Geotech Geol Eng (2012) 30:85–94
DOI 10.1007/s10706-011-9451-8

ORIGINAL PAPER

A Data Based Model to Predict Landslide Induced by Rainfall in Rio de Janeiro City

Fábio T. de Souza · Nelson. F. F. Ebecken

Received: 8 November 2009 / Accepted: 12 September 2011 / Published online: 21 September 2011

Abstract Landslide prediction is complex and involves many factors, such as geotechnical, geological, topographical, and even meteorological. This work presents a methodology by using a *Data Mining* approach in order to predict landslide occurrences induced by rainfall in Rio de Janeiro city. Landslide and rain data records from 1998 to 2001 were obtained from field technical reports and 30 automatic rain gauges, respectively. It was also collected data regarding soil parameters, including urban areas, forest, vulnerability, among others, and totalizing 46 soil variables. All the information was inserted into a Geographic Information Systems. *Clustering* (*Dendrogram* and *k-means*) and *Statistical* (*Principal Component Analysis* and *Correlation*) techniques were used to regionalize the rain data and select the rain gauges to be input on *Artificial Neural Networks* , which were used to replace the missing rain values. The landslide volume variable also presented missing values and it was completed by the *k-Nearest Neighbor* method. After data preparation, some models were built to predict landslide and rainfall using *Data Mining* techniques. The obtained model's performance is also analyzed.

Keywords Landslide · Data mining · Geographic information systems · Classification rules · Artificial neural network

F. T. de Souza
State Key Lab of Hydroscience and Engineering, Tsinghua University, Beijing 100084, China

F. T. de Souza (✉)
Course of Civil Engineering, PPGTU—Postgraduate Program in Urban Management, PUCPR—Pontifical Catholic University, Curitiba, Brazil
e-mail: fabiocoppe@yahoo.com.br; fabioteodoro@pucpr.br

F. T. de Souza
Department of Hydraulics and Sanitation (DHS), Federal University of Parana (UFPR), Curitiba, Brazil

Nelson. F. F. Ebecken
Civil Engineering Department, COPPE/UFRJ—Federal University of Rio de Janeiro, P.O. Ilha do Fundão, Zip Code 21941-972 Rio de Janeiro, Brazil
e-mail: nelson@ntt.ufrj.br

1 Introduction

The uncontrolled urbanization process in big cities has imposed changes on the environment. Many areas near the slopes are occupied by a vast proportion of the population and these occupations are exacerbating deforestation, destroying soil vegetation cover, accumulating garbage in inappropriate areas etc. The new atypical conditions impose a new setup of the soil hillsides and make it susceptible to landslide occurrences during rainfall.

Menezes et al. (2000) have shown that the city of Rio de Janeiro has physical characteristics (geographical position and topography) and atmospherical patterns, which are favorable to develop meteorological

phenomena triggering severe rainfall. Summer is usually a rainy season and more than 60% of the landslides occur during this period.

The present work shows a detailed landslide study performed by '*Data Mining*' and '*Geographical Information Systems*' techniques to explicitly outline the important patterns of slope stabilization. The success of this approach depends on careful data preparation and the needed tasks to prepare the data consume more than 90% of the required time. The rest of the paper is organized as follows: Sect. 2 describes the dataset related to the landslide registers; Sect. 3 presents the tasks required in the dataset preparation; Sect. 4 shows in details the developed models to predict the landslides and rainfall using the *Artificial Neural Networks* (*ANN*), *Interesting Association Rules* and *Classification Rules*; Sect. 5 shows the results of the landslide dataset preparation, data modeling and a related discussion; Sect. 6 illustrates the conclusions and future works.

2 Landslide-Related Dataset

The reports of landslide occurrences describe their characteristics, such as location (borough), date, time, typology, damage caused and slipped volume (estimated in cubic meter). Table 1 illustrates a landslide register obtained by the report of GEO-RIO, Rio de Janeiro Geotechnical Engineering Office, the agency in charge of slope assessment and hazard.

The rain dataset is composed by measurements of the rainfall volume every 15 min from 30 automatic rain gauges installed in Rio de Janeiro city. This work has studied 28 rainfall periods associated to the same intervals of the landslide occurrences between 1998 and 2001.

Figure 1 illustrates Rio de Janeiro 159 boroughs' map (light gray lines) and the rain gauges (dark triangle) network centered in Thiessen polygons.

The soil parameters related to each city's borough have been monitored by satellite images since 1984. Municipal Secretary of the Environment of Rio de Janeiro (SMAC) classifies these images and computes the 46 soil parameters in each of the 159 boroughs. Figure 2 illustrates a satellite image collected and analyzed in 1984.

The knowledge of the existent patterns among several phenomena related to the landslide occurrences, allows the establishment of criteria to alert emission and subsequent mobilization of the responsible institutions.

