DATA MINING APPLIED FOR DETERMINING STREAM FLOW PERMANENCE

P. Mallmann [1], A. Montibeller [1], F. Hino [1], M. Vilela [1], M. Nadas [1], F. Caruso Jr.[1], H. Delabary [2]

[1] Caruso Soluções Ambientais Inovadoras - 88015-120 Florianópolis, Santa Catarina, Brazil – (pedro, marcos, athila, fabiano, micael)@carusojrea.com.br

[2] Centrais Elétricas de Santa Catarina - CELESC - 88034-900 Florianópolis, Santa Catarina, Brazil – henriquesd@celesc.com.br

Commission III, TC III/1

KEY WORDS: Flow Permanence, Mini-basins, Data Mining, Decision Tree, Morphometric Attributes

ABSTRACT:

Streamflow permanence is an aspect of great legal importance in Brazil, because streams, depending on their flow regime, are protected or not by law. Various methods, from field methods to computational methods, are used to determine the flow regime of streams, however some are too time spending and computational methods usually need gaging information as input. Furthermore, computational methods used to extract drainage networks do not identify the flow regimes of streams, and modelled drainage networks always need to be refined manually, as some authors indicate that up to 55% of modelled drainage length is ephemeral in some cases. This work proposes a semi-automatic computational method to determine the flow regime of first order streams, which uses 11 morphometric attributes of the mini drainage basins of these streams to develop a classification model using decision tree algorithms. WEKA package was used to perform the data mining process, which resulted on the development of a compact 8 node decision tree. Ten-fold stratified cross-validation was used to validate the model, which obtained an accuracy rate of 70%. The drainage network of the study area extracted with the classical approach was refined after the result of the classification was obtained. Quantitative analysis of channel length by Strahler order shows an overall reduction of 25% in channel length after refinement was undertaken, and for 1st order streams, as much as 31% were classified as ephemeral. Modelling the drainage areas of headwater streams represents a new approach to determining stream flow permanence, and inclusion of new attributes in the model may yield better results in future research.

1. INTRODUCTION

The flow regime of rivers is an aspect of great legal importance in Brazil, as the perennial and intermittent streams are protected by the Brazilian Forest Code (BRASIL, 2012), while ephemeral ones are not. However, there is not an established procedure to determine the flow regime of stream, so several methodologies are used by the public and private organizations, which results in diverse interpretations and without standardization.

The methods of determination of the flow regime of stream can be separated into two groups: a) field methods, which evaluate in situ several physical and biological characteristics of drainage; and b) computational methods, which use various data to model the flow regime. The first is applicable to small areas with few streams, while the second can be applied to larger areas, but depends on the availability of the data needed for modeling.

Junior & Andreoli (2015) used hydrological parameters and water balance to determine the best timing of the year for field investigations of stream; Panero *et al* (2006) performed a principal component analysis on hydrochemical data to classify perennial and intermittent rivers; the NC Division of Water Quality (2005) has drawn up a manual and a form to determine the flow regime of stream in the field; Porras & Scoggins (2013) determined the probability of stream being perennial, through statistical analysis of river monitoring data; Jaeger *et al* (2019) used the PROSPER model, developed by the USGS to determine the probability of stream being perennial; Williamson *et al* (2015) used the TOPMODEL model to classify the flow regime of stream. Computational methods often use river monitoring data to model the flow regime and are not applicable to areas where this data is not available.

It is known that the lower the Strahler order of a stream, the greater the chance of it being ephemeral (NC Division of Water Quality, 2005). This is true for hydrological models used in the extraction of drainage networks, in which much of the hydrography generated is ephemeral, because the models either perform the extraction of features from the terrain (data mining models), or are obtained by the classical methodology, based on a contribution area threshold above which the pixels are considered as part of the drainage network (O'CALLAGHAN & MARK, 1984). The consequence of this is that the drainage networks generated by computational methods do not identify the flow regime of the stream, and in some cases the percentage of ephemeral stream reaches 55% of the total modeled length (HANSEN, 2001).

Thus, the present work proposes a methodology of semi-automatic determination of the flow regime of first order streams, by the classification (perennial or ephemeral) of the drainage basins of each of these streams in the study area, according to their individual morphometric characteristics.

