RIVER SEDIMENT YIELD CLASSIFICATION USING REMOTE SENSING IMAGERY

R. Pisani

Federal University of Alfenas
Natural Sciences Institute
pisanigeo@gmail.com

K. Costa, G. Rosa, D. Pereira, J. Papa

São Paulo State University
Department of Computing
gth.rosa@uol.com.br
dpereira@ic.unicamp.br
papa@fc.unesp.br

J.M.R.S. Tavares

Universidade do Porto
Faculdade de Engenharia
tavares@fe.up.pt

ABSTRACT

The monitoring of water quality is essencial to the mankind, since we strongly depend on such resource for living and working. The presence of sediments in rivers usually indicates changes in the land use, which can affect the quality of water and the lifetime of hydroelectric power plants. In countries like Brazil, where more than 70% of the energy comes from the water, it is crucial to keep monitoring the sediment yield in rivers and lakes. In this work, we evaluate some state-of-the-art supervised pattern recognition techniques to classify different levels of sediments in Brazilian rivers using satellite images, as well as we make available an annotated dataset composed of two images to foster the related research.

Index Terms— Sediment Yield, Machine Learning, Optimum-Path Forest

I. INTRODUCTION

With every passing day, the water resources become even more scarce in our planet. Being mostly found in lakes and rivers, fresh water is quite complicated to be drained in some remote and/or pollute regions. Even in places with plenty of water like Brazil, there is a need to monitor its quality for further usage by humans and industries.

Hydroelectric power plants are in charge of producing more than 70% of the energy used in Brazil, and they strongly depend on the quality of the water that flows through the rivers. Different levels of sediments in the water, for instance, may cause the dam to get silted. Also, such sediments can influence the water turbidity, which may affect the lifetime of the power plant. Another problem related to different levels of sediments in rivers concerns changes in the land use, where the bare soil gets flushed down to the river, thus pushing forward the sediment yield.

Therefore, to automatic identify the levels of sediments has become crucial to monitor the quality of water, as well as whether there have been changes in the land use behavior around the region of interest or not. Cigizoglu and Alp [1], for instance, used a Generalized Regression Neural Network for river suspended sediment estimation. The results obtained by means of neural networks were considerably superior when compared against multilinear regression and a conventional sediment rating curve technique. Nagy et al. [2] used a neural network trained with backpropagation to estimate the load concentration of sediments in rivers, and Shamaei and Kaedi [3] employed Genetic Programming and Neuro-Fuzzy to estimate sediment concentration in water flow.

Recently, Lafdani et al. [4] used Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) to predict daily suspended sediments in Doiraj River, Iran. Based on an 11-year dataset (1994-2004), the authors used information from rainfall and streamflow to build a regression model to estimate the amount of sediments. Both ANNs and SVMs were able to find very suitable results. Later on, Adib and Mahmoodi [5] used a hybrid approach composed of a neural network and Genetic Algorithm (GA) to predict suspended sediment load at flood conditions. Roughly speaking, the authors employed GA to optimize the architecture of a network trained with the Levenberg-Marquard algorithm. The authors stated GA can reduce the Normalized Mean Square Error of the network up to 80%, which is further used to predict floods together with Markov chains.

A similar work was conducted by Kisi et al. [6], which used Genetic Programming (GP) to estimate suspended sediments. The daily water discharge and river sediment load data of two stations on Cumberland River (USA) were used to build the proposed model, which was compared against an Adaptive Neuro-Fuzzy Inference System, ANNs and SVMs. The results evidenced that GP can be more effective to estimate daily suspended sediment load. Gupta et al. [7], in 2002, used satellite imagery to evaluate the geomorphology and to map environmental degradation and sediment transfer in some parts of the Mekong River. Also, the authors aimed at studying the possible impacts of some development projects on the river. However, as far as we are concerned, machine learning techniques have not been used in this work. Similarly, related works based on radar and satellite images can be referred as well [8], [9], but not considering machine learning-based interpretation either.

Therefore, as we have observed, the works usually focus on regression models mostly, and with data provided by rainfall and previous sediment load information. In order to fill this gap, this work has two main contributions: (i) firstly, we compared some state-of-the-art supervised pattern recognition techniques to estimate different levels of sediments in rivers using satellite images, and (ii) secondly, we made available an annotated dataset composed of two satellite images covering Brazilian rivers in order to foster the research on satellite-based sediment yield estimation in rivers.