3 Data Preparation

Data preparation is the most important task in any Data Mining strategy. A dataset carefully prepared can better expose the information contained to the modeling tools (Pyle 1999). The cumulative rain indexes associated to each landslide were computed according to the geographical proximity of their occurrence and considering different levels of rainfall in the preceding 6 days. For example, a rain index of 6 h (h_6) is related to the cumulative rain volume from 6 h before the landslide. Seventeen cumulative rain indexes were computed for 15, 30, 45 and 90 min prior; 1, 2, 3, 4, 6, 8 and 12 h prior; 1, 2, 3, 4, 5 and 6 days prior to each event. However, the rain dataset presented missing values and their fulfillment improves the possibility to estimate the rain patterns (cumulative rain indexes) related to the landslides. This task may be performed using *ANN* architecture as follows.

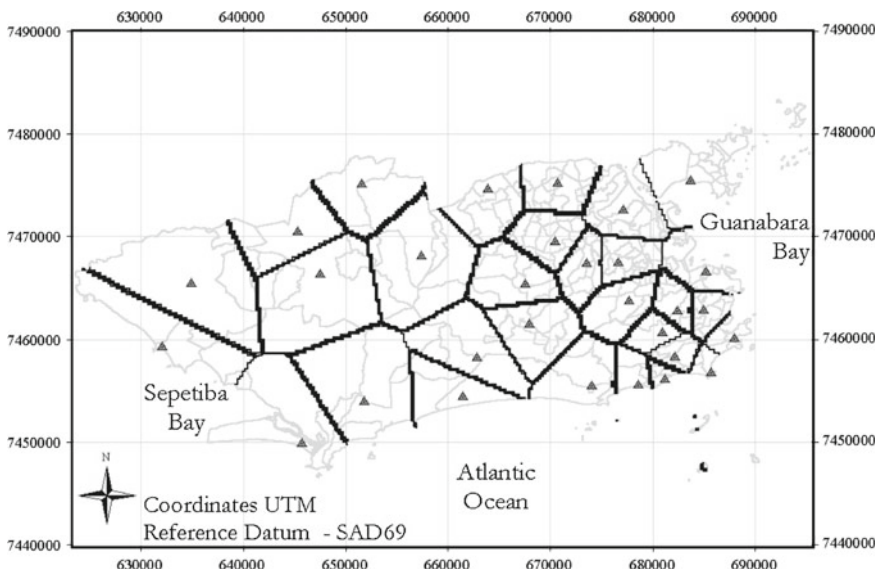
The *ANN* architecture—*Multi Layers Perceptrons* (*MLP's*)—was chosen due to its large learning capacity, generalization, and mainly because of the automatic form of knowledge extraction (Haikin 2001; Kennedy et al. 1997; Bishop 1995; Rumelhart and McClelland 1986). It is less dependent on subjectivity when compared to methodologies commonly used in hydrological literature. In order to replace the missing rain values, the *ANN* was built containing the rain gauge data with missing values in the output layer (dependent variable) and the rain gauges without missing values in the input layer (independent variables).

Table 1 Example of landslide register

Location (address/borough)	Date (year/day/month)	Time	Occurrence description	Class	Volume (m^3)	Consequence
Antônio Rego, 1447/Ramos	(1998/08/Jan)	20:00	Soil slipped in the nature hillside	Es/tc	15	A damaged house

Fig. 1 Automatic rain gauges network

1 - Vidigal
2 – Urca
3 – São Conrado
4 – Tijuca
5 – Santa Tereza
6 – Copacabana
7 – Grajaú
8 – Ilha do Governador
9 – Penha
10 – Madureira
11 – Irajá
12 – Bangu
13 – Piedade
14 – Tanque
15 – Saúde
16 – Jardim Botânico
17 – Itanhangá
18 – Cidade de Deus
19 – RioCentro
20 – Guaratiba
21 – Gericinó
22 – Santa Cruz
23 – Cachambi
24 – Anchieta
25 – Grota Funda
26 – Campo Grande
27 – Sepetiba
28 – Sumaré
29 – Mendanha
30 – Itaúna



The training should not be performed utilizing all the 30 rain gauges data as input because it would introduce bias or noise by using data of distant rain gauges from those with missing values. Thereby, it was previously necessary to perform a rain regionalization. Figure 3 shows the adopted methodology to prepare the landslide dataset (Souza 2004).

4 Rainfall Spatial Analysis (Regionalization)

The selection of the attributes (rain gauges data) for the training dataset was carried out (Souza and Ebecken 2004a, b, c) using four different methods: two statistical, *Principal Component Analyses—PCA* (Wherry 1984) and *Correlation* (Pearson 1896); and two clustering, *K-means* and *Dendrogram* approaches (Lin et al. 2009; Gorsevski et al. 2003; Melchiorrea et al. 2008).

