A data mining approach to landslide prediction

F. T. Souza & N. F. F. Ebecken
COPPE / Federal University of Rio de Janeiro, Brazil

Abstract

The study of landslides is a very difficult task due to the huge space-temporal variety of the involved parameters. The study of Rio de Janeiro's city landslide problem has been performed by a Data Mining approach. The dataset related to the landslides registers between 1998 and 2001, including meteorological and soil parameters, are the basis of this work. The cumulative rain patterns related to the landslides depend on the missing data replacement, which was analyzed by several methods, including Clustering and Statistical Analysis. The rain spatial analysis selected the rain gauges to be input on Neural Networks (NN), which were used to replace the missing rain values. The landslide volume variable also presents missing values and their completion has been performed by a *K*-nearest neighbor method. After data preparation, some models (using NN, Interesting Association Rules and Classification Rules) were built to predict the accidents and rainfall, improving the existing alert system.
Keywords: data mining, geographical information systems (GIS), landslides.

1 Introduction

The decontrolled urbanization process in the large cities imposes alterations in the environmental nature equilibrium. Many areas near the slopes are occupied by a vast part of population and these occupations promote the deforesting, soil vegetable cover destroying, garbage accumulation, etc. The new atypical conditions impose a new setup of the soil hillsides, which turned it susceptible to the occurrences of landslides during the rainfall.

Studies Menezes *et al* [1] have been shown that Rio de Janeiro's city has physical characteristics (geographical position and topography) and atmospherically patterns favorable to development of meteorological phenomenon that causes the rainfall. The summer months are rainy and more than 60% of the landslides generally occur during these months.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

This work shows a landslides study performed by *GIS Data Mining* techniques to explicit the important patterns of slope stabilization. The data preparation tasks are very important to the model success, consuming more than 90% of the required time and better expose the information content on the data to the modeling tools. Next section describes the dataset related to the landslides registers and the relevancy of this work, whereas Section 3 presents the tasks required in the dataset preparation, such as: the rainfall spatial analysis; the replacement of the rain missing values using artificial neural network (*ANN*) prediction; the cumulative rain indexes computation associated to each landslide; the jointing of the all involved parameters (rain and soils) into a matrix; the volume missing values prediction using the *K-nearest neighbor* (*KNN*) method. Section 4 shows the models to predict the landslides and rainfall using the *NN*, *Interesting Association Rules* and *Classification Rules*. Section 5 shows the results of the landslides dataset preparation, data modeling and relates a brief discussion, while Section 6 illustrates the conclusions and future works.

2 Viability and relevancy

The landslides occurrences reports describe the characteristics of the accidents, such as location (neighborhood), date, time, typology, caused damage, and volume slipped (estimated in meter cubic). Table 1 illustrates a landslide register obtained at the report of *Geo-Rio*, the Rio de Janeiro Geotechnical Engineering Office, which is responsible for the slopes and hazard assessment.

Table 1: Example of landslide register.

Location (address / neighbor hood)	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	Damage to house

The precipitation registers compose the rain dataset, each 15 minutes interval, collected in 30 automatic rain gauges installed in the Rio de Janeiro's city. This work has been studied 28 rainfall periods or events, extracted from the same period that occurred landslides between 1998 and 2001.

Figure 1 illustrates the Rio de Janeiro neighborhoods map and the rain gauges (dark triangle) network centered in the Thiessen polygons (surroundings area).

The soil parameters related to each city's neighborhood have been monitored by satellite images since 1984. The knowledge of the existent patterns between the several phenomenon related to the landslides occurrences allows the establishment of decision criterions to alert emission and subsequent mobilization of the responsible institutions.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

Legend:

