A Proposed Rule-Discovery Scheme for Regionalisation of Rainfall-Runoff Characteristics in New South Wales, Australia

J.M. Spate[a,b], B.F.W Croke[b,c], J.P. Norton[a,b]

[a]*Mathematical Sciences Institute, The Australian National University, Canberra ACT 0200, Australia (jessica.spate@anu.edu.au)*

[b]*Integrated Catchment Assessment and Management Centre, The Australian National University, Canberra ACT 0200, Australia*

[c]*Centre for Resource and Environmental Studies, The Australian National University, Canberra ACT 0200, Australia*

Abstract: A project which sets out to explore alternative methods of regionalisation for prediction of streamflow in ungauged catchments is described. A commonly applied regionalisation methodology estimates model parameters for the ungauged basin from multiple linear regressions between parameters and catchment attributes, calibrated on gauged catchments. Its effectiveness is limited by the availability of sufficiently similar, well gauged and modelled catchments. In this paper a different approach is adopted, employing a data mining methodology to seek a useful set of rules in the catchment attribute space with parameter values as consequent. The outline of a hypothesis testing algorithm is given, along with a suggestion for a new discretisation method. The techniques are both based on the information theoretic concepts of cross entropy and mutual information.

Keywords: Regionalisation; Information Theoretic Measures; Hypothesis Testing.

1. INTRODUCTION

The problem considered here is prediction of streamflow in an ungauged catchment, through application of regionalisation rules relating streamflow characteristics to rainfall and measurable catchment attributes. Most commonly, regionalisation estimates rainfall-runoff model parameters for the ungauged basin from multiple linear regressions between parameters and topographical and land-use catchment attributes, calibrated on a large number of gauged catchments. Parameter values are used with rainfall estimated from adjacent or regional gauged data, adjusted based on a long-term mean rainfall surface [e.g. Schreider *et al*., 1997]. The results are typically not good, so an alternative, less prescriptive approach is being explored, based on data mining. As this study is in its early stages, only regionalisation of the rainfall-runoff coefficient *R* is considered (*cf*. Croke and Norton, [2004]). Knowledge of *R* for a given catchment is valuable for water management although by no means sufficient. However, regionalisation for a full set of rainfall-runoff model parameters is far more complex and introduces the additional issues of model-structure selection and parameter identifiability.

The study data are described in the next section. Section 3 is devoted to the transformations applied to them to obtain the information for regionalisation. Rule-extraction approaches are outlined in Section 4. Early results will be reported later.

2. DATA

Daily rainfall and streamflow series are available for 45 catchments in the South-East coastal region of New South Wales, Australia. Each catchment is also described by catchment area, percentage woody cover, relief, perimeter, channel density, and mean elevation above sea level. Climate records add two more descriptive features: average evapotranspiration (ET) and average yearly rainfall. The dataset covers a wide range of physical and geographical characteristics, from

small, mountainous basins to larger, low-relief coastal catchments.

Area, relief, slope, perimeter, channel density and elevation were obtained from a nine-arcsecond DEM [Hutchinson *et al.* 2000], which is well validated. Woody cover values from Lu *et al.* [2003] were used. These estimates were obtained data from multiple LANDSAT passes, with vegetation assumed to be herbaceous and not woody wherever significant yellowing occurred in summer. Point measurements were then used to calculate an estimate for the percentage woody cover over 5km by 5km cells. The cell measurements were then aggregated over each catchment.

Certain soil properties, such as saturated conductivity and soil type, were also available in some areas, but only as point measurements which often do not reflect the composition of the basin well. These data were therefore not included.

The rainfall and streamflow time series are of variable quality, and gauge density is not always as high as one would wish. The study area encompasses several built-up regions and one city (Bega) where gauge coverage is dense, but also National Park and State Forest catchments with little more infrastructure than fire trails and one or two rainfall measurement stations.

Rainfall values for each catchment were estimated by scaling gauged data by a mean surface derived using thin-plate smoothing splines [Hutchinson, 1998], interpolating and integrating over all relevant gauges to approximate the total rainfall incident on the catchment. The daily values were aggregated to a coarser time scale, reducing the effect of the errors incurred at this step.

Streamflow-measuring equipment is prone to occasional failure. The streamflow records ranged in length from less than five years to several decades, almost all peppered with missing days, weeks, or months. Accordingly, the standard procedure of selecting a time period common to all records was not practicable. Instead, the longest unbroken run of days was selected for each gauge. While this introduces some inconsistency, it was preferable to dubious interpolation to fill breaks in the record.