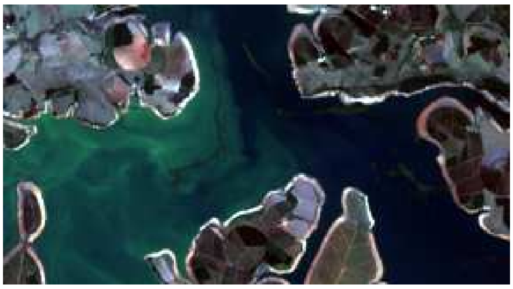
We have the feeling the related area of research lacks on ground-truth data, since few works can be considered similar to ours. As a matter of fact, we have not observed any work that attempted to use Support Vector Machines [10], Naïve-Bayes (NB) [11] and Optimum-Path Forest (OPF) [12], [13] to classify sediment yield in rivers by means of satellite images. The reason for using such techniques concerns the fact that SVMs are considered one of the best techniques to date, and NB and OPF offer competitive results at the price of being much faster for training, since they are parameterless, thus not needing fine-tuning.

The remainder of this paper is organized as follows. Section II presents the datasets used in this work, and Section III discusses the methodology and experiments adopted to validate the datasets and the supervised pattern recognition techniques used for sediment yield classification purposes. Finally, Section IV states conclusions and future works.

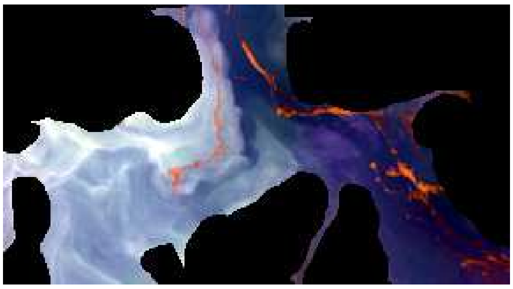
II. DATASET

In this paper, we make available a dataset composed of two images from distinct Brazilian covering areas: (i) Furnas and (ii) Tietê[1]. As aforementioned, the main idea is to foster the research on sediment yield identification by means of remote sensing images. We used images obtained from Landsat-8 satellite, OLI sensor, covering the area of Tietê river, city of Botucatu, São Paulo state, Brazil; as well as another image covering the area of Furnas, city of Alfenas, Minas Gerais state, Brazil. Notice both covering areas represent important water catchment sources for electricity generation in Brazil. The images were collected from INPE (National Institute of Spatial Research) site catalog, and processed and labeled in ENVI 5. Further, the images were geo-referred using the ArcGIS 10.4 tool.

Figures 1 and 2 depict the images covering the areas of Furnas and Tietê, respectively. Additionally, the aforementioned figures display the area of interest of this work, i.e. the rivers. We used bands 2 (blue), 3 (green) and 4 (red) for the image composition process, as follows: 4R3G2B concerning Furnas, and 2R3G4B with respect to Tietê image. Such bands describe better the sediment yields in both rivers, thus making the process of image labeling (i.e. ground truth) easier.



(a)



(b)

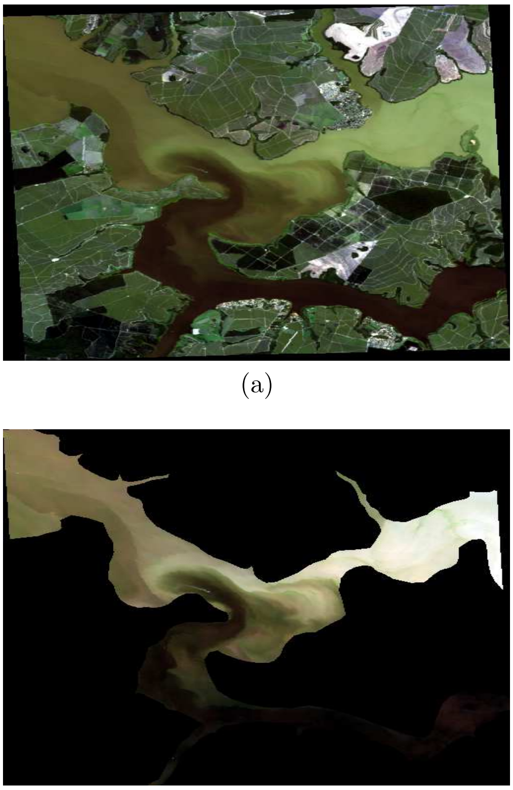
Fig. 1. Image covering the area of Furnas: (a) original image, and (b) the mask containing the area of interest (river) only.