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

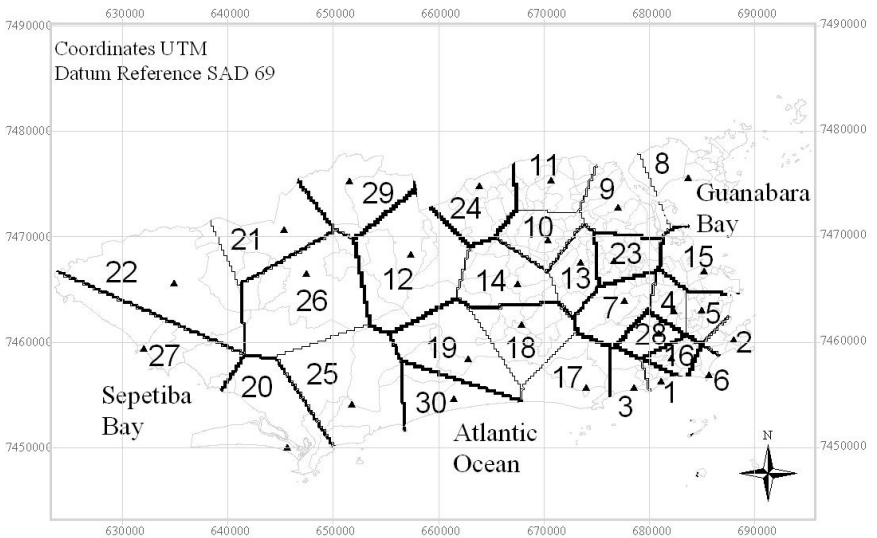

Figure 1: Automatic rain gauges network.

3 Data preparation

The data preparation is the most important part of any project and the dataset carefully prepared better exposes the information content to the modeling tools Pyle [2]. The cumulative rain indexes associated to each landslide have been computed according to the proximity of the accident occurrence (as well the soil parameters). The rain dataset has missing values and their replacement improves the estimative of the rain patterns related to the landslides. This task may be performed using *ANN* prediction and this work has been proposed the *Multi Layers Perceptrons* (*MLP's*) conforming Haikin [3].

The *ANN* method was chosen due its large capacity of learning, generalization and mainly due to the automatic form of knowledge extraction, minus dependent

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

of the subjectivity if compared to the methodologies adopted in the rainfall studies found in the hydrology literature. The used *ANN* architecture to replace the rain missing values must be built with the following set up: the rain gauge data with missing values in the output layer (variable of prediction) and the rain gauges without and with missing values in the input layer (variables of training). Nevertheless the training may not be performed using all the possible data (30 rain gauges) because this issue would use data from distant rain gauges those with missing values and introducing bias or noise. Hence it's previously necessary to perform a rain regionalize. Figure 2 shows the methodology adopted Souza [4] in this work to prepare the landslides dataset.

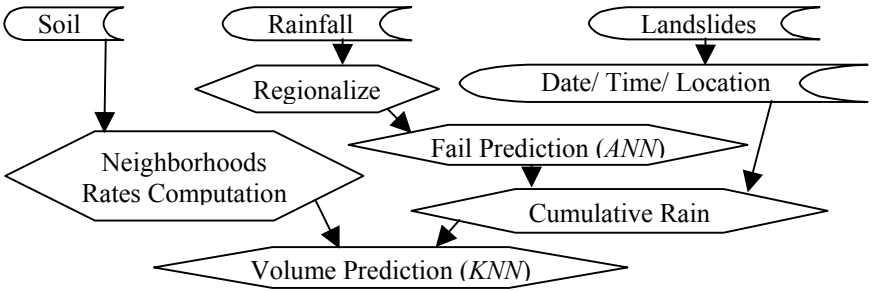

Figure 2: Landslides data preparation tasks.

3.1 Rainfall spatial analyses (regionalize)

The choice of the attributes (spatial rain gauges data) to the training dataset was carried out using four different methods: two statistical (*Principal Component Analyses-PCA* and *correlation*) and two clustering (*K-means* and *tree*) approaches, Souza & Ebecken [5].

PCA can be seen as an axes rotation strategy obtaining a pattern of interpretation easier and clear, through the factors association with high loads to some variables and low to others, Wherry [6].

The *correlation* coefficients express a measured of the relation between two or more variables, Pearson [7]. *Correlation* coefficients can range from -1 to +1. The values of -1, +1 and 0 represent a perfect negative correlation, perfect positive correlation and lack of correlation, respectively.

Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to another, Han and Kamber [8]. *Tree clustering* is a hierarchical method that works by grouping data objects into a tree of clusters using the distances or dissimilarities between the objects [8]. When the data contain a clear "structure" in terms of clusters of objects that are similar to each other, then this structure will often be reflected in the hierarchical tree as distinct "branches". *K-means* partitioning algorithm divides 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 objects.

© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

Once the rain gauges have been grouped by these regionalize technical, the rain dataset may be prepared to the training, test and prediction of the missing values.