In this work perennial river are those which maintain streamflow over at least 90% of the year; intermittent rivers maintain streamflow only during the rainy season and are associated to arid climates; and ephemeral streams only exist during or immediately after a rainfall event.

2. MATERIALS AND METHODS

2.1 Study Area

The study area is located in the southern part of Brazil, city of Joinville/SC, between the coordinates 26°13'01"S, 49°04'06"W and 26°24'44"S, 48°51'44"W (Figure 1), where there are different patterns of relief, from flat to mountainous. At the flat areas rice crops dominate, whilst at the mountains natural forest cover prevails. The climatic regime is humid subtropical, and the yearly pluviosity rates are around 2.500mm. There are perennial and ephemeral streams in the area, but there are not intermittent streams, for these occur in association with arid climatic regimes.

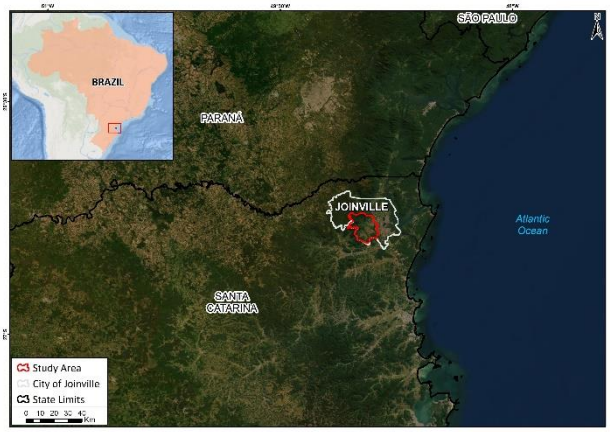

Figure 1: Location of the study area in Joinville/SC.

2.2 Method

The main idea is that basins with similar morphometric characteristics will yield streams with similar hydrologic regimes. Thus, samples (Regions of Interest - ROIs) of perennial and ephemeral streams were collected at the study area in a data mining process. Each ROI contains information of the eleven morphometric attributes of the basins and serves as input to building the classification model in WEKA.

The sequence of procedures for obtaining the classification model of the mini-basins comprises nine main steps: a) extraction of the drainage network; b) delimitation of mini-basins; c) extraction of morphometric attributes; d) band stacking; e) establishment of field control points (perennial and ephemeral streams); f) collect ROIs over the control points; (g) obtaining the decision tree; h) classification of the mini-basins and; (i) evaluation of the results.

The drainage network of the study area and the limits of the mini-basins were extracted from a 1m resolution DEM (SDS, 2011) using the hydrological tools of TerraHidro, and the morphometric attributes used in data mining were generated in ArcGIS, including area, perimeter, circularity index, hypsometric amplitude and average slope. Collection of ROIs was done in ENVI, and data analyses were performed by WEKA. Once the decision tree was built by J48 algorithm, it was transposed to ENVI once again to read classify the mini-basins as perennial or ephemeral based on the morphometric attributes bandstack.

ROI sampling is the most important part of the process, for it is from these samples that the J48 algorithm performs the statistical analyses which results in a decision tree customized for this data. A field investigation was undertaken, to check the flow regime of streams both in the plains and in the hilly areas. It was done after a 10-day dry period, to ensure there was no influence from recent rainfall events on the streams. As expected, the drainage density in the hilly areas was much higher than in the plains, where most rivers are channelized.

2.3 Extraction of Mini-basins

According to Guerra (1993), drainage basins are areas drained by a river and its tributaries from the headwaters to the outfall, and are limited by watersheds, where rainwater flows, forming rivers and stream, or seeps into the soil recharging aquifers. Similarly, mini-basins are the drainage areas of individual segments of streams, from the watershed to the point of confluence with another stream.

TerraHidro application (ROSIM, 2008) is available in TerraView software (CÂMARA *et al*, 2000), and makes hydrological analyses from DEMs, including the delimitation of mini-basins, based on the flow directions and segments of the drainage network extracted from the DEM. In this work, a drainage network extracted with the classical method (based only on hydrological attributes) was used, with a contribution area threshold equal to 15000m². If a higher threshold was used (e.g. 30.000), the springs would be displaced downstream and the drainage density would be reduced, which would result in the delimitation of mini-basins larger than those generated with threshold 15000 (Figure 2).

