Identification of Long Term Patterns in Hydrologic Data Using Fuzzy and Neural Network Techniques

N. Lauzon and **B.J. Lence**

Department of Civil Engineering, The University of British Columbia, 2324 Main Mall, Vancouver, BC V6T 1Z4, Canada (lauzon@civil.ubc.ca)

Abstract: In the 1990s, several studies have been undertaken to determine whether the hydrologic regime in Canada and the United States has changed over time. The purpose has been to detect the presence of shifts or trends in streamflow sequences through the use of common univariate statistical tests. This paper shows that the Kohonen neural network and fuzzy c-means approaches can be used for the detection of shifts and trends in data sequences. A performance analysis based on synthetic data samples, with properties that are similar to those found in natural streamflow sequences recorded in Canada, is accomplished to analyze the value of the conventional statistical tests, the Kohonen network and the fuzzy c-means approach for the detection of shifts and trends. The results show that all techniques perform similarly. All techniques perform very poorly when the coefficient of variation of the data set is very high or when the shift or the trend is very small. However, unlike the statistical tests, improvement of the performance is likely with the Kohonen network and fuzzy c-means using multiple types of inputs.

Keywords: Shifts; Trends; Kohonen network; Fuzzy c-means; Inflows

1. INTRODUCTION

Anthropogenic and natural changes can influence the water inflow regime produced on watersheds over time. As a water resource system analyst, it is important to be aware of such long-term influences or patterns in order to estimate water availability as accurately as possible. The common practice is to divide these long-term patterns into two categories: (1) shifts, which are sudden changes over time in the statistical properties of the inflow sequences, and (2) trends, which represent continuous variations of the statistical properties over time. Conventional statistical tests, such as the Mann-Whitney, Mann-Kendall and Spearman tests, have been used in several studies in order to determine the presence of shifts or trends in hydrometric data sequences from measurement networks in Canada [Yulianti and Burn, 1998, Zhang et al., 2001] and the United States [Lettenmaier et al., 1994; Lins and Slack, 1999].

With shifts, the task is to identify the various clusters present in the data sequence. Each cluster represents a subset of continuous data between two shifts in the sequence, and the individuals in each subset are assumed to come from the same population. With trends, the goal can also be to separate the data sequence into clusters even though the statistical properties of the individuals in the sequence constantly evolve. Indeed, subsets of continuous data can be formed so that the variability of each subset is smaller than the variability of the whole sequence, and this is the basic objective of any clustering technique in trend analysis. Among the various clustering techniques, the Kohonen neural network and fuzzy c-means approaches, originating from the development of artificial intelligence techniques, are among the most recently developed ones and have been applied for the definition of homogeneous hydrologic regions [Hall and Minns, 1999]. This paper develops these techniques for identifying shifts and trends in hydrometric data sequences and compares their performance with that of conventional statistical tests.

In this paper, statistical tests and artificial intelligence techniques for the detection of shifts and trends are described first. Then the performance of these techniques are evaluated using synthetic data samples, with properties that vary within the range of those from historical hydrometric records in Canada.

2. CONVENTIONAL TESTS

A brief description of each test is presented in this paper. The Student's test on the mean is a standard test of hypotheses that can be found in any statistical textbook. The other three tests are less common, but detailed descriptions can be found in Conover [1980] or Salas [1993].

2. 1 Tests for Shifts

Either the Student's or the Mann-Whitney tests can be used for detecting shifts. Given a sample of individuals y_t, t = 1, ..., N, divided into two continuous sub-samples of size n_1 and n_2 ($n_1+n_2 = N$), both tests assume a null hypothesis stating that both sub-samples come from the same population. With the Student's test, the null hypothesis is rejected if the absolute value of the standardized difference of the means of the sub-samples is greater than $T_{1-\alpha/2,v}$, which is the 1-α/2 quantile of the Student's distribution, with $v = N$-2 degrees of freedom and α as the significance level of the test. With the Mann-Whitney test, the whole sample is sorted in ascending order, and the mean (R) of the position in the sorted sequence of the individuals of the first sub-sample is calculated. This value of R is standardized, and the null hypothesis is rejected if this absolute standardized value is greater than $u_{1-\alpha/2}$, which is the 1-α/2 quantile of the standard normal distribution, with α as the significance level of the test.