3.2 Replacement rain missing values

The substitution of missing values is an important subtask in the data preparation step, Hruschka *et al* [9]. This activity allows the computing of the cumulative rain indexes associated to the landslides. The simulations with *ANN* were performed with 28 rainfall events occurred between 1998 and 2001, because all events presented rain-missing values. When the rain dataset was obtained, only 6 days before the first landslide were furnished. In this sense, the maximum rain index also was limited to six days. Figure 3 shows the issue to the replacement of the rain missing values, Souza & Ebecken [10].

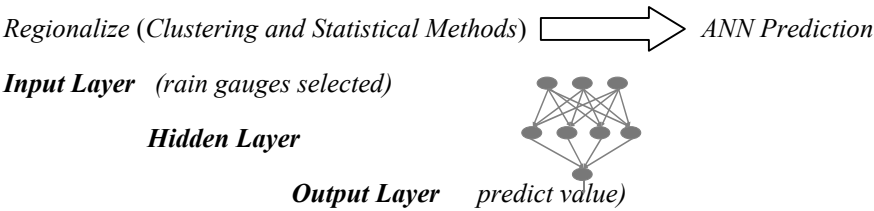

Figure 3: *ANN* architecture setup.

The number of samples used to training, test and validation is variable, because it depends on the beginning of the specific event. Once replaced the rain missing values, it is possible to compute the cumulative rain indexes associated to each landslide. These indexes have been computed according to location (rain data from nearest rain gauge), date and hour of the accident report.

The modeling tools used in this work require a cognitive map that contains all possible parameters related to studied phenomenon. So, the cumulative rain indexes were jointed into a matrix including all involved soil parameters.

This matrix is composed by more than 60 attributes (landslides, rain and soil variables) and 1266 samples. Among these samples, 1033 registers are from landslides or panic occurrences between 1998 to 2001, and 233 patterns were artificially inserted to train also the no landslides occurrences pattern.

As earlier described, there are several registers with volume missing values. To replace the volume missing values *K-nearest-neighbor* (*Knn*) method was adopted, Mitchell [11].

3.3 Replacement volume slipped missing values

Knn method considers that missing values can be substituted by the corresponding attribute value of the most similar complete object in the dataset. This method was applied in this study using two different distance definitions (Euclidean and Manhattan).

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

Let consider 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 will be called $d(i, j)$. Suppose that the *k-th* attribute value ($1 \leq k \leq N$) of the object m is missing. Thus, the Nearest Neighbor Method (*NMM*) will compute the distances $d(m, i)$, for all $i \neq m$, according to the Euclidean or Manhattan distance.

Several simulations with *knn* algorithm were performed to test this method in a sample of 98 registers (10% of the 977 total full registers). These simulations were carried out varying the number of variables (according to the clustering methods), the number of k or neighbors and reported the average error and the correlation. Among these several simulations was extracted the best result (the minimum average errors and the maximum correlation) to complete the volume-missing values. The proposed *knn* method can be easily adapted to datasets formed by discrete attributes, changing the Euclidean / Manhattan distance by the *Simple Matching Approach*, Kaufman & Rousseeuw [12].

4 Data modeling

After data preparation, some models were built, involving three approaches: (1) landslides prediction; (2) association rules extraction; and (3) rainfall prediction. First approach corresponds to models that use two techniques: *NN* and *Classification Rules*, Liu et al. [13]. Second approach extracts rules from the landslides database. Third approach generates a model to predict rainfall considering the past rainfall. The two last approaches furnish one more landslide prediction model, i.e., given a rule describing a landslide occurrence by a particular rain index, and if is possible predict this index; it will be possible associate the landslide risk according to the rule confidence. Next rule illustrates an example:

Rule 1:

IF h_6_>_43.7mm
THEN ->> ACCIDENT
(9.2% 90.6% 117 106) (1)

(in 117 times that six hours cumulative rain index were above 43.7 mm, occurred accident in 90.6% of the times).

5 Results and discussion

As early described, the regionalized rain was carried out considering the whole net of rain gauges and applying data mining techniques to select the rain gauges (input layer) to be used in the *ANN* simulations.

Figures 4 shows an example of the *PCA* method applied to the rain dataset of the Rio de Janeiro's city. This map have a white polygon, at east of the city, which corresponds the surroundings area of the rain gauge with missing values during a rainfall event (January, 1 to 13 - 1998).