Table I presents a brief description of the images, which were labeled according to the following classes and colors:

- Furnas:
  - Class 1: background - white
  - Class 2: sediment yield level 1 - blue (high level of turbidity);
  - Class 3: sediment yield level 2 - light green

---

[1] The datasets are available at http://wwwp.fc.unesp.br/~papa/recogna/remote_sensing.html.


(a)

(b)

Fig. 2. Image covering the area of Tietê: (a) original image, and (b) the mask containing the area of interest (river) only.

  (average level of turbidity);
  - Class 4: sediment yield level 3 - green (low level of turbidity);
  - Class 5: water plants - rose.
- Tietê:
  - Class 1: background - white
  - Class 2: sediment yield level 1 - blue (high level of turbidity I);
  - Class 3: sediment yield level 2 - red (high level of turbidity II);
  - Class 4: sediment yield level 3 - dark blue (average level of turbidity);
  - Class 5: sediment yield level 4 - green (low level of turbidity);

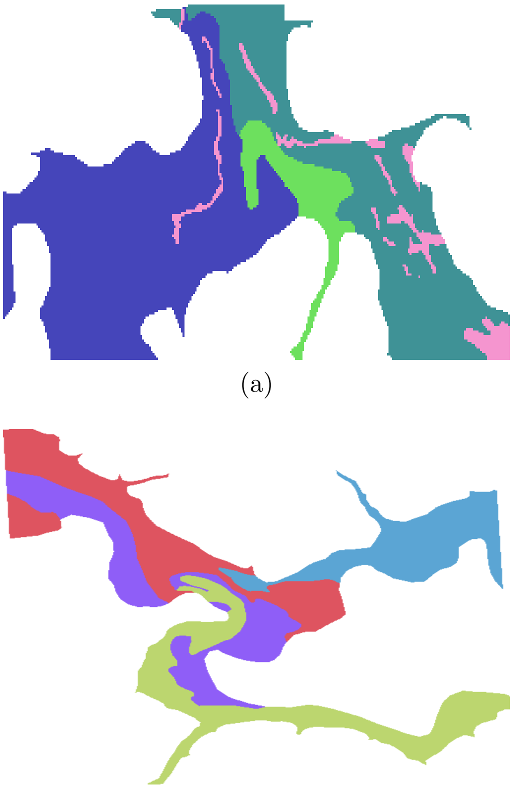
Figure 3 displays the ground-truth images labeled according to the aforementioned colors. Both images were analyzed and labeled by an expert in geography from our research group.

III. EXPERIMENTAL SECTION

In this section, we present the methodology employed to evaluate the effectiveness and efficiency of the classifiers adopted for sediment classification purposes, as well as the experimental results. In regard to the pattern recognition techniques, we considered the Optimum-Path Forest, a Bayesian classifier, and Support Vector Machines. Also, we evaluated the influence of different training set sizes with respect to the accuracy: we considered training sets with 5% and 10% of the entire image, being the remaining pixels (samples) used to compose the test set. Notice each pixel was described by its RGB values to compose the feature vector.

Table I. Description of the images that compose the dataset.

Image	Size	Classes
Furnas	288 × 156	4
Tietê	515 × 549	4


(a)

(b)

Fig. 3. Ground-truth images: (a) concerning Figure 1b, and (b) with respect to Figure 2b.

In order to allow a robust statistical analysis, we employed a hold-out validation approach over 15 randomly generated training and testing sets. Further, the Wilcoxon signed-rank test [14] with significance of 0.05 was used for validation purposes. In regard to SVM source code, we used the open-source library LibSVM with a Radial Basis Function kernel[2]. The searching range of parameter C was defined within $[-32, 32]$, and the searching interval of parameter γ was restricted within $[0, 32]$. Notice the step-size for both parameters is equal to 2. With respect to OPF implementation, we employed the LibOPF [15], and concerning the Bayesian classifier, we employed our own implementation.