3. TRANSFORMATION OF TIME SERIES

The first action performed on the rainfall and streamflow series was conversion from daily to yearly records over the selected period for each gauge for input into calculation of the rainfall-runoff coefficient *R*. Although daily or monthly values would potentially give better insight into catchment behaviour and help provide information for short-term catchment management, they pose the need to consider catchment dynamics and are much more susceptible to error due to local climatic variability.

The rainfall-runoff coefficient *R* is a function of rainfall *P*, streamflow *Q* potential evapotranspiration *ET*, (all as mean annual values expressed as millimetres per year), woody cover *W*, mean slope m and other catchment attributes such as relief, soil type, drainage patterns, and typical rainfall intensity. Rainfall can be partitioned into runoff, evapotranspiration, the change δM in soil moisture and groundwater recharge *G* (which may be negative). Aggregating the time series to yearly figures permits the assumption that δM is small, *i.e.* that the net change in soil moisture over each year is a negligible component of the water balance. In the absence of better information, *G* is also assumed to be negligible. With these assumptions, Equation 1 may be applied, and hence Equation 2 obtained.

$$P = Q + ET + \delta M + G \cong Q + ET \quad (1)$$

$$\text{so} \quad R \cong 1 - ET/P \quad (2)$$

Evapotranspiration can be expressed as the product of a function ET_{lz} of woody cover and rainfall [Zhang *et al.*, 2001], and a function *g* of topographical and other catchment characteristics:

$$ET = g(PE, m, \ldots) * ET_{lz}(W, P) \quad (3)$$

ET_{lz} was calculated and added to the list of descriptive features.

To summarise, the preprocessed data consist of 45 records, each consisting of annual values for *R*, and a descriptive vector with eight spatially aggregated variables: slope, percentage woody cover, relief, elevation, channel density, perimeter, catchment area, and ET_{lz} as features.

4. HYPOTHESIS-TESTING ALGORITHM

Rather than adopting the conventional linear regression method of relating hydrological parameter values to measurable attributes of ungauged catchments, a more flexible, rule-based approach is taken. The rules are to be discovered from the data, without strong prior assumptions about model structure. There are, of course, limitations to the rule syntax, and the complexity of the rule set must be subject to conditions of demonstrable predictive power, just as a parametric model should be parsimonious in its number of parameters.

Each rule is composed of a simple set of conditions. An example of a possible rule is

$$\text{Slope} > 0.3 \text{ AND } 800\text{m} < \text{Elevation} \leq 950\text{m} \quad \text{AND Area} < 5\text{km}^2 \text{ THEN R} > 0.5 \tag{4}$$

Individual conditions, eg *Slope > 0.3*, or groupings of conditions are called predicates. Note that not all variables need to be represented in a given rule, although for the sake of computational simplicity, trivial conditions stating that the value of a variable not forming part of a rule is between the maximum and minimum bounds of that quantity can be added.

The space from which rules must be extracted is potentially large, the number of discrete values growing exponentially with the number N of descriptive attributes. While at present N is only 7, further catchment attributes may be added later. Even with seven attributes, the number of discrete values, and hence possible rules, is $n_1 n_2 n_3 n_4 n_5 n_6 n_7$ where n_i is the number of values attribute i may take, *e.g.* 10^7 if each n_i is 10. Clearly an exhaustive search for each possible rule in the space is impractical. There exists a large number of widely used rule-extraction algorithms. Perhaps the best known is the Apriori algorithm [Agrawal and Srikant, 1994] developed for supermarket basket analysis. The method isolates small 'frequent itemsets', or combinations of attribute values that occur together often, and uses them to build large frequent itemsets that can be considered rules describing features in the data. Terabyte-sized databases can be efficiently processed in this way.

Like most rule-extraction methods, the Apriori algorithm usually yields a large number of rules, most of which are irrelevant or intuitively obvious. Sorting and ranking the rules for interestingness and relevance is a problem in itself. See Freitas, [1999] for a discussion. Also, the basic Apriori algorithm produces rules with unspecified consequents (explained items), whereas only those that say something about R are of interest here. Simple interdependence among the small number of descriptive data (catchment characteristics) can be investigated by simpler methods such as visualisation.