Clustering is the process of grouping data into classes or clusters in which objects within a cluster have high similarity in comparison to another (Han and Kamber 2001). *Tree clustering (Dendrogram)* is a hierarchical method that works by grouping data into a tree of clusters using the distances or dissimilarities between the objects. When there is a clear "structure" in terms of clusters of objects, according to a degree of similarity among the objects, then this structure would often be reflected in the hierarchical tree as distinct "branches". On the other hand, *K-means* partitioning

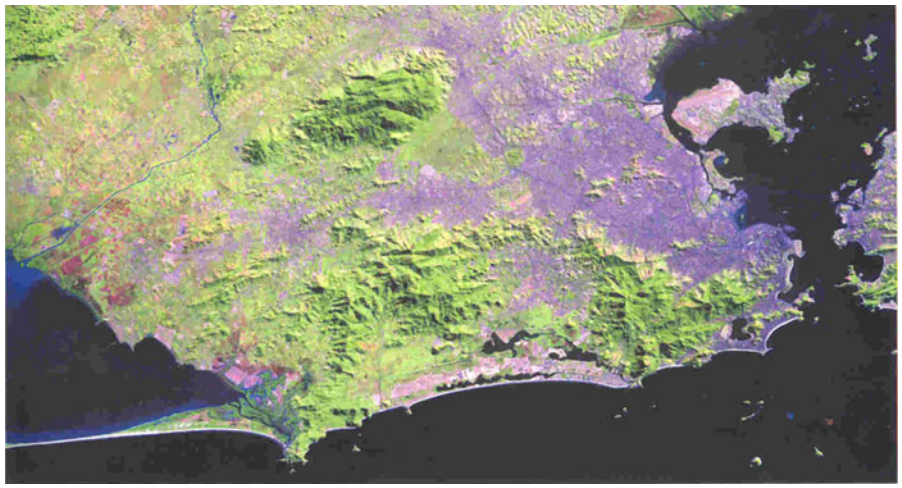


Fig. 2 Satellite image collected in 1984

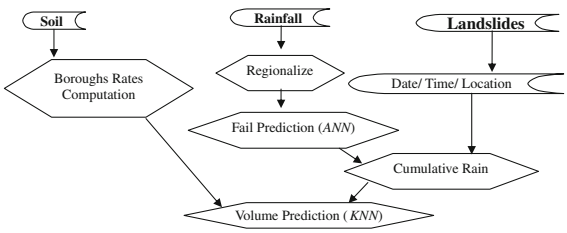


Fig. 3 Landslides data preparation tasks

algorithm splits a set of n objects into k classes, where each partition represents a cluster $k \leq n$, i.e., it classifies the data into k groups and uses an iterative relocation technique that attempts to improve the partitioning by moving the objects through different clusters.

After the rain gauges have been grouped by these regionalization techniques, the rain dataset was prepared for the training, test and prediction of the missing values.

5 Rain Missing Values Replacement

Substitution of missing values is an important subtask for data preparation steps (Hruschka et al. 2003). The rain missing values replacement improves the possibility to estimate the rain patterns (cumulative rain indexes) related to the landslides. The simulations with *ANN* were performed in 28 rainfall events occurred between 1998 and 2001, because all these

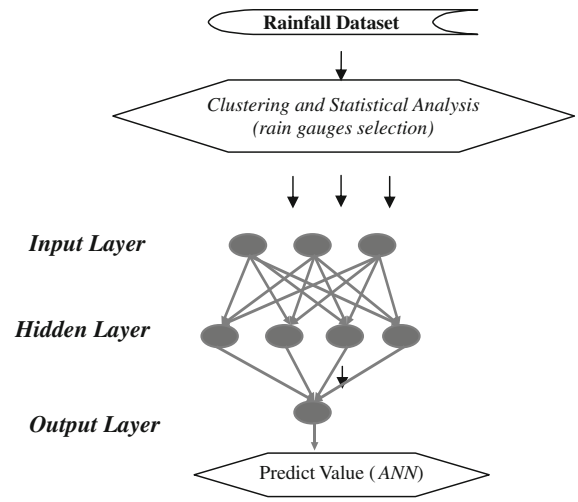


Fig. 4 *ANN* architecture setup

events contained rain missing values. Figure 4 shows the strategy of the rain missing values replacement (Souza and Ebecken 2003).

The number of samples used for training, testing and validation was variable, because it depends on the size of the event in question; however, the percentage for each one of those sample were 80, 10 and 10%, respectively. After the rain missing values have been replaced, it is possible to compute the cumulative rain indexes associated to each landslide. These indexes were computed according to location (rain data from the nearest rain gauge), date and hour of the reported accident.

The modeling tools used in this research require a cognitive map comprising all possible parameters related to studied phenomenon. In this way, the cumulative rain indexes were joined into a matrix including all associated soil variables, such as urban areas, forest, vulnerability, among others.

