Analysis of environmental thresholds for runoff and erosion through qualitative reasoning methods

C.G. Vallejo[1], F. Rodríguez[2], S. Bautista[3] & J.C. Riquelme[1]
[1]*Dpto. de Lenguajes y Sistemas Informáticos, Facultad de Informática y Estadística, Universidad de Sevilla, Spain.*
[2]*Dpto. de Análisis Matemático y Matemática Aplicada, Universidad de Alicante, Spain.*
[3]*Dpto. de Ecología, Universidad de Alicante; and CEAM (Centro de Estudios Ambientales del Mediterráneo), Spain.*

Abstract

Qualitative reasoning methods have proven useful to extract qualitative information from quantitative data in a wide variety of applications. In this work, we analysed a 30 months record of runoff and sediment yield from an experimental station in Benidorm, SE Spain, comprising a total of 104 rainfall events. We adopted a decision-tree approach to identify the environmental variables, basically rainfall characteristics, which were the most informative to predict the production of runoff or sediment and, for these variables, to determine thresholds for different levels of production. A set of 11 rainfall parameters, most of them highly correlated, were simultaneously presented to the C4.5 software tool. Using the variable AI30, which combines the information of height and intensity, 98% of the cases were correctly classified as productive or non-productive for runoff, with an estimate of 5% for the expected error. For the production of sediments, the kinetic energy of the rainfall and the storm duration were incorporated into the prediction vector, giving a 1% of misclassifications and an expected error of 6%. For a subset of cases, the length of the dry period and the antecedent soil moisture were also considered, but no improvement was detected. We also considered the level of production, which was automatically classified as low, medium or high, and obtained decision trees of higher order with higher error rates. The usefulness and drawbacks of the methods for the problems in erosion research considered in this work are discussed.

1 Introduction

The modelling of the hydrological or erosive response of a system may be tackled with many different tools and approaches. Once the system under study is defined (erosion plot, small catchment, river basin, etc.), the modelling tools available range from black-box models, where the inputs of the system are related to its outputs without analysing the system itself, to physically based models, where the relation between the inputs and the outputs emerge from a thorough description of the physical processes occurring inside the system, with many intermediate examples (Bennett[1], Kirkby[2]). To centre the problem considered in this work, let the inputs of the system exclusively be the rainfall events, the outputs of interest being runoff and sediment yield. Whichever type of model is selected, there will be threshold values of the input variables below which there will be no runoff, no sediment yield or both. Moreover, there may be other thresholds defining different regimes of the model, that is, the functions relating the outputs and the inputs may vary in different subsets of the input parameters space. Thus, for instance, a raising in the rainfall intensity may result in an extremely high level of sediment yield.

These thresholds are of special interest in many erosion and hydrological studies. For example, when testing soil conservation measures the best possible result is getting the threshold values increased, so that there is no erosion under normal rainfall conditions and the erosion level is low for heavy rainfall events. On the contrary, the erosion effects of disturbances like wildfires are usually reflected in lower threshold values for a rainfall to be erosive.

The estimation of threshold values for the input parameters that determine, say, no runoff, a low level of production or a high one, involves the classification of the continuous input parameters space into disjoint subsets, so that two different values of one parameter, say rainfall intensity, are assigned to the same class if the labels of their outputs are the same (no, low or high runoff). In this setting, the problem considered is the extraction of qualitative information from quantitative data, a problem that lies under the wide scope of qualitative reasoning methods.

To be precise, the problem we focus on is tackled in this work with machine learning methods, as we do not pursue the analysis further, using the classification obtained to explore the qualitative dynamics of the system, although our results can be easily transformed into a completely qualitative model based on linguistic terms (Aguilar et al.[3]). More specifically, we should speak of supervised machine learning methods, since the labels used for the outputs are assigned using external rules (Mitchell[4]). In particular, we selected the C4.5 software tool (Quinlan[5]), a decision tree builder deemed to be one of the state-of-the-art machine learning classification systems.

