cover changes on temperature trends over India. *Land Use Policy*, v. 89, p. 104238, 2019. Available in: <https://www.sciencedirect.com/science/article/pii/S0264837719300407>.

Novo, E. M. L. de M. Sensoriamento Remoto: Princípios e Aplicações. 4º ed. 2010.

Rajendran, G. B.; Kumarasamy, U. M.; Zarro, C.; Divakarachari, P. B.; Ullo, S. L. Land-Use and Land-Cover Classification Using a Human Group-Based Particle Swarm Optimization Algorithm with an LSTM Classifier on Hybrid Pre-Processing *Remote-Sensing Images*. **Remote Sensing**, v. 12, n. 24, 2020. Available in: <https://www.mdpi.com/2072-4292/12/24/4135>.

Saini, R.; Rawat, S. Land Use Land Cover Classification in Remote Sensing Using Machine Learning Techniques. 2023 1st International Conference on Innovations in *High-Speed Communication and Signal Processing* (IHCSP). Annals. p.99–104, 2023.

Singh, S.; Bhardwaj, A.; Verma, V. K. Remote sensing and GIS based analysis of temporal land use/land cover and water quality changes in Harike wetland ecosystem, Punjab, India. *Journal of Environmental Management*, v. 262, p. 110355, 2020. Available in: <https://www.sciencedirect.com/science/article/pii/S0301479720302905>.