This matrix is composed by more than 60 attributes (landslides, rain and soil variables) and 1,266 samples. Among these, 1,033 samples are from *landslides* or *panic situation* that have occurred between 1998 and 2001. In order to train also the *non landslides* pattern the remained 233 samples were computed by considering a time-lag previous to the landslide event. For example, if a landslide occurred at 8 am its cumulative rain index of 6 h is the rain volume between 2 am and 8 am. The computation of the cumulative rain index of 6 h for the *non landslide* pattern considered, for example, a time-lag between 1 am and 7 am, or the cumulative rain volume before the landslide. During a severe rainfall, the predictive model should be able to make a decision and answer some of the three possibilities: *landslides, panic situation* or *non landslides*.

As previously mentioned, there are several samples with landslide volume missing values. In order to replace them, the *K-nearest-neighbor* (*Knn*) method was adopted (Mitchell 1997).

6 Slipped Volume Missing Values Replacement

Knn method focuses on missing values substitution by a combination of the corresponding attributed values of the most similar complete objects found in the dataset. This method was applied in this study using two different distance definitions, Euclidean and Manhattan.

Considering two objects i and j, both described by a set of N continuous attributes $\{x_1, x_2, \ldots, x_N\}$, the distance between object i and object j is called $d(i, j)$. Supposing that the kth attribute value ($1 \leq k \leq N$) of the object m is missing; thus, the Nearest Neighbor Method (*NNM*) should compute the distances $d(m, i)$, for all $i \neq m$, according to the Euclidean or Manhattan distance. Finally the algorithm combines the k objects with the lowest distance in order to replace the missing value.

To validate this method several simulations with *knn* algorithm in a sample of 98 registers of the matrix (10% of the 977 total full registers) were performed. These simulations were carried out varying the number of attributes (according to the clustering methods), varying the number of k or neighbors and reporting the average relative error and the correlation. Among these several simulations, the best result, which presented the minimum average error and the maximum correlation, was extracted to complete the volume-missing values. The results are shown on Sect. 5. The proposed *knn* method could be easily adapted to datasets formed by discrete attributes, changing the Euclidean/Manhattan distance by *Simple Matching Approach* (Kaufman and Rousseeuw 1990).

7 Data Modeling

After data preparation, some models were built, involving three approaches: (1) landslides prediction; (2) association rules extraction; and (3) rainfall prediction. The first approach corresponds to construct models using two techniques: *ANN* and *Classification Rules*—an algorithm that finds a set of interesting rules based on two indexes: *support* and *confidence* (Liu et al. 1998). The second approach extracts rules from the landslides database and the proposed algorithm is based on algorithm Apriori for finding association rules (Agrawal and Srikant 1994; Liu et al. 2000). The third approach generates a model to predict the rainfall volume for the next hours/days and considering the previous ones. The last two approaches furnish one more landslide prediction model, i.e., if a rule describes a landslide occurrence by a particular rain index. If it is possible to predict this rain index, it would be possible to associate the landslide risk according to the confidence rule. The next rule illustrates an example:

IF *h_6_>_43.7mm*
THEN *->>* *LANDSLIDE* (1)
(90.6% *117* *106)*

[Within 6 h of cumulative rain index (h_6) measuring above 43.7 mm (presented by the total amount of 117 registers), 106 out of 117 registers (90.6%) would predict a landslide].

8 Results and Discussion

As described earlier, the rain regionalization was carried out considering the whole network of rain

gauges. In order to select the nearest rain gauges (input layer) around the rain gauge containing missing value (output layer) data mining techniques were applied.

Figures 5 and 6 show examples of regionalization methods applied to the rain data set of the Rio de Janeiro city. These maps have a white polygon, in the eastside of the city, which corresponds to the surrounding area (Thiessen polygon) of the rain gauge with missing values during the rainfall event occurred from 1st to 13th on January 1998 (first rainfall event). This example was chosen due to its high precipitation values (peak of 26.8 mm in 15 min).

Landslides occurred in all the boroughs with a white outline during this event (67 from a total of 159 boroughs). The dark grey polygons correspond to the surrounding rain gauges that were selected together with the rain gauge with missing values. In Fig. 5a these dark grey polygons (rain gauges) were grouped by the *PCA* technique. The *PCA* can be seen as a data reduction method by the association of two or more correlated variables into factors. This method consists in an axes rotation strategy, and creates a pattern of interpretation easier and clear through the factors visualization with high loads to some variables and low loads to others. In Fig. 5b the dark grey polygons (rain gauges) were grouped by the *Correlation* technique which considers the coefficients of correlation ≥ 0.70, if compared to the rain gauge with missing values.

The dark grey polygons grouped by the *Tree-Clustering* technique were selected to illustrate the rain gauge on the "branch" that contains the rain gauge with missing values, as illustrated in Fig. 6a. Figure 6b shows the dark grey polygons grouped by the *K-means* technique, which considers the cluster that contains the rain gauge with missing values.