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

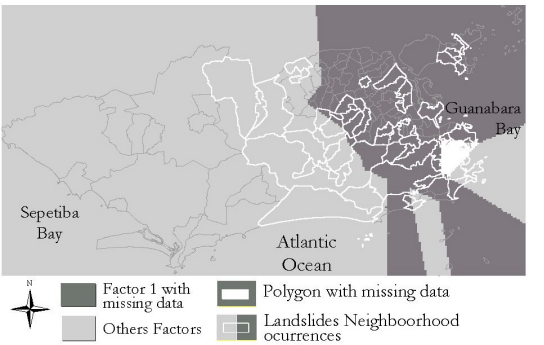

Figure 4: Regionalize with *PCA* Method.

The dark gray polygons correspond to the rain gauges which were selected by the *PCA* regionalize factor, containing the rain gauge with missing values. The neighborhoods with white outline were reached by landslides during this rainfall event. As observed in this figure, the rain spatial analyses showed a good relationship between the Thiessen polygons grouped by *PCA* method and the areas affected by landslides. Table 2 shows the results obtained from simulations with *ANN* during the predictions of the rain missing values (validation of the first rainfall event). The *Standard Deviation Ratio* (*SDR*) is a metric related to error and measured data; a low *SDR* value indicates better prediction. *Pearson-R Correlation* (*PRC*) is a metric related to prediction and measured data; a value nearest to 1 indicates better prediction.

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

The best predictions were obtained from rains regionalize by the statistical methods (*PCA* and *Correlation*).

The *ANN* setup used to completion the rain missing values was chosen considering the minor *SDR*, the major *PRC* and the best adhering curves measured data vs. prediction. The results presented at the Table 2 show clearly the reliability of the *ANN* methods and that the adopted strategy can be considered a good choice to replace the rain missing values.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

Knn simulations were carried out to estimate the volume missing values. It was considered several sets of input attributes, the number of *k* neighbors, and the algorithm also have been modified trying to find the best combinations. The implemented algorithm to this task considered the traditional *knn* method, but the novelty is to consider three-dimensional spaces search, Souza and Ebecken [14]. Each space is driven by the taxonomies of the landslides: i) typology, ii) damage caused, iii) month and iv) diurnal patterns. All performed simulations have utilized as input attributes rain fall indexes and soil parameters. After several simulations some configuration didn't promote any meaningful enhance or improvements in the results, and then the algorithm parameters were considered acceptable. 9, 13 and 6 are the *k* optimal neighbors number related to first, second and third space, respectively. The estimated errors (cubic meters) by Manhattan method in the best setup may be considered excellent (< 4m3) and presents a correlation coefficient greater than 0.83. This information is acquired in the field by a visual inspection and this methodology of measurement introduces data uncertainty.

Once the method was validated, the missing values were replaced using the best setup. The classes' distribution not suffered with bias introduction and this methodology can be applied to this task.

After data preparation, some models were built to predict landslides and rainfall. Table 3 shows the correct classification rate using *ANN* and *Rules Classification*.

As can be observed in this table, due to the lack of data, the obtained results can be classified as satisfactory.

Table 3: Classification results.

Taxonomies	**Classes**	**Correct Classification Rate (%)**	
		ANN	**Rule Classification**
Typology	No 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**
	No Damage	70.8	**88.1**

Figure 5 shows the results of the rainfall modeling. This rainfall model showed very good result (*PRC* 0.99) and can be applied together that Rule 1 (Equation 1) to predict accident or landslides.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

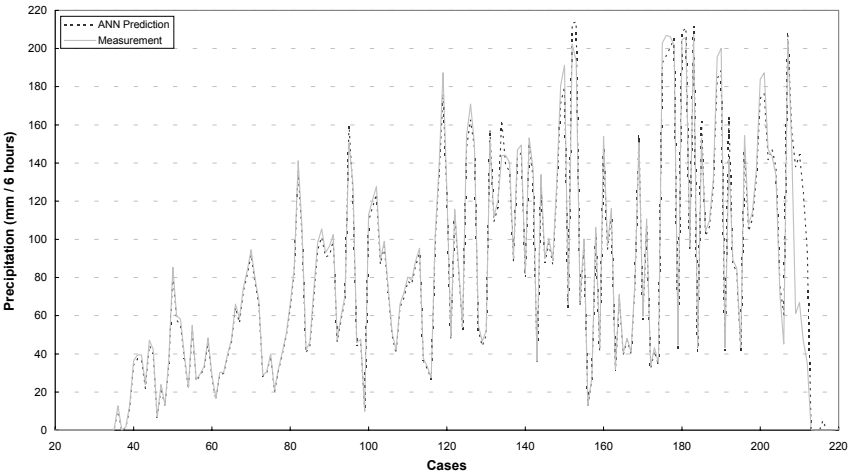

Figure 5: *ANN* prediction (rainfall).