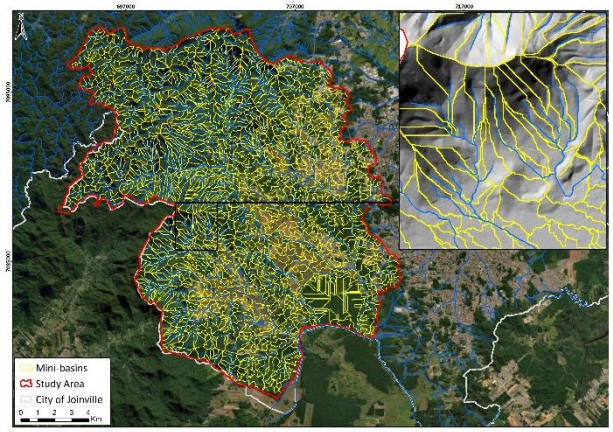

Figure 2: Mini-basins and hydrography of the study area.

In total, 1635 mini-basins were modeled, with a mean area of about 16 hectares. At the mountainous areas they tend to be smaller and more elongated than at the plain, which indicates that geologic structures play an important role in the formation of valleys and streams in the study area. First order mini-basins area account for almost half (45%) of the study area.

It is important to highlight that the modeling proposed in this work applies only to first order mini-basins, which are the ones with the greatest chance of being ephemeral, according to the NC Division of Water Quality (2005). In addition, first-order streams generally account for about 50% of the total river length in drainage basins (HANSEN, 2001), with great importance for the maintenance of water resources in a region (WOHL, 2017).

2.4 Extraction of Attributes

Eleven morphometric attributes were used: area, perimeter, mean slope, standard slope deviation, hypsometric amplitude, circularity index, compactness coefficient, relative perimeter, relative relief, mean roughness and roughness concentration index (RCI). Morphometric attributes represent characteristics of the mini-basins and are extracted from the DEM using geoprocessing tools. Table 1 presents the attributes, formulas and corresponding references.

Attribute	Formula	Reference
Area (A)	GIS Analysis	Schumm (1956)
Perimeter (P)	GIS Analysis	Schumm (1956)
Mean Slope (Dm)	GIS Analysis	Wenthworth (1930)
Slope Standard Deviation (Sstd)	GIS Analysis	-
Hipsometric Amplitude (H)	H = Z - z	Strahler (1952)
Circularity Index (Ci)	$Ci = 12.57 \times (A/P^2)$	Miller (1953)
Compacity Coefficient (Cc)	Cc = 0.2841 x (P/VA)	Gravelius (1914)
Relative Perimeter (Pr)	Pr = A/P	Schumm (1956)
Relativo Relief (Rr)	Rr = H x 100/P	Schumm (1956)
Mean Roughness (Rm)	GIS Analysis	-
Roughness Concentration Index (RCI)	GIS Analysis	Sampaio (2014)

Table 1: Morphometric attributes used in the data mining process, extracted from the DEM of the study area (SDS, 2011).

The rasters of the morphometric attributes were processed in ArcGIS, using zonal statistics and map algebra tools. Pixel size was reduced to 10m, to increase processing speed. The eleven attribute raster bands were stacked in ENVI and formed the database on which ROIs were collected.

Mean slope, standard slope deviation hypsometric amplitude, relative relief, mean roughness and RCI (Roughness Concentration Index) represent changes on the terrain, as it gets steeper or flatter. Area, perimeter, circularity index and relative perimeter describe geometrical aspects of the basins.

At the end, only six attributes were considered for modelling: RCI, relative perimeter, hypsometric amplitude, area, medium slope and relative relief.

Area and perimeter were obtained from simple calculate geometry tools in GIS. Mean slope and Slope Standard Deviation were obtained from zonal statistics tools, using the mini-basins as input zone. Circularity Index, Compacity Coefficient, Relative Perimeter and Relative Relief were obtained applying the formulas presented in Table 1 using map algebra.

RCI represents the degree of roughness of the terrain, which changes as relief gets flatter or steeper. At the study area, its values vary from 0 to 4 and represent five classes of terrain types: flat, undulated, strongly undulated, mountainous and escarpments.