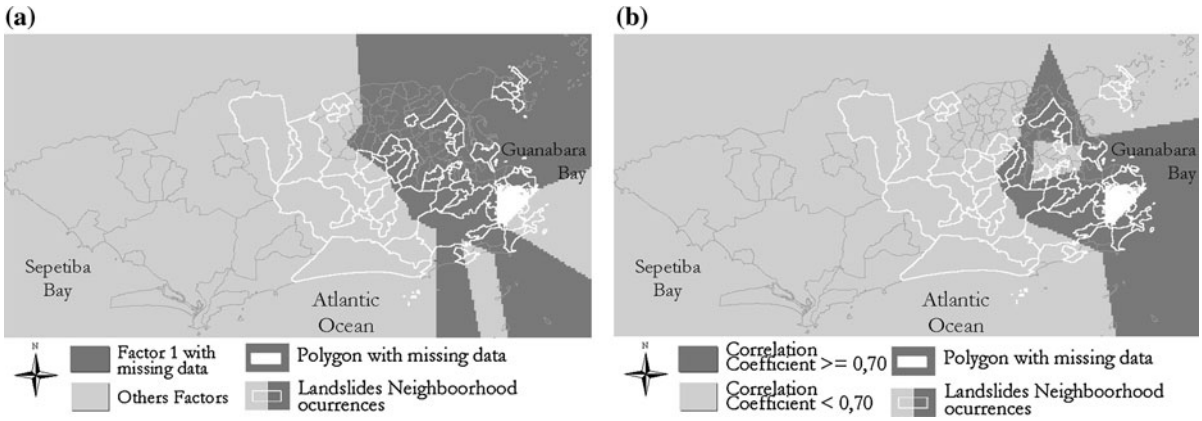


Fig. 5 Regionalize with PCA and correlation methods

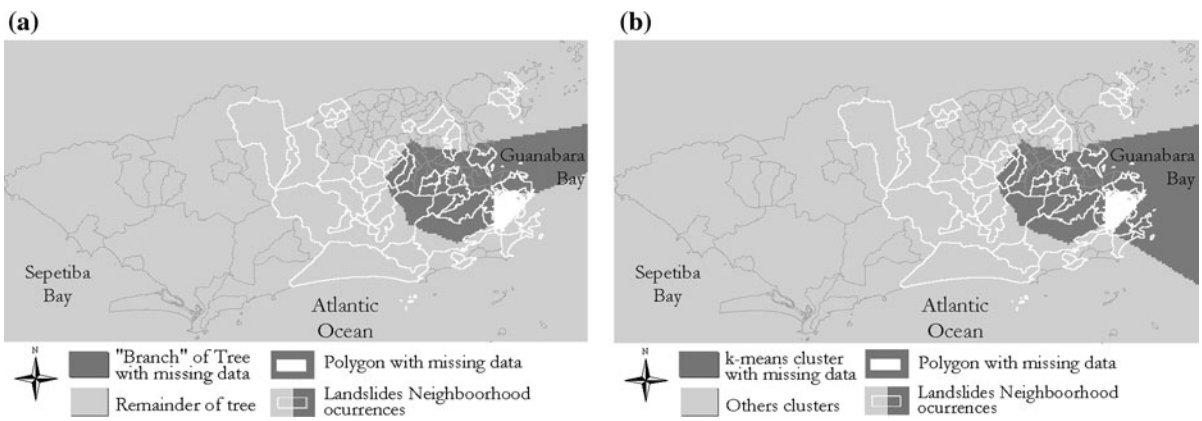


Fig. 6 Regionalize with Tree-Clustering and K-means methods

Table 2 Results from *ANN* predictions (validation)

Methods	SDR	PRC
PCA	0.34	0.94
Correlation	0.34	0.95
Tree	0.55	0.83
K-means	0.48	0.89

As observed in these figures, the rain spatial analyses show a good relationship between the Thiessen polygons grouped by all methods and the areas affected by landslides (Babu and Mukesh 2001). However, the *PCA* technique presented better results and embraced the highest number of boroughs reached by landslides if compared to other methods.

Table 2 shows an example of result during the first rainfall event for the rain missing values prediction on validation data through *ANN* simulations. In this table, the best results are related to the *Standard Deviation Ratio* (*SDR*) values nearest to 0 and *Pearson-R Correlation* (*PRC*) values nearest to 1. The rows of this table concerns the results for each statistical method used on rain regionalization. *PCA* and *Correlation* are the two techniques which provide the best results on *ANN* prediction.

The *ANN* setup used to replace the rain missing values was chosen considering the minor *SDR*, the major *PRC*, and the best curves concerning the adherence measured data verus prediction. The results presented in Table 2 clearly show good accuracy of the *ANN* method and the adopted strategy could be considered to replacing the rain missing values.

