RECONSTRUCTION OF WATER MASK FROM OPTICAL CLOUDY LANDSAT SCENES USING NAIVE BAYES ALGORITHM

Vinícius Kuchinski[1], *Rodrigo Cauduro Dias de Paiva*[2]

[1]PhD Candidate at IPH/UFRGS, eng.vinicius@outlook.com, [2]IPH/UFRGS, rodrigocdpaiva@gmail.com

ABSTRACT

Floods are natural events that can generate major disasters. Remote sensing data can be used for mapping flood events. However, atmospheric conditions in images captured by optical sensors, such as the presence of clouds, affect the recorded data and impair the use of these images in hydrological studies. In this context, a machine learning algorithm – Naive Bayes – was used to determine the extent of water masks under two hypothetical cloud cover. The results, compared to the observed data, indicated that the proposed methodology could be used for the determination of water masks, since the performance metrics reached values for accuracy above 0.98 and for the critical success index above 0.70. Further tests and validations should be performed to prove the potential of reconstruction of water masks by the proposed methodology.

Key words – *flood, inundation, machine learning, remote sensing*

1. INTRODUCTION

Flooding is a natural event being the most frequent type of disaster [1]. Some tools, such as maps, help the spatial representation of floods [2] , however for the majority of places there are no sufficient information to determine the extension of these events.

Remote sensing is an important tool for mapping floods because of the ability to observe water resources on large scale, with high temporal coverage and reasonable accuracy [3]. Optical sensors such as those embedded in *Landsat* satellites have been proven capable of extracting water masks from flood events [4], [5]. However, despite the proven ability, information obtained from these sensors are highly affected by atmospheric conditions such cloud presence or particulate matter, for example [3].

The objective of this work is to reconstruct water masks from images, whose data were obtained by optical sensors with the presence of clouds, using the machine learning algorithm Naive Bayes (NB). To achieve this, a flood event without the presence of clouds was used as a benchmark and two other images with the presence of clouds were used to apply a cloud mask to the benchmark image. Finally, predicted values for the pixels under the cloud mask were compared with the original values to verify the accuracy of the reconstruction.

2. MATERIAL E METHODS

This work is part of a research in a larger context on flood mapping. Because of this, some previously performed analyses will not be described in the methodology.

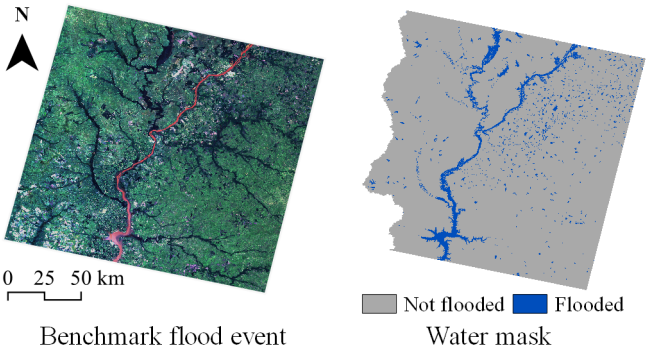


Figure 1: Study area presenting the benchmark image (left) and the water mask extracted by mNDWI (right).

2.1. Study area

The study area of this work is a reach of Uruguai river (Figure 1, left), located in the west portion of the Rio Grande do Sul state, southern Brazil and South America. The Uruguai river makes the border of two states of Brazil and also Brazil, Argentina and Uruguay.

The region is hydrologically representative to southern Brazil, since the Uruguai river is the largest in extesion in the Rio Grande do Sul state. In the river reach under analysis, due its topographic carachteristics, flood events have been recurrently recorded.

2.2. Data used

Five features were used as input data to the model: digital surface model (DSM), land use and land cover (LULC), height above nearest drainage (HAND), frequency of flood occurrence (FO) and discharge's return period (QRP). These features were selected because of their significance in the determination of flood inundation map by the NB model in a previously performed analysis.

With the exception of QRP, the features were obtained from the Google Earth Engine (GEE) cloud computing platform by the application programming interface (API) available for GEE use in *Python* programming language. The data used are shown in Fig. 2

Elevation is a basic factor that influences the intensity of the flood inundation [6]. For its representation was chosen the digital surface model ALOS DSM [7], with spatial resolution of approximately 30 meters.