2.5 Training Samples

In the study area, 84 training samples were selected, divided equally between perennials and ephemera (Figure 3). Sampling was done based on field surveys and visual interpretation. A good spatial distribution was observed for the relief units present in the area.

The samples were collected only in the first order mini-basins, as the focus of this work is to identify the flow regime of first order streams.

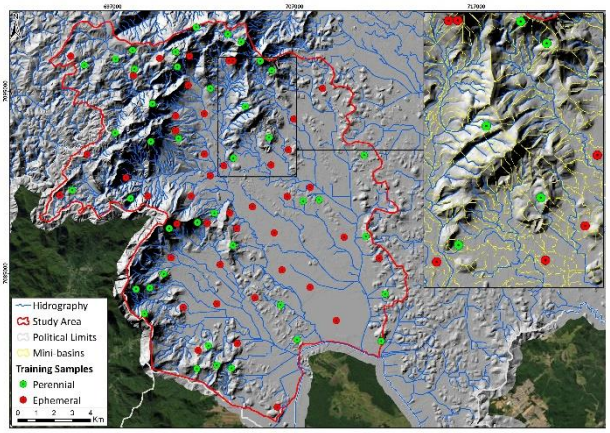

Figure 3: Distribution of sample points in the study area.

The Regions of Interest (ROIs) file was then exported to a CSV file, which was imported into WEKA.

2.6 Data Mining

Data mining techniques are currently in evidence because they are ways to recognize patterns and identify new information in large databases. There are several fields of application of data mining, such as medicine, marketing and hydrology. In the latter, the work of Banon *et al* (2013) is highlighted, which was the starting point for the development of the methodology proposed in this work.

The Waikato Environment for Knowledge Analysis (WEKA) application package (WITTEN & FRANK, 2005) provides several data mining algorithms, including decision tree classifiers, which predict the classes predefined by the operator from the values of the sample set attributes. These algorithms organize the classification rules in a tree structure, with nodes and leaves bound by branches. On each node there is a rule associated with a morphometric attribute, which can branch to other nodes with distinct rules and attributes until it reaches a leaf. The leaves represent the classes (perennial or ephemeral) that will be assigned to each mini-basin of the study area.

Tests were performed with the algorithms J48, Random Tree, Random Forest, RepTree, Hoeffding Tree, LMT And Decision Stump, and the first two obtained the best results.

The first tests performed obtained very complex decision trees and accuracy rates lower than 50%. Most of the incongruities were found in the plain, where the existence of springs is conditioned by the existence of depressions in the terrain. After the inclusion of the RCI attribute, the classification improved in all areas, because this attribute assists the algorithm in identifying relief patterns.

Five sample sets were tested, which underwent adjustments after each classification, until the result was satisfactory. The decision tree generated by J48 was used, with an accuracy rate of 70%.

The decision tree obtained has 7 nodes and 8 leaves, and uses only the attributes RCI, relative perimeter, hypsometric amplitude, area, medium slope and relative relief (Figure 4). Because it is small, with few nodes and leaves, the decision tree tends to be quite generalist, which means that its accuracy rate is good both in training data and in new areas (WITTEN *et al*, 2014).

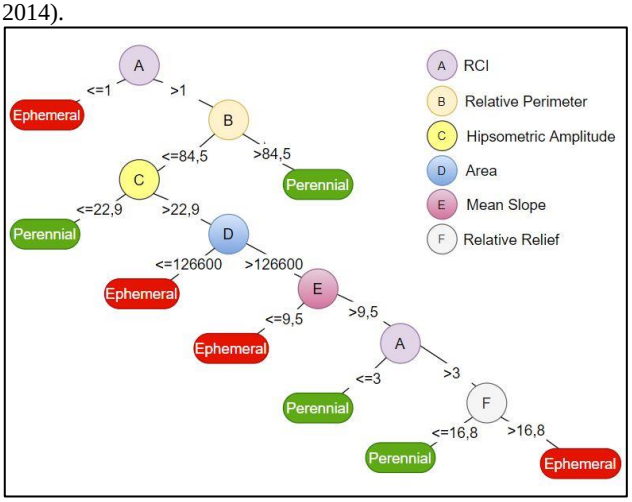

Figure 4: Decision tree generated by J48 from the sample data.