Knn simulations were carried out to estimate the landslide volume missing values. Several sets of attributes by combining different group of variables to be input of the model was considered. The number of k neighbors was changed during the simulations. The algorithm was also modified in trying to find the best parameters combination. The implemented algorithm is basically the traditional *knn* method, but this new approach considers three search spaces (Souza and Ebecken 2004a, b, c). Each one of these spaces is driven by four taxonomies of the landslides: typology; damage caused; month; daily patterns. All simulations were performed utilizing as input attributes cumulative rainfall indexes and soil parameters. After several simulations, the relative average error decreased below 4 m^3 and the algorithm parameters were considered acceptable. This slipped volume information is acquired in the field by visual inspection and this measurement methodology is very simple implying in data uncertainty. Due to this uncertainty, the proposed method could be considered a great strategy to replace landslide volume missing values. Nine, thirteen, and six are the k optimal number of neighbors related to first, second and, third search space, respectively. The obtained relative average error (in cubic meters) and correlation coefficient were 3.857 and 0.83, respectively for the Manhattan method.

After validating the method, the missing values were replaced by using the best configuration. The Fig. 7 illustrates the results of the statistical distribution before and after knn methods replace-ment. The classes' distribution did not suffer from bias introduction; therefore, this methodology could be successfully applied for this task.

After data preparation, some models were built to predict landslides and rainfall (Souza and Ebecken 2004a, b, c). Table 3 shows the correct classification rate (%) using *ANN* and *Rules Classification*. Besides the *typology* output (*landslides, panic situation* or *non landslides*), models to predict the *volume* (*volume* $= 0$ or *volume* > 0) considering the replacement performed by Euclidian and Manhattan distances were also built, and *damage caused* (*with damage* or *non damage*).

As seen in Table 3, due to the small amount of data to develop these models, only 4 years worth of recordings were available, the obtained results are satisfactory.

Figure 8 shows the results of the next 6 h rainfall volume modeling. This rainfall model shows a very good result (*PRC* $= 0.99$), and, along with *Rule 1*, also could be applied to predict landslides.

The results have shown that adopted methodology can be considered a very efficient and accurate means of landslide prediction, even working with a small amount of data, imbalanced classes, missing values etc. Techniques described in this paper could be integrated to an automatic system in order to improve the alert emission criterion during the rainfall season.

9 Conclusions and Future Works

The landslides study is a very difficult task due to its huge variety (space and temporal) of involved

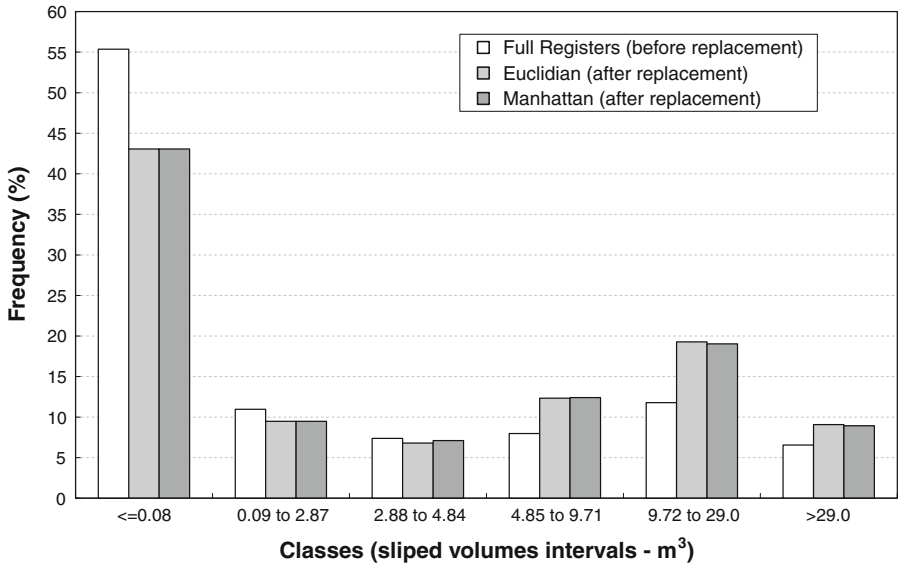


Fig. 7 Statistical distribution—*knn* prediction on validation

Table 3 Classification results

Taxonomies	Classes	Correct classification rate (%)	
		ANN	Rule classification
Typology	Non occurrence	94.1	80.7
	Panic	93.6	89.4
	Landslides	72.4	79.0
Volume (Euclidian)	$V = 0\ m^3$	87.1	89.3
	$V > 0\ m^3$	75.9	88.1
Volume (Manhattan)	$V = 0\ m^3$	90.4	87.3
	$V > 0\ m^3$	74.6	91.3
Consequence	With damage	80.2	91.5
	Non damage	70.8	88.1

parameters. The rain spatial analysis has shown a good relationship between the Thiessen polygons grouped by several methods and the affected areas by landslides (mainly by *PCA* and *Correlation* methods). The *ANN's* prediction used to replace rain missing values also showed very good results demonstrating high reliability in this specific application.

The *Knn* method also presented good performance to estimate the volume missing values. Due to uncertainties in the measurement methodology of the slipped volume (visual inspection), the calculated error was considered excellent.