HAND normalizes the topography according to the drainage network [8], representing in a complementary way the distance of the analysis point to the drainage [9]. The HAND selected is from MERIT Hydro [10], which is a global hydrographic dataset with spatial resolution of 90 meters.

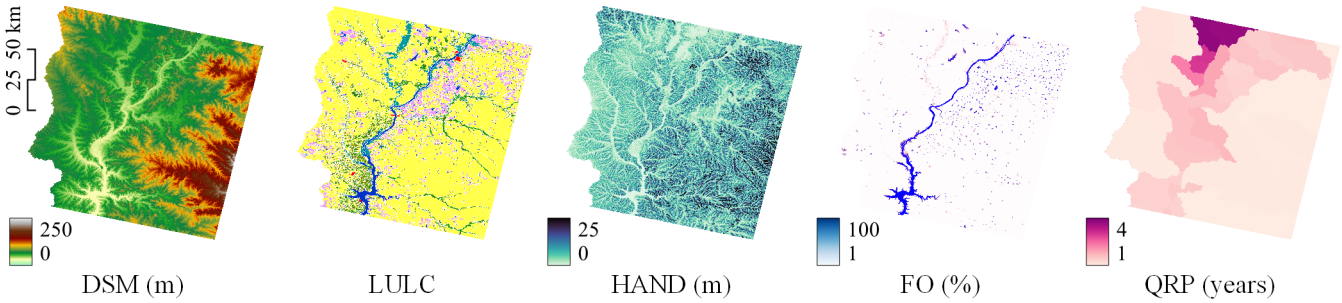


Figure 2: Data used as input to the NB model.

LULC is an important factor to identify flood-susceptible areas: vegetated areas tend to be less prone to flooding, unlike urban and cropland areas [11]. The LULC data, produced by [12], are available in GEE with spatial resolution of 100 meters.

The frequency of flood occurrence over time (FO) is a dataset with 30 meters of spatial resolution presented by Pekel et al. (2016). This feature was used to directly identify which areas tend to be more flood prone.

The return period of the discharge associated to the river reach in that minibasin area (QPR), which comes from the hydrological and hydrodynamic model called *Modelo de Grandes Bacias* (MGB). The version of the model used is the MGB-AS, applied to all South America by [13]. The return period was used to eliminate the differences in quantities between discharges of different minibasins.

As observed data for the model, a water mask was used, which is a binary map resulting from a flood inventory, where the values indicate whether the pixel is considered not flooded (pixel=0) or flooded (pixel=1) [9]. The observed water mask was extracted from an image of a reference event, without the presence of clouds by applying the modified version of normalised difference water index (mNDWI) [14].

Two other images of the same scene, but with the presence of clouds in distinct coverage percentage were used to extract a cloud mask (by using the *CloudMasking*, a *QGIS* package). These cloud masks were applied to the reference water mask, excluding pixels that were under the cloud mask, creating a hypothetical water mask that simulates the presence of clouds. The hypothetical water masks remaining pixels were considered the observed data for the models' training.

2.3. Naive Bayes model

Due to its simplicity and elegance, the NB algorithm for categorical data was adopted. NB is the simplest Bayesian network, based on Bayes' theory (equation 1) [15].

$$P(y|Feature) = \frac{P(y) \times P(Feature|y)}{P(Feature)} \quad (1)$$

where y is the class (not flooded/flooded), $Feature$ is used to describe the flood event, $p(y/Feature)$ is the probability of the not flood/flood occur since the $Feature$ occurs, $p(y)$ and $p(Feature)$ are the *a priori* probabilities.

The NB algorithm is known as naive due the conditional independence, where all features are independent given a class variable. When applying the algorithm to a real-world problem, this condition is difficult to be respected (JIANG et al., 2013). However, it has been proven that it is possible to obtain satisfactory results even if there is correlation between the attributes [15].

The NB classifier algorithm is a simplification of equation 1. Considering an i number of features, the result of the classification is $\hat{y}$, presented by the equation 2, adapted from [16]:

$$\hat{y} = \arg\max_{y} p(y) \prod_{i=1}^{n} p(Feature_i|y) \quad (2)$$

where $\hat{y}$ is the class determined by the algorithm, according to the highest value of the equation's product, $p(y)$ is the *a priori* probability of the class (not flooded/flooded), $p(Feature_i/y)$ is the likelihood of $Feature_i$ depending on the y class.