Under data mining problem, regionalisation could fall as well under the heading of classification as under rule extraction, because the target class is specified at the outset. However, it is not necessary here to completely cover every instance in the dataset, in contrast to normal classification. The algorithm described below will borrow from general-to-specific classification by rule methods like the AQ family of algorithms (see Wnek and Michalski, [1991]). A hypothesis-test-driven route will be taken rather than the more usual divide-and-conquer approach to classification.

From physical considerations, certain tentative rules giving the rainfall-runoff coefficient R can be postulated. For example, catchments with small area, high elevation, and high slope are likely to respond to a rainfall event with a sudden peak in streamflow that accounts for most of the incident rainfall. Thus, it seems probable that R be high for these catchments. This intuitive rule is quantified in Equation 4.

Clearly, there exists a large gulf between the rough description of the trend and the sharply defined rule. The domain expert, in this case a catchment management scientist familiar with the area in question, can make an educated guess as to what the rule should look like and estimate the bounds on each condition.

The proposed rules are unlikely to be optimal, but form an initial candidate set from which to move. It is here that the algorithms described below come into play. Each candidate rule must be tested against two aspects of rule performance, accuracy and support. Accuracy, often called confidence in the literature, is simply the percentage of times the consequent occurs when the rule's conditions are fulfilled, and support is the number of times the rule is applied in the training dataset, normalised by the number of records N. If support is too low, the rule is not interesting because it rarely applies, and if accuracy is insufficient, the rule does not consistently represent a trend.

From a simple hypothesis-testing point of view the expert-generated candidate rules could be tested for support and confidence against predefined thresholds. Rules meeting both criteria would be accepted and those with poorer performance rejected. However, as noted above, these rules are unlikely to be optimal. The quantitative relations presented by the domain expert are likely to translate imperfectly into rules with valid syntax. Therefore, a local search will be performed around each rule, tightening and generalising each condition (or predicate) until a rule that identifies quantitatively and precisely the trait suggested by the domain expert is arrived at. A possible execution scheme, sketched in pseudocode, is:

1. Select a candidate rule
2. Test confidence and support
3. Select a new condition. If none, go to 1

3.1 Relax condition by factor f_1

3.2 Test confidence and support of predicate

3.3 If support is still acceptable and confidence has increased by *d* or more, go to 3.1

4. If confidence has increased over initial value, go to 5 else continue

4.1 Tighten bounds by factor f_2

4.2 Test confidence and support

4.3 If confidence has increased by *d* or more, go to 4.1 else go to 5

5. Store new condition, go to 3.

The question of precisely how individual conditions should be varied should be considered in detail. The order in which the conditions should be taken, and the parameters of the relaxation or tightening of each condition, must be chosen. Here the method diverges from the most of the usual approaches, which involve random, greedy or exhaustive searches through the space of possibilities. Two different methods for evolving the candidate rules will be explored at this step.

The first and simplest is to rate each condition by its (normalised) support, relaxing those conditions with low support and further constraining those with high support. This way, it is not unreasonable to expect to find a balance between support and confidence, but such ideal behaviour is by no means guaranteed. It may well be that greater performance can be achieved by loosening a condition with high support, or conversely by tightening one with low support. It may be optimal to shift the upper or lower bound only, but this possibility will be ignored for the moment.

When using this method, the rather bold assumption is made that the optimal rule is most likely to consist of a number of predicates (individual conditions) with close to the same mean level of support, and that a simple tradeoff relationship of the type illustrated in Figure 1 exists for each.

Figure 1. Confidence and Support as assumed by simple method

Of course, this is not in general true, nor are the contributions of each predicate usually separable as has been implicitly assumed thus far, and will be assumed below.

In the second method, a more sophisticated approach is taken, using the information-theoretic properties of the data to guide adjustment of conditions. The most important measures used are the entropy *H(X)* of a variable *X*, and cross entropy *H(R,*X) between a descriptive feature *X* and the target *R*. Entropy and cross entropy are defined [Henery, 1994] as

$$H(X) = -\sum_{ij} \pi_i \log_2 \pi_i \quad (5)$$

$$H(R, X) = -\sum_{ij} p_{ij} \log_2 p_{ij} \quad (6)$$

Here, π_i is the prior probability that *X* takes its *i*th value and p_{ij} is the joint probability of the target taking its *i*th value while the descriptive feature takes value *j*. If the dataset is very large, these values could be estimated from a subsample rather than the entire dataset. For our 45 samples reducing the sample size is not necessary or desirable, but the hypothesis-testing method outlined is transferable to other problems.