Table II presents the mean accuracy and class-specific accuracy results considering a training set composed of 5% and 10% of the entire image, being the most accurate results according to Wilcoxon statistical test in bold. The recognition rates were computed using an accuracy measure proposed by Papa et al. [12], which considers unbalanced data. In case of Furnas dataset, for instance, one can clearly observe this problem, where the "pink" class has way less samples than the "blue" class (Figure 3a).

The best results were obtained by SVM for both training set configurations, followed by Bayes and OPF. We did no observe a clear difference among the two training set configurations, i.e. it is usually expected better recognition rates when using larger training sets. The only situation observed concerns OPF over the Tietê image using 10% of the image for training purposes, which obtained an accuracy of 66.87% with a considerably high standard deviation. Such behaviour can be explained by taking a look at the central region of Figure 2b reveals similar colours, tough representing different sediment classes. One of the strongest skills of the OPF classifier turns out to be its main weakness: a theoretical property says OPF minimizes the classification error over the training set, which can be close to zero depending on the configuration (distribution of samples) of the training set [16]. Roughly speaking, OPF training step aims at partition the graph induced by the dataset samples by means of a competition process among *prototype* samples (key samples chosen from each class). Therefore, OPF is quite susceptible to the quality of such prototypes, which means it can obtain suitable results when the prototypes are chosen at the regions with highest probability of misclassification, i.e. the central region of Figure 2b. As we are creating training and test sets at random, we can no longer guarantee the prototypes will be placed at those regions every time.

Figures 4 and 5 depict some images classified by SVM, OPF and Bayes using 5% of the entire image for training purposes over Furnas and Tietê datasets, respectively. We can observe a better performance concerning SVM with respect to Tietê image, i.e. with low spreading (confusion) among the sediment classes. If we consider classes 1 and 2 only, Bayes obtained better recognition results than OPF with respect to the central region of Figure 5a. A similar performance applies to Furnas image either, with better recognition rates obtained by SVM. In this case, OPF results were spread from classes 3 and 4 to class 1. Also, classes 1 and 3 were better recognized by all classifiers.

Tables III and IV present the mean computational load in seconds concerning all techniques employed in this work over Furnas and Tietê images, respectively. Clearly, the Bayesian classier and OPF were considerably faster than SVM, since they are parameterless. Notice SVM training time also considers the fine-tuning parameters procedure. In regard to the test step, SVM has been the fastest classifier, closely followed by OPF and then Bayes. If one considers the whole computational load, i.e. training+testing, OPF classifier has been the fastest one, followed by Bayes and SVM. The main problem related to OPF concerns it needs to go over the whole training set (in the worst case) to verify the sample that will conquer each test sample, which does not happen with SVM.

Table III. Mean training time (seconds).

5%

Image	Bayes	OPF	SVM
Furnas	0.06 ± 0.00	0.13 ± 0.00	164.34 ± 3.64
Tietê	1.92 ± 0.00	5.46 ± 0.11	2638.73 ± 29.80

10%

Image	Bayes	OPF	SVM
Furnas	0.25 ± 0.00	0.61 ± 0.01	673.27 ± 20.17
Tietê	7.89 ± 0.02	22.23 ± 0.77	10206.96 ± 253.72

Table IV. Mean testing time (seconds).

5%

Image	Bayes	OPF	SVM
Furnas	9.16 ± 0.04	1.26 ± 0.05	1.18 ± 0.18
Tietê	334.46 ± 0.73	38.36 ± 0.30	22.84 ± 3.43

10%

Image	Bayes	OPF	SVM
Furnas	18.28 ± 0.21	3.03 ± 0.15	2.02 ± 0.31
Tietê	638.01 ± 3.78	73.39 ± 0.24	38.39 ± 3.30

IV. CONCLUSIONS

In this paper, we dealt with the problem of sediment yield classification in remote sensing images by means of supervised pattern recognition techniques. The main idea is to employ images acquired by satellites to automatic classify the level of sediments in Brazilian rivers, since such information is of extreme importance to measure the turbidity of the water, which has a strong influence in the lifetime of power plants. Also, the amount of sediments are usually related to the land-use behaviour nearby the river. Another main contribution of this work is to make available a dataset composed of

---

[2] https://www.csie.ntu.edu.tw/˜cjlin/libsvm

Table II. Mean recognition rates and class-specific accuracy. The recognition rates in parenthesis are displayed as follows: (a_1,a_2,a_3,a_4,a_5), where a_i stands for the men recognition rate of class i according to the definition presented in Section II.