At the first node of the decision tree, the division rule is whether the minibasin are plain or not. If yes, they are considered as ephemeral, if no, the second rule is applied, which divides the remaining basins on the basis of their relative perimeter, whether it's greater or smaller than 84,5. If it's greater the basins are classified as perennial, if it's smaller, the third rule is applied to the remaining basins and so on.

2.7 Validation

WEKA offers some validation tools: stratified cross-validation with 10 partitions, percent split, and validation with a test dataset. Stratified Cross Validation was used, which divides the samples into 10 sets, of which 9 are used to train the classifier and one to

validate the classification. The process repeats until all 10 sets have been tested, and at the end WEKA provides the accuracy rate and the confusion matrix of the classes obtained in the result. After cross-validation, the classification was evaluated qualitatively by the specialist to verify classification errors, i.e., basins classified as perennials that do not have perennial stream; and basins classified as ephemeral that house a perennial watercourse. The samples mistakenly classified by the WEKA were repositioned in order to increase the accuracy of the model. In cases where the samples were not considered representative of the study area, sampling of ROIs was redone.

3. RESULTS

To run the model and classify the mini-basins as perennial or ephemeral, the decision tree built in the WEKA was transposed to the IDL/ENVI environment, which allows the raster classification to be performed. The result (Figure 5) is a raster with a binary classification of the mini-basins (perennial or ephemeral), in which only the first-order basins should be evaluated, and the remainder ignored.

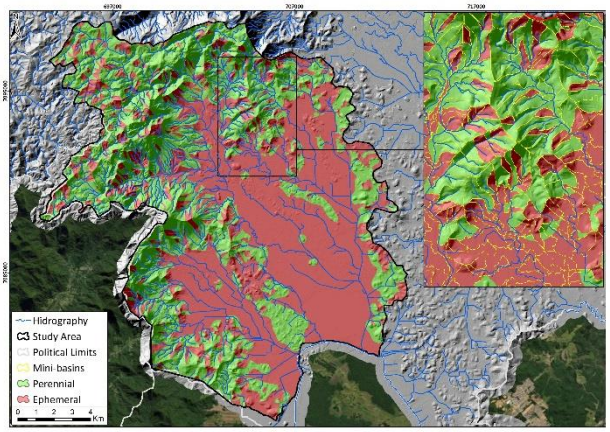

Figure 5: Result of the classification.

In general, the model yields better results for the mountainous and undulated areas rather than the plain areas. It happens because relief influences the hydrography strongly at the slopes, but at the plain the formation of springs is associated with local depressions or geological aspects, which are not included in the model. The model considers that all mini-basins at the plain are ephemeral, as established at the first node of the decision tree.

In quantitative terms, the accuracy obtained from the stratified cross-validation was 70%.

Out of the 775 first order mini-basins, 366 (47%) were considered as perennial and 409 (53%) ephemeral. Some basins didn't have a modelled stream in it, so they were considered ephemeral anyway.

Once the classification of the mini-basins in perennial or ephemeral was done, the drainage network extracted at the beginning of the process was refined using the results of the streamflow permanence model.

Previous to refinement, the gross hydrography was much denser than the refined hydrography.

Table 2 presents a quantitative analysis of total stream length before and after refinement of the hydrography modelled with the classic method.

Order	Gross (Km)	Refined (Km)	Reduction (%)
1	555	383	31
2	282	212	25
3	172	136	21
4	99	86	13
5	34	31	8
6	27	24	11
Total	**1169**	**872**	**25**

Table 2: Length of channels per Strahler order of streams before (gross) and after (refined) refinement, and overall reduction after refinement of the drainage network.

Ephemeral first order streams accounted for 31% of total channel length modelled, and overall, 25% of the channel length was classified as ephemeral. As Strahler order of streams increases, the channel length decreases, and reduction from gross to refined drainage length also decreases.