The remainder of this paper is organised as follows. In Section 2 we briefly describe the C4.5 system. Section 3 is dedicated to the experimental data and the rainfall parameters we work with. In Section 4 the results of the analysis are summarised. Finally, in Section 5, we discuss the application of the methods used in this work to problems in hydrology and erosion research.

2 The C4.5 system

The problem of supervised classification can be stated in the language of machine learning, in rough terms, as follows. Consider a set of cases, the training set, each of them specifying values for a set of attributes or features and a label. The attributes may have continuous or discrete values, while the label may only have discrete values, which correspond to different classes. The problem is to obtain decision rules, data structures or algorithms from the training set so that the class of a case is predicted in terms of the values of its attributes. Of course, there are many properties that a useful classification algorithm should have, more or less critical depending on the intended use (computation efficiency, scalability, low classification error, high prediction accuracy, etc.).

A common approach to supervised classification is the use of decision trees. From the root to the leaves (the terminal nodes), each non-leaf node is associated with a test on the values of the attributes, the different outcomes corresponding to the children of the node (the subtrees emerging from that node). Each leaf is associated with a class. For a given case, the values of its attributes imply a path from the root to a certain leaf, whose associated class is the predicted class for the case. A decision tree can also be converted to a set of rules.

The C4.5 system is the "classic" decision tree tool. It traces back to the ID3 system (Quinlan[6]) and is the reference to compare with when new methods are proposed. We only give here a brief description of the system; for details and the code see Quinlan[5].

C4.5 uses the standard technique for building classification trees from data, the so-called recursive partitioning algorithm, a divide and conquer strategy also used in ID3 and CART (Breiman et al.[7]), a classifier widely used in applied statistics and data mining. In C4.5, the test at each node is performed over one attribute. For a continuous attribute A, the test is of the form $A \leq t$, with the mutually exclusive outcomes *true* and *false*. The selection of the test is based on a splitting criterion that is to be maximised, the gain-ratio, an information-based measure that takes into account the different possible outcomes. This selection implies choosing one of the attributes and finding the best threshold t. Thus, two problems of interest for the applications considered in this work are resolved at once: the selection of the most informative features from a set of attributes, possibly numerous and highly correlated, and the estimation of threshold values for the selected parameters.

Once the decision tree has been constructed, the C4.5 system analyses whether it can be simplified (pruned), cutting paths without increasing the classification error over a certain confidence level. C4.5 evaluates the tree on the training set, obtaining the percentage of cases misclassified, and also on a random subset (test set) to estimate the expected classification error or prediction accuracy of the tree. The C4.5 system also includes a rule constructor (C4.5 rules) to convert the tree into a set of rules, so that the output is easily understood by a human being. An example of C4.5 output is shown in Figure 1.

The most recent versions of C4.5 have overcome a certain weakness of the original system in the treatment of continuous attributes (Quinlan[8]). Moreover, in recent comparisons with other classification algorithms (Lim et al.[9]), C4.5 has provided good classification accuracy combined with a fast execution on large datasets.

Machine learning classification algorithms, and specifically the C4.5 system, have been applied in a wide variety of fields. The most typical application to be thought of might be the diagnosis and prognosis in medicine (Masic et al.[10], Zupan et al.[11], Laurikkala et al.[12]), but some recent examples of applications include such diverse fields as chemical and electrical engineering (Mulholland et al.[13], Talaie et al.[14], Karunadasa et al.[15]), image processing (Linhui & Kitchen[16]), veterinary science (Scott et al.[17], Stark & Pfeiffer[18]), meteorology (Tag & Peak[19]), software engineering (De Almeida et al.[20]) or speech and hand-written language recognition (Samouelian[21], Amin & Singh[22], Amin[23]).

3 Experimental data

Data were recorded in an experimental station located at Benidorm, on the southeast coast of Valencia Region, Spain. The mean annual precipitation is 293 mm, with 50% coming as autumn storms. The mean annual temperature is 19° C. The soil at the study site was a Xeric Torriorthent developed over marls and limestone colluvium. Stone surface cover was about 50%. A sizeable forest fire occurred in August 1992 affecting near 500 ha of pine forest. The experimental station included erosion plots in different zones, rain gauges and an autographic gauger recorder (see Bautista[24] or Bautista et al.[25] for details). The data used in this work comes from an erosion plot located in the burned area. Runoff and sediment yield were collected from September 1993 to May 1996. The record comprises 104 rainfall events.