Entropy tells us about the distribution of instances in *R*, and cross entropy provides an intuitively simple measure of the information *X* is capable of providing about *R*. The presence of the logs to base 2 stems from the use of information theory to quantify the bits of information required to determine an outcome uniquely. Intuitively, the entropy as a number of bits is also the number of binary yes/no or greater than/less than questions needed to identify a result. MacKay (2003) uses the example of simple number guessing game and demonstrates that *H(X)* is the minimum number of questions required to correctly identify any number in *X*.

The mutual information *M(R,X)* [Henery, 1994] between feature *X* and *R* can be calculated as Equation 7. Mutual information gives a useful measure of how dependent the feature and classifier are. If *R* and *X* are completely independent, *M(R,X)* is zero and *X* alone provides no information about *R*. If *M(R,X)* is maximal (equal to the lesser of *H(R)* and *H(X)*), then the value of *X* determines the value of *R* completely [Hamming, 1994].

$$M(R, X) = H(R) + H(X) - H(R, X) \quad (7)$$

Given a particular rule hypothesis, the predicates for each feature X^i will first be ranked according to $M(R; X^i)$. If for any i this value is close to zero, it may be decided that conditions based on X^i will be omitted. Alternatively, they may be allowed on the grounds that each predicate relates to only a small part of X^i. This question will be considered more in subsequent work. The predicate operating on the feature with greatest mutual information with R will be operated upon first.

Next, individual predicates must be expanded, contracted or shifted. At this stage many quantities, most importantly the p_{ij} calculated in the construction of $H(R,X)$, are known. These values can be used to decide how to shift the boundary of the predicate. Consider the condition C, part of some hypothetical rule:

$$C = X \subset X_k, X_{k+1}, \ldots, X_m \Rightarrow R = R_j \quad (8)$$

Note that this statement does not require X to be a numerical or ordered feature, although ordering is useful in the boundary-shifting step. If ordering is not present and a distance metric cannot be defined by the domain expert, some kind of search through candidate categories must be performed, increasing the computational complexity of the calculation.

For each X, a p_{ij} has already been calculated. Therefore, we can readily test p_{kj} and p_{mj} to ascertain which way each boundary should be shifted: on the left, if p_{kj} is small, X_k can be removed from the condition, and if it is large and p_{k+1j} is also, X_{kj1} could tentatively be added. Parameters to govern the various thresholds on p_{ij} need to be defined, but as this is a qualitative discussion we shall leave this process to the experimental stage.

In advocating this method, the assumption has been made that the domain expert has produced qualitatively good rules. This may not be the case, and, if so, varying the predicate boundaries is unlikely to remedy the situation. The tests for confidence and support must be sufficiently strict to remove rules that are irrelevant when compared to some benchmark values. The expected confidence and support of a few randomly generated rules would be a sound choice to provide a base level of performance.

Locally and temporarily, the process of adding and removing blocks to a condition amounts to adjusting the discretisation of the feature (predicate granularity may vary between rules covering the same attribute). This observation suggests an alternative, to define the discretisation at the data-processing stage, where any floating-point features need to be discretised anyway.

Many methods exist for discretising continuous data, ranging from simple binning to quite sophisticated distribution-based methods. For a small number of instances as in this project, the more complex methods are not really useful. A simple entropy-based discretisation method is:

1. Sort entries by feature value
2. Divide dataset in half; call the subsets record A and record B
3. Consider record B one bin, and define two small bins S and T at the left-hand extreme of record A
4. Calculate $H(B \cup S, X)$ and $H(B \cup S \cup T, X)$ by Equation 1
5. If adding T to the union decreases the cross-entropy, aggregate S and T into S'. If not, go to 8.
6. Select a small partition to the right of T and call it T'
7. Repeat from 4. with T' and S' in the places of T and S, until the condition fails or the rightmost edge of record A is reached, in which case go to 10
8. Leave the small block S as it is and redefine T as S'. Select a small block immediately to the right of T to be labelled T'
9. Repeat from 4
10. Swap records A and B. When the end of dataset B is reached, stop.

Note the minus sign in the definition of cross-entropy (Equation 6). Although simplistic and biased towards defining partition edges further to the right than may be optimal, this scheme provides a rough way of decreasing (although perhaps not minimising) the cross-entropy of a feature with the target during the discretisation stage, and thus increasing the mutual information between the two.

While repeated calculations of cross-entropy are computationally expensive, the summation is separable into individual parts, and these can be stored at first calculation and re-used. For example, the contributions of records A and B need only be calculated once each. The cost of the above algorithm would probably be governed by memory accesses rather than floating-point calculations.