Two models were trained, one for each dataset obtained by applying the cloud masks. The models were conditioned using the input features, with the water masks after the application of the cloud masks considered the observed data. Validation occurred by predicting the flood extensions that were under the cloud mask, comparing them with the observed pixels that had been excluded from the training dataset.

2.4. Performance metrics

Two metrics were used to verify the accuracy of the models generated from the estimation of flood extents: accuracy (AC) and critical success index (CSI). Accuracy is defined by Equation 3, as per [17], and the CSI, which is also known as fit [18] or threat score [19], is defined by Equation 4:

$$AC = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

$$CSI = \frac{TP}{TP + FP + FN} \quad (4)$$

where: TP are the true positives, pixels that are predicted and observed flooded, TN are the true negatives, pixels that are predicted and observed not flooded, FP are the false positives, pixels predicted flooded and observed not flooded, and FN are the false negatives, pixels predicted not flooded and observed flooded.

AC was chosen because it is a metric that generally represents the percentage of success on a predictive model, widely used in flood mapping works. Likewise, the choice for

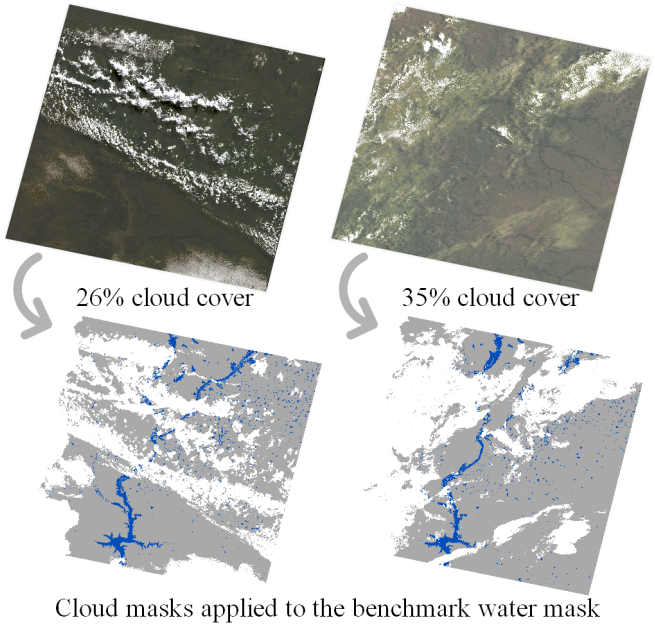


Figure 3: The two images used to create the cloud mask (top) and the respective hypothetical water mask (bottom)

CSI was due its metric that varies between 0 and 1, 0 meaning total disagreement between the observed and predicted flood extent, and 1 meaning total agreement. Therefore, it disregards true negatives, predicted and observed non-flooded, which is the value that causes accuracy imbalance, penalizing the result with false predictions (positive or negative) [18].

3. RESULTS

The flood event selected to be the benchmark is the one that occurred in January 4th, 2015. The image that represents this event is a *Landsat* 8, Collection 2, Level 2 located in path 225, row 081. This event was selected due to the absence of clouds (Figure 1, left). The water mask of this flood event was extracted using mNDWI and is showed in the right side of Figure 1. This is the benchmark water mask.

The images used to create the cloud mask are both from the same *Landsat* scene of the benchmark. Cloud masks extracted are from August 29, 2020 with 26% cloud cover and January 18, 2020 with 35% cloud cover, and they are presented in the upper side of Figure 3. Both hypothetically created water masks by applying these cloud masks to the benchmark are presented in the lower side of Figure 3.

Two NB models were trained with the input data and the remaining pixels of each hypothetical water mask used as observed values. The prediction using the NB model trained with 26% cloud cover achieved an AC of 0.9817 and a CSI of 0.706, considering only the pixels under the cloud mask. For the prediction using the NB with 35% cloud cover model, the performance metrics were 0.9775 for AC and 0.713 for CSI.

The results of the reconstructed pixels are showed in Figure 4, left side for the reconstruction of the 26% cloud cover image and 35% for the right side. More detailed results are shown in lower part of the Figure 4, visually comparing the models performance.