From the bands recorded by the autographic gauger, several basic rainfall descriptors were obtained. These included the height (amount) of precipitation (Precip), the duration of the rainfall event (Dur), the effective duration (EfDur) and the maximum intensity in periods of 10 minutes (I10) and 30 minutes (I30). Some other indexes, proposed in the literature as being good predictors of rainfall erosivity, were also computed. The kinetic energy of the rainfall is one of the factors affecting erosivity. There are empirical linear relations between the kinetic energy and the logarithm of the intensity, obtained in different climatic zones. We used the relation estimated by Zanchi & Torri[26] in the Mediterranean region. To compute the index Ketot (total kinetic energy), the storm is divided into small time increments of uniform intensity. For each time period, the kinetic energy is estimated through the empirical relation with the intensity and, summing for all the time periods, the total kinetic energy of the storm is obtained (Morgan[27]). Several authors have proposed to compute only the kinetic energy of those periods where the intensity is greater than a certain value. We used the index Ke>5, where only the periods with intensity greater than 5 mm h^{-1} are considered, and the index Ke>10, accounting for those periods

with intensity greater than 10 mm h^{-1}. Some compound indexes have also been proposed as more accurate predictors for the erosivity of a storm. The index EI30 (Wischmeier[28]) is the product of kinetic energy and the maximum 30-min rainfall intensity (Ketot x I30), and is widely used as the factor R (erosivity factor) in the well known Universal Soil-Loss Equation (Wischmeier[28], Wischmeier & Smith[29]). The index AIm (Lal[30]) is the product of the amount of precipitation and the maximum intensity in periods of *m* minutes. We used the indexes AI10 (Precip x I10) and AI30 (Precip x I30). As could be expected, most of these rainfall parameters were highly correlated (Table 1).

Table 1: Spearman rank correlation coefficients between the rainfall parameters considered (see text for description of the parameters). All correlations were highly significant (P<0.001, n=104).

	I10	I30	Ketot	Ke>5	Ke>10	EI30	AI10	AI30	Dur	EfDur
Precip	0.840	0.907	0.990	0.718	0.486	0.972	0.970	0.983	0.796	0.836
I10		0.960	0.900	0.816	0.659	0.929	0.942	0.906	0.467	0.456
I30			0.945	0.808	0.630	0.974	0.964	0.965	0.570	0.572
Ketot				0.760	0.541	0.993	0.990	0.995	0.729	0.759
Ke>5					0.770	0.791	0.796	0.772	0.464	0.409
Ke>10						0.587	0.599	0.567	0.244	0.153
EI30							0.995	0.997	0.683	0.701
AI10								0.991	0.681	0.699
AI30									0.721	0.741
Dur										0.918

For a subset of cases (47 rainfall events) the length of the dry period (number of days between two consecutive storms) and the antecedent soil moisture were also considered. Soil moisture was determined by the gravimetric method, at different moments during the recording period. When two or more soil moisture measures were available for a dry period, an exponential decay model was fitted to the data, thus allowing to estimate the soil moisture just before the next storm (antecedent soil moisture).

4 Results

Each case (rainfall event) was labelled as *productive* or *non-productive*, for runoff or sediment, accordingly with the corresponding value (positive or zero) of runoff or sediment yield. The 11 rainfall parameters described in the previous section (see Table 1) were simultaneously presented to the C4.5 system for the set of 104 cases.