5%

Image	Bayes	OPF	SVM
Furnas	95.88 ± 0.06(100/95.9/94.9/98.4/90.0)	95.63 ± 0.13(100/95.3/95.3/98.3/89.0)	**96.33 ± 0.29(100/93.6/98.2/99.7/89.4)**
Tietê	97.56 ± 0.04(100/99.1/89.9/99.2/99.7)	82.71 ± 29.24(100/98.5/86.6/72.6/72.7)	**98.07 ± 0.03(100/99.7/90.7/99.8/99.9)**

10%

Image	Bayes	OPF	SVM
Furnas	95.97 ± 0.17(100/96.2/95.8/97.7/89.9)	95.77 ± 0.20(100/95.0/97.0/98.9/88.0)	**96.68 ± 0.14(100/97.1/96.0/98.9/91.0)**
Tietê	97.71 ± 0.03(100/98.3/92.5/99.0/99.0)	66.87 ± 36.92(100/58.0/55.9/53.8/66.2)	**98.14 ± 0.02(100/99.2/93.8/99.0/99.0)**

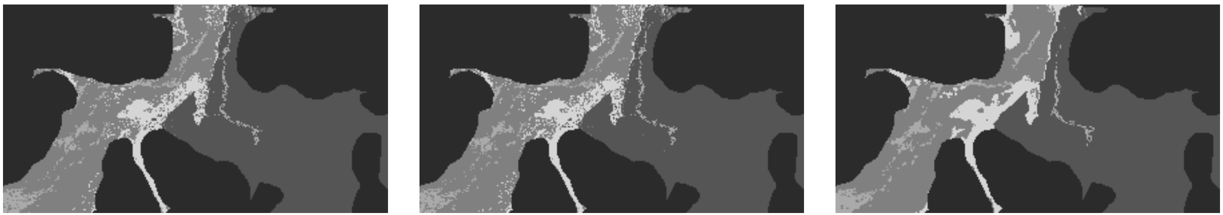

(a) (b) (c)

Fig. 4. Furnas: classified images using 5% of the entire image for training by means of (a) Bayes, (b) OPF and (c) SVM.

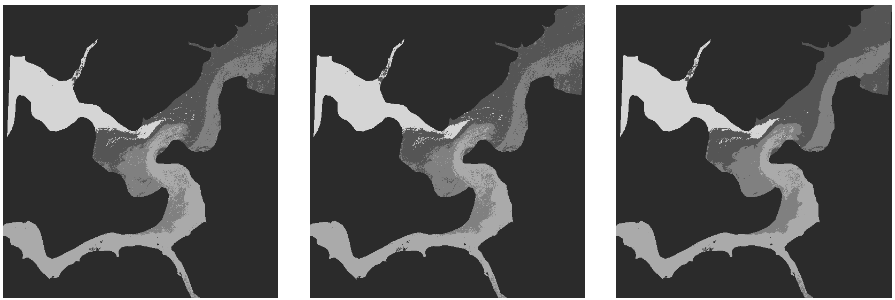

(a) (b) (c)

Fig. 5. Tietê: classified images using 5% of the entire image for training by means of (a) Bayes, (b) OPF and (c) SVM.

two annotated images to foster the research related to sediment yield classification in rivers.

We have compared the performance of three state-of-the-art classifiers using two different training sets: one composed of 5% of the entire image for training purposes, and another with a larger training set (10% of the image). SVM classifier obtained the best results for both images and percentages of training sets, but at the price of being the costly technique for training purposes. The best trade-off concerning training+testing computational load was obtained by the OPF classifier, but at the price of being placed in third concerning the recognition rates, right after SVM and Bayes.