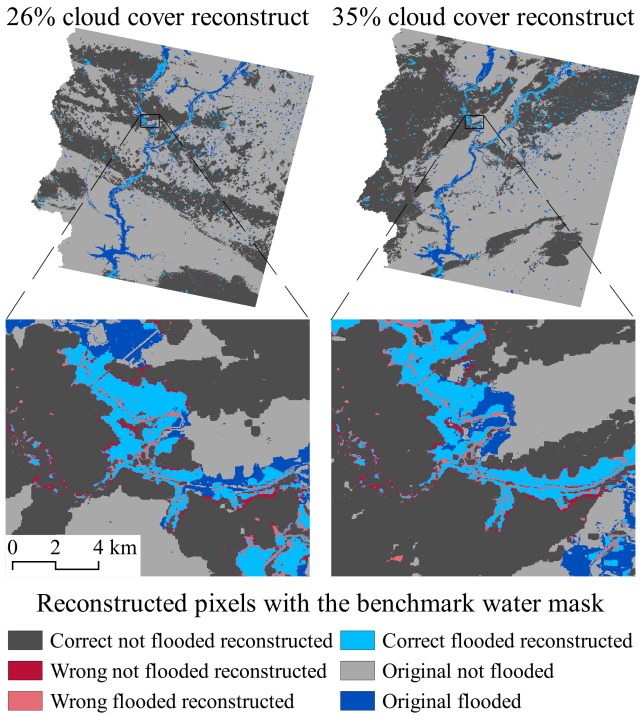


Figure 4: Reconstruction of pixels hypothetically considered as cloud. In the left the 26% and in the right the 35% cloud cover predicted values under the cloud mask.

4. DISCUSSION

A water mask from an image in which clouds would hypothetically be present was reconstructed using a NB model, estimating the lost information due the clouds based on the remaining values, that were assimilated by the model. Although the temporal substitution method is the most usual for reconstructing remote sensed images with the presence of clouds [20], this is not applicable to the case, as a flood event has short duration and temporarily modifies the landscape, especially in the floodplain.

A threshold of 0.65 for the CSI was proposed by [21] to consider flood maps as locally relevant. The results obtained by the methodology presented here achieved higher values of CSI than the proposed threshold. This could indicate the possibility of using the NB model to reconstruct images for use in studies on larger scales such as regional or continental, without prejudice to accuracy and using a flood event that, depending on the size of the cloud cover would not be used.

5. CONCLUSIONS

This work explored the potential of the machine learning Naive Bayes algorithm to reconstruct water masks derived from optical remote sensed images with presence of clouds.

The NB model trained with the remaining data of a 26% of cloud cover was able to reconstruct a water mask, extracted from a Landsat 8 image, with an AC of 0.9817 and a CSI of 0.706. Another NB model, trained with 35% of cloud cover image, extracted from the same Landsat scene, achieved the proposed performance metrics of 0.9817 and 0.706 for AC and CSI, respectively.

These results demonstrated the feasibility of using the proposed model in this context. For future work, a recommendation would be to investigate the use of the model in other regions. In addition, in the same flood event, compare water masks obtained by images captured by optical sensors with masks obtained by microwave sensors such as those present in Sentinel satellites.

6. REFERENCES

[1] Tawatchai Tingsanchali and Thanasit Promping. Comprehensive Assessment of Flood Hazard, Vulnerability, and Flood Risk at the Household Level in a Municipality Area: A Case Study of Nan Province, Thailand. *Water*, 14(2):161, January 2022.

[2] Giuliano Di Baldassarre, Guy Schumann, Paul D. Bates, Jim E. Freer, and Keith J. Beven. Flood-plain mapping: a critical discussion of deterministic and probabilistic approaches. *Hydrological Sciences Journal*, 55(3):364–376, April 2010.

[3] Ila Chawla, L. Karthikeyan, and Ashok K. Mishra. A review of remote sensing applications for water security: Quantity, quality, and extremes. *Journal of Hydrology*, 585:124826, June 2020.

[4] Hamid Mehmood, Crystal Conway, and Duminda Perera. Mapping of Flood Areas Using Landsat with Google Earth Engine Cloud Platform. *Atmosphere*, 12(7):866, July 2021.

[5] Eliakim Hamunyela, Martin Hipondoka, Frans Persendt, Hilma Sevelia Nghiyalwa, Chris Thomas, and Kenneth Matengu. Spatio-temporal characterization of surface water dynamics with Landsat in endorheic Cuvelai-Etosha Basin (1990–2021). *ISPRS Journal of Photogrammetry and Remote Sensing*, 191:68–84, September 2022.