Figure 1 shows the C4.5 output for the analysis of runoff. The simplified decision tree is based only on the parameter AI30, with an estimated threshold value of 11.7 mm^2 h^{-1} below which there is no production of runoff. This extremely simple tree correctly classified more than 98% of cases (only two cases were misclassified), with an estimate of 5% for the expected classification error.

```
Read 104 cases (11 attributes) from s3-sn.data

Decision Tree:
AI30 <= 11.7 : NO (63.0)
AI30 > 11.7 :
|   Precip > 4.9 : YES (34.0)
|   Precip <= 4.9 :
|   |   AI10 > 33.6 : YES (3.0)
|   |   AI10 <= 33.6 :
|   |   |   I30 <= 4.4 : YES (2.0)
|   |   |   I30 > 4.4 : NO (2.0)

Simplified Decision Tree:
AI30 <= 11.7 : NO (63.0/1.4)
AI30 > 11.7 : YES (41.0/3.8)

Tree saved

Evaluation on training data (104 items):
    Before Pruning           After Pruning
   ----------------   ---------------------------
   Size      Errors   Size      Errors   Estimate
      9    0( 0.0%)      3    2( 1.9%)    ( 5.0%)
```

Figure 1: C4.5 output for runoff data analysis. Cases were labelled as 'YES' (productive) or 'NO' (non-productive).

The decision tree for sediment yield also included the index AI30, with a threshold value of 14.82 $mm^2\ h^{-1}$ below which there is no sediment yield, but the kinetic energy (through the index EI30) and the duration of the storm (Dur) were also incorporated into the prediction vector. If AI30>14.82, then the case was classified as *productive* if EI30>508.75 J m^{-2} mm h^{-1}; else, the case was classified as *productive* or *non-productive* depending on the duration of the storm (*productive* if Dur≤8.5 h; *non-productive* if Dur>8.5 h). This decision tree only misclassified one case (less than 1% of error), with an expected classification error of 6%. Anyway, C4.5 computes these expected errors in a very pessimistic way.

When the length of the dry period and the antecedent soil moisture were also presented to C4.5 (for the subset of 47 cases), none of these parameters were incorporated into the decision tree. Moreover, the decision tree for runoff was very similar to the tree obtained with the whole dataset –only the index AI30 was used, with a threshold value of 11.66 $mm^2\ h^{-1}$. Unlike this, the trees for sediment yield were different for the whole dataset and for the subset of 47 cases, where a simple test on the index AI10 with a threshold value of 29.4 $mm^2\ h^{-1}$ correctly classified all cases.

We also considered the level of production, which was classified as low, medium or high using the quartiles of the distribution of productive events. Thus,

the label of each case had four possible values (*non-productive*, *low*, *medium* and *high*). The decision trees obtained were more complex than in the previous analysis (the size of the tree, that is, the total number of nodes, was 13 both for runoff and erosion), with higher classification errors (5.8% and 2.9% for runoff and erosion, respectively) and very high expected errors (14.6% for runoff and 11.3% for erosion).

5 Discussion

The results described in this work prove the usefulness of machine learning supervised classification methods, and specifically of C4.5, to select the most relevant rainfall parameters that determine the existence of runoff or erosion, as well as to estimate the threshold values that define productive and non-productive storms.

In erosion studies, it is usual to choose the rainfall parameter that best correlates with runoff or sediment yield, or that provides a good fitting in regression analysis relating the production (of runoff or sediments) and the rainfall variable. Then, the functional relation estimated in this, possibly non-linear, regression analysis may be used to estimate threshold values of the rainfall variable for different levels of production. Obviously, having a good functional relationship with production is unnecessary for a rainfall parameter to be a good discriminant of the levels of production. Also, thresholds estimated from regression analysis heavily depend on the particular form of the functional relation and the quality of the fitting –and, moreover, threshold values for the production to be zero or positive are usually estimated extrapolating a functional relation obtained from productive storms only. Another problem with regression analysis when trying to incorporate more than one rainfall variable is the existence of colinearities between highly correlated variables. In sum, the more direct and qualitative approach used in this work seems to be preferable when only qualitative information, like threshold levels, is the main concern.