6 Conclusions and future works

The landslides study is a very difficult task due to a huge variety (space and temporal) of the involved parameters. The rain spatial analysis showed a good relationship between the Thiessen polygons grouped by several methods and the areas affected by landslides (mainly by the *PCA* and Correlation methods). The *ANN* prediction replacing rain missing values also showed very goods results and proves that can be used to this task.

The *knn* method also presented a good performance to estimate the volume missing values. Due to uncertainties in the measurement methodology of the volume slipped (visual inspection) the calculated error was considered excellent. The quality of the obtained results recommends the insertion of the developed models in the existent alert system. The accuracy will be continually improved as new registers were inserted in the database.

Boosting [15], Bagging [8] and Weighting [16] can also generate new classifiers ensemble to achieve better performance.

Acknowledgements

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

Reference

[1] Menezes, W. F., Paiva, L. M. S, Silva, M. G. A. J. & Belassiano, M., 2000, Estudo do Ambiente Favorável à Propagação de Sistemas

© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

Convectivos de Mesoescala sobre o Município do Rio de Janeiro, In XI Congresso Brasileiro de Meteorologia, Rio de Janeiro.
[2] Pyle, D., 1999, Data Preparation for Data Mining. San Francisco, California: Morgan Kaufmann Publishers.
[3] Haikin, S, 2001, Redes Neurais – Princípios e Prática. Porto Alegre: Bookman, 2.ed.
[4] Souza, F. T., Predição de Escorregamentos das Encostas do Município do Rio de Janeiro através de Técnicas de Mineração de Dados (*in Portuguese*), 2004, doctoral thesis, Federal University of Rio de Janeiro, Rio de Janeiro.
[5] Souza, F. T & Ebecken, 2004, Preparação de Dados de Chuvas Intensas utilizando Técnicas de Mineração de Dados (*in Portuguese*), Paper in press to publish at the Revista Brasileira de Recursos Hídricos, vol. 9, no. 1, ABRH, RS.
[6] Wherry, R. J., 1984, Contributions to correlational analysis, New York: Academic Press.
[7] Pearson, K., 1896, Regression, heredity, and panmixia, Philosophical Transactions of the Royal Society of London, Ser. A, 187, pp. 253-318.
[8] Han, J., Kamber, H., 2001, Data Mining - Concepts and Techniques – Chapter 7 and 8, San Francisco, California: Morgan Kaufmann, Chapter 7.
[9] Hruschka, E. R., Hruschka Jr, E. R. & Ebecken, N. F. F., 2003, Evaluating a Nearest-Neighbor Method to Substitute Continuous Missing Values, Australian Conference on Artificial Intelligence, 723-734.
[10] Souza, F. T & Ebecken, 2003, A Data Mining Approach for Landslide Analysis Caused by Rainfall in Rio de Janeiro. Paper Published at the International Conference on Slope Engineering, Hong Kong, 8 - 10 December, pp. 611-616.
[11] Mitchell, T. M., 1997, Machine Learning. McGraw-Hill.
[12] Kaufman, L. & Rousseeuw, P. J., 1990, Findings Groups in Data – An Introduction to Cluster Analysis, Wiley Series in Probability and Mathematical Statistics.
[13] Liu, B., Hsu, W., Chen, S., Ma, Y., 1998, "Integrating Classification and Association Rule Mining", KDD-98, August, New York.
[14] Souza, F. T & Ebecken, 2004, Landslides Data Preparation for Data Mining, Paper Published at the IX International Symposium on Landslides, Rio de Janeiro, June and July, Balkema, vol. 1, pp. 429-434.
[15] Ting, K. M., Zheng, Z., 1998, "Boosting Trees for Cost-Sensitive Classifications", Proceedings of the Tenth European Conference on Machine Learning, LNAI-1398, Berlin: Springer-Verlag, pp. 190-195.
[16] Duda, R. O., Hart, P. E., Stork, D. G., 2001, Pattern Classification, Wiley Interscience, Second Edition.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9