[6] Rui Liu, Yun Chen, Jianping Wu, Lei Gao, Damian Barrett, Tingbao Xu, Linyi Li, Chang Huang, and Jia Yu. Assessing spatial likelihood of flooding hazard using naïve Bayes and GIS: a case study in Bowen Basin, Australia. *Stochastic Environmental Research and Risk Assessment*, 30(6):1575–1590, August 2016.

[7] T. Tadono, H. Ishida, F. Oda, S. Naito, K. Minakawa, and H. Iwamoto. Precise Global DEM Generation by ALOS PRISM. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-4:71–76, April 2014.

[8] A.D. Nobre, L.A. Cuartas, M. Hodnett, C.D. Rennó, G. Rodrigues, A. Silveira, M. Waterloo, and S. Saleska. Height Above the Nearest Drainage – a hydrologically relevant new terrain model. *Journal of Hydrology*, 404(1-2):13–29, June 2011.

[9] Maria Kaiser, Stephan Günnemann, and Markus Disse. Regional-scale prediction of pluvial and flash flood susceptible areas using tree-based classifiers. *Journal of Hydrology*, 612:128088, September 2022.

[10] Dai Yamazaki, Daiki Ikeshima, Jeison Sosa, Paul D. Bates, George H. Allen, and Tamlin M. Pavelsky. MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset. *Water Resources Research*, 55(6):5053–5073, June 2019.

[11] Seyed Mostafa Mousavi, Behzad Ataie-Ashtiani, and Seiyed Mossa Hosseini. Comparison of statistical and MCDM approaches for flood susceptibility mapping in northern Iran. *Journal of Hydrology*, 612:128072, September 2022.

[12] Marcel Buchhorn, Myroslava Lesiv, Nandin-Erdene Tsendbazar, Martin Herold, Luc Bertels, and Bruno Smets. Copernicus Global Land Cover Layers—Collection 2. *Remote Sensing*, 12(6):1044, March 2020.

[13] Vinícius A. Siqueira, Rodrigo C. D. Paiva, Ayan S. Fleischmann, Fernando M. Fan, Anderson L. Ruhoff, Paulo R. M. Pontes, Adrien Paris, Stéphane Calmant, and Walter Collischonn. Toward continental hydrologic–hydrodynamic modeling in South America. *Hydrology and Earth System Sciences*, 22(9):4815–4842, September 2018.

[14] Hanqiu Xu. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14):3025–3033, July 2006.

[15] Harry Zhang. The Optimality of Naive Bayes. *American Association for Artificial Intelligence*, 1(2):6, 2004.

[16] Jun Li, Lixin Ding, and Bo Li. A Novel Naive Bayes Classification Algorithm Based on Particle Swarm Optimization. *The Open Automation and Control Systems Journal*, 6(1):747–753, December 2014.

[17] Sk Ajim Ali, Farhana Parvin, Quoc Bao Pham, Matej Vojtek, Jana Vojteková, Romulus Costache, Nguyen Thi Thuy Linh, Hong Quan Nguyen, Ateeque Ahmad, and Mohammad Ali Ghorbani. GIS-based comparative assessment of flood susceptibility mapping using hybrid multi-criteria decision-making approach, naïve Bayes tree, bivariate statistics and logistic regression: A case of Topl'a basin, Slovakia. *Ecological Indicators*, 117:106620, October 2020.

[18] P.D Bates and A.P.J De Roo. A simple raster-based model for flood inundation simulation. *Journal of Hydrology*, 236(1-2):54–77, September 2000.

[19] Rodrigo C. D. Paiva, Walter Collischonn, and Diogo Costa Buarque. Validation of a full hydrodynamic model for large-scale hydrologic modelling in the Amazon: AMAZON LARGE-SCALE HYDRODYNAMIC MODELLING. *Hydrological Processes*, 27(3):333–346, January 2013.

[20] Huanfeng Shen, Xinghua Li, Qing Cheng, Chao Zeng, Gang Yang, Huifang Li, and Liangpei Zhang. Missing Information Reconstruction of Remote Sensing Data: A Technical Review. *IEEE Geoscience and Remote Sensing Magazine*, 3(3):61–85, September 2015.

[21] Ayan Fleischmann, Rodrigo Paiva, and Walter Collischonn. Can regional to continental river hydrodynamic models be locally relevant? A cross-scale comparison. *Journal of Hydrology X*, 3:100027, April 2019.