The task of sediment classification by means of satellite images seems to be fruitful, since we obtained results nearly to 99%, but at the price of a high computational load when considering SVM classifiers. In regard to future works, we plan to make available more images to the scientific community, as well as to study the influence of the sediments with the land-use, since we do not have labeled images at the very same region concerning both information to date.

ACKNOWLEDGMENTS

The authors are grateful to FAPESP grants #2014/16250-9, #2015/25739-4, #2015/00801-9 and #2015/00801-9, as well as CNPq grant #306166/2014-3. Authors gratefully acknowledge the funding of Project NORTE-01-0145-FEDER-000022 - SciTech - Science and Technology for Competitive and Sustainable Industries, co-financed by “Programa Operacional Regional do Norte” (NORTE2020), through “Fundo Europeu de Desenvolvimento Regional” (FEDER).

V. REFERENCES

[1] H. K. Cigizoglu and M. Alp, "Generalized regression neural network in modelling river sediment yield," *Advances in Engineering Software*, vol. 37, no. 2, pp. 63–68, 2006.

[2] H. Nagy, K. Watanabe, and M. Hirano, "Prediction of sediment load concentration in rivers using artificial neural network model," *Journal of Hydraulic Engineering*, vol. 128, pp. 588–595, 2002.

[3] E. Shamaei and M. Kaedi, "Suspended sediment concentration estimation by stacking the genetic programming and neuro-fuzzy predictions," *Applied Soft Computing*, vol. 45, pp. 187–196, 2016.

[4] E. K. Lafdani, A. M. Nia, and A. Ahmadi, "Daily suspended sediment load prediction using artificial neural networks and support vector machines," *Journal of Hydrology*, vol. 478, pp. 50–62, 2013.

[5] A.Adib and M. Mahmoodi, "Prediction of suspended sediment load using ANN GA conjunction model with markov chain approach at flood conditions," *KSCE Journal of Civil Engineering*, pp. 1–11, 2016.

[6] O. Kisi, A. H. Dailr, M. C., and J. Shiri, "Suspended sediment modeling using genetic programming and soft computing techniques," *Journal of Hydrology*, vol. 450–451, pp. 48–58, 2012.

[7] A. Gupta, L. Hock, H. X., and C. Ping, "Evaluation of part of the mekong river using satellite imagery," *Geomorphology*, vol. 44, no. 3–4, pp. 221–239, 2002, geomorphology on Large Rivers.

[8] L. C. Smith and D. E. Alsdorf, "Control on sediment and organic carbon delivery to the arctic ocean revealed with space-borne synthetic aperture radar: Ob' river, siberia," *Geology*, vol. 26, no. 5, pp. 395–398, 1998.

[9] O. C. Montanher, E. M. L. M. Novo, C. C. F. Barbosa, C. D. Rennó, and T. S. F. Silva, "Empirical models for estimating the suspended sediment concentration in amazonian white water rivers using landsat 5/tm," *International Journal of Applied Earth Observation and Geoinformation*, vol. 29, pp. 67–77, 2014.

[10] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[12] J. P. Papa, A. X. Falcão, and C. T. N. Suzuki, "Supervised pattern classification based on optimum-path forest," *International Journal of Imaging Systems and Technology*, vol. 19, no. 2, pp. 120–131, 2009.

[13] J. P. Papa, A. X. Falcão, V. H. C. Albuquerque, and J. M. R. S. Tavares, "Efficient supervised optimum-path forest classification for large datasets," *Pattern Recognition*, vol. 45, no. 1, pp. 512–520, 2012.

[14] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, Dec. 1945. [Online]. Available: http://dx.doi.org/10.2307/3001968

[15] J. P. Papa, C. T. N. Suzuki, and A. X. Falcão, *LibOPF: A library for the design of optimum-path forest classifiers*, 2014, software version 2.1 available at http://www.ic.unicamp.br/~afalcao/LibOPF.

[16] C. Allène, J.-Y. Audibert, M. Couprie, and R. Keriven, "Some links between extremum spanning forests, watersheds and min-cuts," *Image and Vision Computing*, vol. 28, no. 10, pp. 1460–1471, 2010, image Analysis and Mathematical Morphology.