The quality of the obtained results for landslide and rainfall prediction recommends the insertion of these developed models in the existing alert system. The accuracy would be continuously improved as new registers are inserted in the database. Improvements on the methodology should be also considered because the landslide study is associated to an imbalanced class problem.

The rare objects are often of great interest and great value. It is important to study rarity in the context of Data Mining because rare objects are typically much harder to identify than common objects, and most Data Mining algorithms have a great deal of difficulty dealing with rarity (Weiss 2004).

The imbalanced class problem typically occurs when, in a classification problem, there are many more records from some classes than from others (Chawla et al. 2004). For example, there are very few cases of landslide occurrences if compared to the large number of "*non-occurrences*" at a certain period, even during a rainfall event.

A number of solutions to the imbalanced class problem have been proposed both at the *data* and *algorithm* levels. At the *data* level, these solutions include many different forms of resampling. At the *algorithm* level, solutions include: adjusting the costs

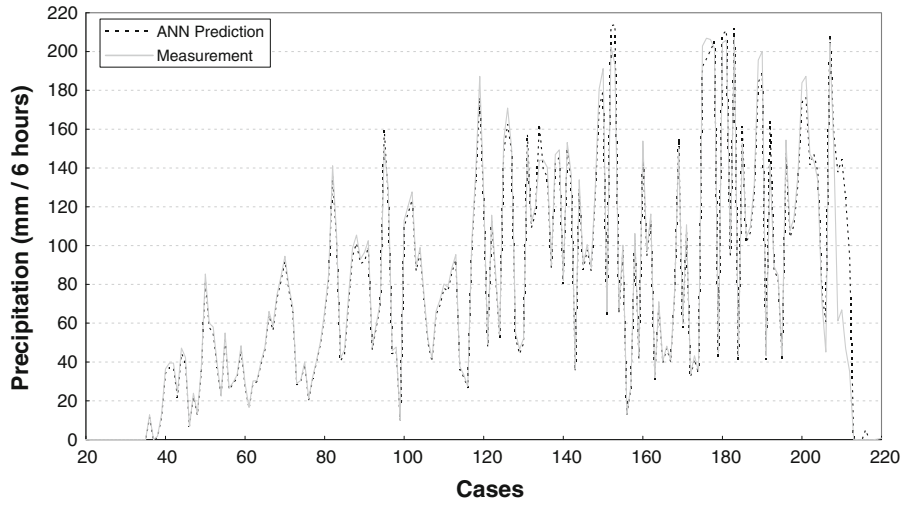


Fig. 8 *ANN* prediction (rainfall)

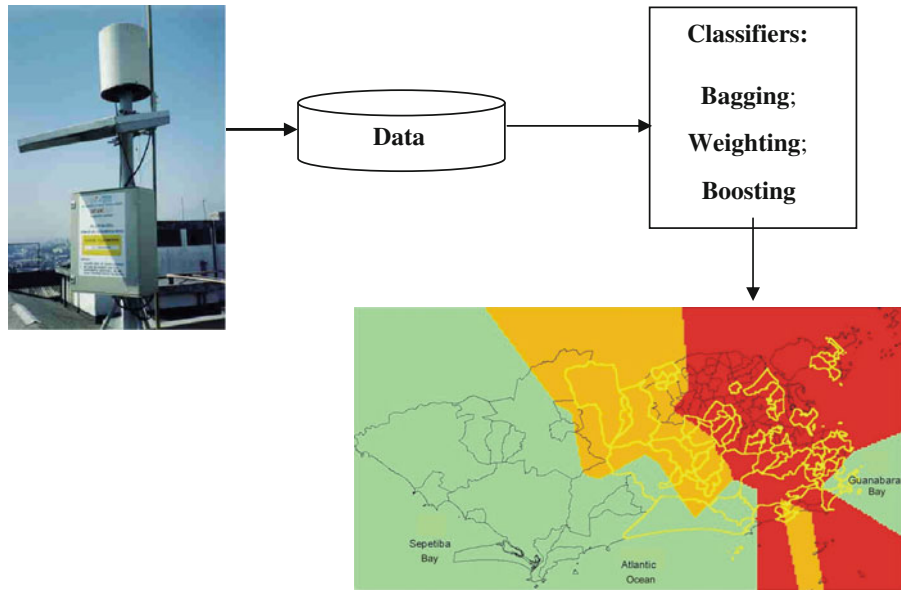


Fig. 9 Proposed model to hazard assessment

of various classes, adjusting the probabilistic estimation, adjusting the decision threshold, among others.

The landslide study is strongly associated with the imbalanced class problem. The adapted *Knn* algorithm used in the volume missing values replacement could be transformed into a classifier also considering the several solutions to treat the imbalanced class problem, such as Boosting (Ting and Zheng 1998), Bagging, Weighting (Duda et al. 2001), etc. A new contribution must also link the results obtained by this ensemble to a *GIS*, generating a landslide risk map with a dynamic provided by rain data in real time, as Fig. 9 illustrates.