The results obtained with C4.5 are not only useful from a formal point of view (simple trees with good prediction accuracies), but they are also sound for the expert in hydrology or erosion. While the amount of precipitation and the rainfall intensity (through the index AI30) explained well the production of runoff, the incorporation of the kinetic energy of the storm (through the index EI30) for sediment yield prediction agrees with the well known importance of this rainfall parameter for a storm to be erosive (Morgan[27], Wischmeier[28], Obi & Salako[31]). Likewise, the exclusion of the length of the dry period and the antecedent soil moisture, when these attributes were considered, support the low importance of these factors in the environmental conditions considered in this work (Bautista[24]).

There are, however, some drawbacks of C4.5, and any machine learning classification method, that should not be forgotten. The main requirement for these methods to be effective is having a training set not too small. The minimum size depends, among other factors, on the number of classes and the number of cases in each class but, as a rule of thumb, at least 100 cases should be

available. Two opposite indications about this size requirement are given in our results. While the decision tree for *productive* or *non-productive* runoff was remarkably the same with 104 and 47 cases, for sediment yield the selected attributes were different. Also, when the level of production was considered, C4.5 constructed complex trees with high expected classification errors. It should be noted that, although the total size of the training set was 104, the number of productive cases –to be assigned to three different classes (*low*, *medium* or *high*)– were only 39 for runoff and 34 for sediment yield.

When there is a very large dataset available, and the behaviour of the system is sufficiently regular, other methods like regression analysis might be of election and give a quantitative information that qualitative methods are no intended for. Thus, machine learning classification methods may be especially useful for moderately large datasets, or for large problems where the system under study shows a high variability that makes difficult to obtain good quantitative relations. Nonetheless, the capacity of C4.5 and other machine learning methods to work with a high number of attributes, discrete and continuous, and to select the most informative even in the presence of high mutual correlations, makes them a tool that may prove worthy for the researcher in hydrology and erosion.

Acknowledgements

The work of S. Bautista was partly funded by Generalitat Valenciana and Fundación Bancaja.

References

[1] Bennett, J.P. Concepts of mathematical modelling of sediment yield. *Water Resources Research*, **10(3)**, pp. 485-492, 1974.

[2] Kirkby, M.J. Hillslope runoff processes and models. *Journal of Hydrology*, **100**, pp. 315-339, 1988.

[3] Aguilar, J., Riquelme, J. & Toro, M. A tool to obtain hierarchical qualitative rules from quantitative data. *Methodology and Tools in Knowledge-Based Systems*, eds. J. Mira, A. Pasqual del Pobil & M. Ali, Lecture Notes in Artificial Intelligence, **1415**, pp. 336-346, Springer Verlag: Berlin, 1998.

[4] Mitchell, T.M. *Machine Learning*, McGraw-Hill: New York, 1997.

[5] Quinlan, J.R. *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers: San Mateo, 1993.

[6] Quinlan, J.R. Induction of decision trees. *Machine Learning*, **1**, pp. 81-106, 1986.

[7] Breiman, L., Friedman, J., Olshen, R. & Stone, C. *Classification and Regression Trees*, Wadsworth: Belmont, 1984.

[8] Quinlan, J.R. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, **4**, pp. 77-90, 1996.

[9] Lim, T.S., Loh, W.Y. & Shih, Y.S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, **40(3)**, pp. 203-228, 2000.

[10] Masic, N., Gagro, A., Rabatic, S., Sabioncello, A., Dasic, G., Jaksic, B. & Vitale B. Decision-tree approach to the immunophenotype-based prognosis of the B-cell chronic lymphocytic leukemia. *American Journal of Hematology*, **59 (2)**, pp. 143-148, 1998.

[11] Zupan, B., Stokic, D.S., Bohanec, M., Priebe, M.M. & Sherwood, A.M. Relating clinical and neurophysiological assessment of spasticity by machine learning. *International Journal of Medical Informatics*, **49(2)**, pp. 243-251, 1998.

[12] Laurikkala, J., Juhola, M., Lammi, S. & Viikki, K. Comparison of genetic algorithms and other classification methods in the diagnosis of female urinary incontinence. *Methods of Information in Medicine*, **38(2)**, pp. 125-131, 1999.