4. CONCLUSION

Modelling stream flow permanence for large areas is a challenge for researchers, due to the randomness in the occurrence of perennial, intermittent and ephemeral streams. Legal protection of natural streams promotes litigation between public and private sectors, and the development of stream flow models will be of great importance to the management of water resources hereafter. The method proposed in this work represents a new approach to modelling stream flow permanence in large areas, and obtained a satisfactory result using only DEM derived morphometric attributes of mini-basins.

5. ACKNOWLEDGEMENTS

This research is thankful to Celesc S.A for providing data, material and support for this research happening. We would also like to hank ANEEL, for providing funding for R&D projects within the electricity industry in Brazil.

6. REFERENCES

BRASIL. Lei Federal n° 12.651, de 25 de maio de 2012, alterada pela Lei 12.727, de 17 de outubro de 2012. Brasília, DF. Congresso Nacional, 2012.

CÂMARA, G.R.; SOUZA, C.M.; PEDROSA, B.; VINHAS, L.; MONTEIRO, A.M.V.; PAIVA; J.A.; CARVALHO, M.T.; GATTASS, M. TerraLib: Technology in Support of GIS Innovation. In: II Worhshop Brasileiro de Geoinformática, Centro Anhembi, São Paulo, 12 e 13 de junho de 2000, p. 126-133.

GUERRA, A.T. Dicionário Geológico-Geomorfológico. Rio de Janeiro: IBGE, 8ª ed. 1993.

GRAVELIUS, H. Flusskunde. Goschen Verlagshan dlung Berlin. In: Zavoianu I (ed) Morphometry of drainage basins. Elsevier, Amsterdam. 1914.

HANSEN, W.F. Identifying stream types and management implications. Forest Ecology and Management, 2001, ed. 143, p. 39-46.

JUNIOR, J.J.; ANDREOLI, C.V. Uso de dados climáticos e hidrológicos como subsídio na determinação do regime de fluxo de canais de drenagem. Nota Técnica. Revista Brasileira de Geomorfologia, v. 16, n°. 1, 2015. ISSN 2236-5664.

MILLER V.C. Quantitative geomorphic study of drainage basin characteristics in the Clinch Mountain area, Virginia and Tennessee. Technical report (Columbia University. Department of Geology), n. 3, 1953.

NC Division of Water Quality. Identification Methods for the Origins of Intermittent and Perennial streams, Version 3.1. North Carolina Department of Environment and Natural Resources, Division of Water Quality. Raleigh, NC. 2005. Disponível em https://files.nc.gov/ncdeq/Water%20Quality/Surface%20Water%20Protection/PDU/Headwater%20StrDEMs/NC%20Stream%20ID%20Manual_Ver%203.1.pdf. Acesso em 22/01/2020.

O'CALLAGHAN, J. F; MARK, D. M. The extraction of drainage networks from digital elevation data. Computer Vision, Graphics and Image Processing, v. 28, p. 323−344, 1984.

PANERO, F.S.; SILVA, H.E.B.; ROLIM, N.M. Aplicação de análise exploratória de dados na classificação de rios perenes e intermitentes do estado de Roraima. XLVI Congresso Brasileiro de Química, Salvador/BA, 25-29 de setembro de 2006. Disponível em: http://www.abq.org.br/cbq/2006/trabalhos2006/5/1001-1153-5-T1.htm. Acesso em 22/01/2019.

ROSIM, S.; MONTEIRO, A.M.V.; RENNÓ, C.D.; OLIVEIRA, J.R.F. Uma ferramenta open source que unifica representações de fluxo local para apoio à gestão de recursos hídricos no Brasil. Informática Pública, v. n. 1, p. 29-49, 2008.

SCHUMM S.A. Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey. Bulletin of the Geological Society of America, 67:597–646. 1956.

SECRETARIA DE ESTADO DO DESENVOLVIMENTO ECONÔMICO SUSTENTÁVEL DE SANTA CATARINA (SDS). Aerolevantamento do estado de Santa Catarina, 2011.

STRAHLER, A.N. Hypsometric (area-altitude) analysis of erosional topography. Geol. Soc. Am. Bull. 63, 1117–1142. 1952.

WENTWORTH C.K. A simplified method of determining the average slope of land surfaces. American journal of science 117:184-194. 1930.

WOHL, E. The significance of small streams. Frontiers in Earth Science, 2017, ed. 11, p.447-456.