Acknowledgments We are deeply grateful to FAPERJ, who provided the financing, as well as the following institutes, which have furnished the data to perform this work: GEORIO; SMAC; SERLA; INMET; UERJ; UFRJ; Wyoming Univ.; DHN; DECEA and CPRM.

References

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. VLDB-94, 1994

Babu GLS, Mukesh MD (2001) Landslide analysis in geographic information systems. Department of Civil Engineering, Indiana Institute of Science, Bangalore

Bishop C (1995) Neural networks for pattern recognition. University Press, Oxford

Chawla N, Japkowicz N, Kolcz A (2004) Special issue on learning from imbalanced datasets (Editorial). In: SIGKDD Explorations, vol 6, pp 1–6

Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley Interscience, New York

Gorsevski PV, Gessler PE, Jankowski P (2003) Integrating a fuzzy k-means classification and a Bayesian approach for spatial prediction of landslide hazard. J Geogr Syst 5(3):223–251

Haikin S (2001) Redes Neurais—Princípios e Prática, 2nd edn. Bookman, Porto Alegre

Han J, Kamber H (2001) Data mining—concepts and techniques, Chapters 7 and 8. Morgan Kaufmann, San Francisco

Hruschka ER, Hruschka Jr ER, Ebecken NFF (2003) Evaluating a nearest-neighbor method to substitute continuous missing values. In: Australian conference on artificial intelligence, pp 723–734

Kaufman L, Rousseeuw PJ (1990) Findings groups in data: an introduction to cluster analysis. In: Wiley series in probability and mathematical statistics, Wiley, New York

Kennedy R, Lee Y, Reed C, Roy BV (1997) Solving pattern recognition problems. Unica Technologies, Lincoln

Lin CY, Chuang CW, Lin WT, Chou WC (2009) Vegetation recovery and landscape change assessment at Chiufenershan landslide area caused by Chichi earthquake in central Taiwan. J Nat Hazards 53(1)

Liu B, Hsu W, Chen S, Ma Y (1998) Integrating classification and association rule mining. In: KDD-98, New York

Liu B, Hsu W, Chen S, Ma Y (2000) Analyzing the subjective interestingness of association rules. In: IEEE Intelligent Systems, National University of Singapore, Singapore

Melchiorrea C, Matteuccib M, Azzonic A, Zanch A (2008) Artificial neural networks and cluster analysis in landslide susceptibility zonation. Geomorphology 94(3–4):379–400

Menezes WF, Paiva LMS, Silva MGAJ, Belassiano M (2000) Estudo do Ambiente Favorável à Propagação de Sistemas Convectivos de Mesoescala sobre o Município do Rio de Janeiro. In: XI Congresso Brasileiro de Meteorologia, Rio de Janeiro

Mitchell TM (1997) Machine learning. McGraw Hill, New York

Pearson K (1896) Regression, heredity, and panmixia. In: Philosophical transactions of the royal society of London, Ser. A, vol 187, pp 253–318

Pyle D (1999) Data preparation for data mining. Morgan Kaufmann Publishers, San Francisco

Rumelhart DE, McClelland J (1986) Parallel distributed processing, vol 1. MIT Press, Cambridge, MA

Souza FT (2004) Predição de Escorregamentos das Encostas do Município do Rio de Janeiro através de Técnicas de Mineração de Dados (in Portuguese). Doctoral thesis, COPPE/UFRJ–Federal University of Rio de Janeiro, Rio de Janeiro

Souza FT, Ebecken NFF (2003) A data mining approach for landslide analysis caused by rainfall in Rio de Janeiro. In: Proceedings of the international conference on slope engineering, Hong Kong, 8–10 December, pp 611–616

Souza FT, Ebecken NFF (2004) Preparação de Dados de Chuvas Intensas utilizando Técnicas de Mineração de Dados (in Portuguese), vol 9, 1st edn. Revista Brasileira de Recursos Hídricos, ABRH, RS, pp 181–187

Souza FT, Ebecken NFF (2004) Landslides data preparation for data mining. In: Proceedings of the IX international symposium on landslides, vol 1, Rio de Janeiro, June and July, Balkema, pp 429–434

Souza FT, Ebecken NFF (2004) A data mining approach to landslides prediction. In: Proceedings of the data mining V—data mining, text mining and their business applications, Malaga, 15–17 September, WITPress, pp 423–432

Ting KM, Zheng Z (1998) Boosting trees for cost-sensitive classifications. In: Proceedings of the tenth European conference on machine learning, LNAI-1398. Springer, Berlin, pp 190–195

Weiss GM (2004) Mining with rarity: a unifying framework. In: SIGKDD Explorations, vol 6, pp 7–19

Wherry RJ (1984) Contributions to correlational analysis. Academic Press, New York