Once discretisation has been performed in the data processing stage, converting the rules to suit the new format is trivial, and the condition-size adjustments detailed above would be superseded. Of course, the discretisation does not take into account specific interesting values of the target R

as the process based around p_{ij} does, and some unnecessary work may have been done in regions away from the hypothesised predicates, but the two methods are not claimed to be equivalent, merely two ways of approaching the problem. Also, the unnecessary adjustment of bin edges in regions thought to be non-informative by the domain expert may highlight non-intuitive, interesting aspects of the system.

The information theoretic approach we have taken has parallels with the noisy-channel coding theory developed by Shannon (see MacKay [2003]) and others for information transmission, compression, and storage. It is hoped that the rigorous foundation of that work will be transferable to this class of problems, and thus that theoretical performance bounds can be calculated for the algorithms described above.

Preliminary experiments with a standard dataset taken from the University of California at Irvine machine learning repository indicate that both the discretisation routine and information-theoretic rule evolution algorithm are in principle valid. However, the UCI datasets are well understood and much-studied, whereas environmental data of the kind we aim to use for regionalisation are complex-full of noise and interdependencies. It remains to be seen if the simple procedures described above can extract useful rules from such a system.

5. FURTHER WORK AND CONCLUDING REMARKS

A brief look has been taken at one approach to the regionalisation problem as posed by Croke and Norton [2004], and an algorithm suggested to develop vague hypothesised rules into firm quantitative rules. It is thought that the output rules will optimally capture the features suggested by the domain expert. The natural next step in the process is to apply the rule-extraction algorithm with and without information-theoretic guidance, to validate the method outlined in Section 4.

If the validation is positive, the next task will be to look into extending the regionalisation (and accordingly the algorithm) into model parameter space. Also, the method proposed is in no way confined to regionalisation problems, so trials with other datasets with different characteristics would be simple in principle.

The project including this work is at too early a stage for results to be reported in this paper, and performance comparison with any of the plethora of existing rule-extraction methods is not yet possible, but it is hoped that the skeleton of a viable and novel regionalisation method has been presented.

6. REFERENCES

Agrawal, R. and Srikant, R., Fast algorithms for mining association rules. *VLDB 1994: Proceedings of the 20th international conference on Very Large DataBases*, Santiago de Chile, Chile, 487–499, 1994.

Blake, C.L. and Merz, C.J., UCI Repository of Machince Learning Databases, University of California, Irvine, Department of Computer and Information Science, 1998.

Croke, B.F.W. and Norton, J.P., Regionalisation of rainfall-runoff models, *iEMSs2004*, 2004.

Freitas, A.A., On rule interestingness measures. *Knowledge-Based Systems*, 12:309–315, 1999.

Hamming, R.W., Coding and Information Theory, pp 103-118. Prentice-Hall, 1986.

Henery, R.J., In Michie, D., Spiegelhalter, D.J, and Taylor, C.C. editors, *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, Crystal City, USA, 1994.

Hutchinson, M.F., Interpolation of rainfall data with thin plate smoothing splines: II. Analysis of topographic dependence. *Journal of Geographic Information and Decision Analysis*, 2, 168-185, 1998.

Hutchinson, M., Stein, J. and Stein, J., Upgrade of the 9 second Australian Digital Elevation Model, Australian National University, 2000.

Lu H., Raupach M.R., McVicar T.R., Barrett D.J., Decomposition of vegetation cover into woody and herbaceous components using AVHRR NDVI time series, *Remote Sensing Of Environment* 86 (1): 1-18, 2003.

MacKay, D.J., Information Theory, Inference, and Learning Algorithms, pp 68-72. Cambridge University Press, 2003.

Schreider, S. Yu, Whetton, P.H., Jakeman, A.J. and Pittock, A.B., Runoff modelling for snow-affected catchments in the Australian alpine region, eastern Victoria, *Journal of Hydrology*, 200,1-23, 1997.

Wnek, J. and Michalski, R.S. Hypothesis-driven constructive induction in AQ17: A method and experiments. *Proceedings of the IJCAI-91 Workshop on Evaluating and Changing Representation in Machine Learning*, Sydney, Australia, 13–22, 1991.

Zhang, L., Dawes, W.R. and Walker, G.R. Response of mean annual evapotranspiration to vegetation changes at the catchment scale, *Water Resources Research*, 37, 701-708, 2001.