[13] Mulholland, M., Hibbert, D.B., Haddad, P.R. & Sammut, C. Application of the C4.5 classifier to building an expert system for ion chromatography. *Chemometrics and Intelligent Laboratory Systems*, **27(1)**, pp. 95-104, 1995.

[14] Talaie, A., Esmaili. N., Taguchi, T., Romagnoli, J.A. & Talaie, F. Application of an engineering algorithm with software code (C4.5) for specific ion detection in the chemical industries. *Advances in Engineering Software*, **30(1)**, pp. 13-19, 1999.

[15] Karunadasa, P.S., Annakkage, U.D., MacDonald, B.A. & Pahalawaththa, N.C. Static security assessment using a decision tree technique. *Journal of Electrical and Electronics Engineering*, Australia, **19(1-2)**, pp. 17-24, 1999.

[16] Linhui, J. & Kitchen, L. Object-based image similarity computation using inductive learning of contour-segment relations. *IEEE Transactions on Image Processing*, **9(1)**, pp. 80-87, 2000.

[17] Scott, M.R., Sherlock, R.A. & Smith, L.A. An investigation into the use of machine learning for determining oestrus in cows. *Computers and Electronics in Agriculture*, **15 (3)**, pp. 195-213, 1996.

[18] Stark, K.D.C. & Pfeiffer, D.U. The application of non-parametric techniques to solve classification problems in complex data sets in veterinary epidemiology-an example. *Intelligent Data Analysis*, **3(1)**, pp. 23-35. 1999.

[19] Tag, P.M. & Peak, J.E. Machine learning of maritime fog forecast rules. *Journal of Applied Meteorology*, **35(5)**, pp. 714-724, 1996.

[20] De Almeida, M.A., Lounis, H. & Melo, W.L. An investigation on the use of machine learned models for estimating software correctability. *International Journal of Software Engineering and Knowledge Engineering*, **9(5)**, pp. 565-593, 1999.

[21] Samouelian, A. Frame-level phoneme classification using inductive inference. *Computer Speech and Language*, **11(3)**, pp. 161-186, 1997.

[22] Amin, A. & Singh, S. Recognition of hand-printed Chinese characters using decision trees/machine learning C4.5 system. *Pattern Analysis and Applications*, **1(2)**, pp. 130-141, 1998.

[23] Amin, A. Recognition of hand-printed Latin characters based on generalized Hough transform and decision tree learning techniques. *International Journal of Pattern Recognition and Artificial Intelligence*, **14(3)**, pp. 369-387, 2000.

[24] Bautista, S. *Regeneración post-incendio de un pinar (Pinus halepensis, Miller) en ambiente semiárido. Erosión del suelo y medidas de conservación a corto plazo*, Ph. D. Thesis, University of Alicante, 1999.

[25] Bautista, S., Bellot, J. & Vallejo, V.R. Mulching treatment for post-fire soil conservation in a semiarid ecosystem. *Arid Soil Research and Rehabilitation*, **10(3)**, pp. 235-242, 1996.

[26] Zanchi, C. & Torri, D. Evaluation of rainfall energy in central Italy. *Assessment of erosion*, eds. De Boodt & D. Gabriels, John Wiley & Sons: Chichester, pp. 133-142, 1980.

[27] Morgan, R.P.C. *Soil erosion and conservation*, Longman Scientific & Technical: Essex, 1986.

[28] Wischmeier, W.H. A rainfall erosion index for a Universal Soil-Loss Equation. *Soil Science Society of America Proceedings*, **23**, pp. 246-249, 1959.

[29] Wischmeier, W.H. & Smith, D.D. *Predicting rainfall erosion losses –a guide to conservation planning*. Agriculture Handbook **537**. USDA: Washington D.C., 1978.

[30] Lal, R. *Soil erosion problems on an Alfisol in Western Nigeria and their control*, IITA Monograph **1**, International Institute of Tropical Agriculture: Ibadan, 1976.

[31] Obi, M.E. & Salako, F.K. Rainfall parameters influencing erosivity in southeastern Nigeria. *Catena*, **24(4)**, pp. 275-287, 1995.