For both tests, the location of the shift and consequently the size of n_1 and n_2 are assumed to be known. When they are not, the solution is to apply the test at all potential shift locations, and to assume that the shift actually occurs at the location of the largest test value above the value of $T_{1-\alpha/2,v}$ or $u_{1-\alpha/2}$, respectively.

2. 2 Tests for Trends

Either the Mann-Kendall or the Spearman tests can be used for detecting trends. Given a sample of individuals y_t, t = 1, ..., N, both tests assume a null hypothesis stating that there is no trend in the sample. The Mann-Kendall test considers the gradient between each individual y_t, t = 1, ..., N-1, and all the subsequent individuals $y_{t'}$, $t' = t$+1, ..., N, in the sample. The value of the sum of the number of positive gradients minus the sum of the number of negative gradients is calculated, and standardized. The null hypothesis is rejected if the absolute standardized value is greater than $u_{1-\alpha/2}$, which is the 1-α/2 quantile of the standard normal distribution, with α as the significance level of the test. With the Spearman test, the sum of the difference between the actual position of each individual in the sample, and its position in the sample when it is sorted in ascending order, is performed. This sum is standardized, and the null hypothesis is rejected if the absolute standardized value is greater than $\rho_{1-\alpha/2}$, which is the 1-α/2 quantile of the probability distribution related to the Spearman test, with α as the significance level of the test. The distribution can be found in Conover [1980].

The results of the Mann-Kendall test can be affected if the sample exhibits a significant auto-correlation. The solution is therefore to perform the test on a pre-whitened sample. Assuming that ρ is the lag-1 auto-correlation, then the pre-whitened sample is y_2-ρy_1, ..., y_n-ρy_{n-1}. It is suggested that one should pre-whiten the sample if $\rho > 0.1$ [Zhang et al., 2001].

3. ARTIFICIAL INTELLIGENCE TECHNIQUES

3. 1 Kohonen Network

A comprehensive description of the Kohonen network can be found in Kohonen [1990]. Such a network is made of an input layer that receives the data and of an output layer composed of several neurons often structured in a two dimensional plane. The weight vector (w) of each neuron in the output layer is calibrated through an unsupervised learning process to respond to specific input patterns. In the learning process the weight vectors are updated as follows:

$$w_{i,l+1} = \begin{cases} w_{i,l} + \beta_l \left(x_l - w_{i,l}\right) & \text{if } i \in N_l \\ w_{i,l} & \text{if } i \notin N_l \end{cases} \quad (1)$$

where i is the subscript identifying the neurons in the output layer, l is the calibration step, and x_l is the input vector coming from the input layer. Element N_l represents the neighborhood set, which is centered at the winning neuron at step l in the calibration process and defines all the neurons whose weights must be adjusted. Parameter β_l is the learning rate, which indicates how much the weights adapt to the input. The weights are on the same scale as the input vector, which means that each neuron of the output layer can be considered as a cluster whose centroid is defined by its associated weight vector. The calibration process also tends to structure the output layer so that the input pattern can be defined in some meaningful coordinate system [Kohonen, 1990], which is why the Kohonen network is also called a self-organizing map.

As explained in Section 3.3, the application in this paper takes advantage of the mapping properties of the Kohonen network. From the weights attributed to the neurons on the output map, it is possible to differentiate individuals of a data sample if they come from different populations.

3. 2 Fuzzy c-means

A detailed description of the fuzzy c-means approach can be found in Ross [1995]. It is an unsupervised clustering technique which, during the calibration process, allows the input vectors to be attributed to several if not all clusters as defined by a degree of membership. Indeed, given K input vectors and C possible clusters, an input vector x_k is associated with a cluster c with some a membership value μ_{ck}. The objective is to minimize the function:

$$F = \sum_{k=1}^{K} \sum_{c=1}^{C} (\mu_{ck})^r (d_{ck})^2 \qquad (2)$$

where d_{ck} is the distance between input vector k and cluster center c, and r is a weighting parameter controlling the amount of fuzziness in the process of classification. The value of r generally varies between 1.25 and 2. The values of the cluster centers are functions of the sum of the input vectors weighted by their respective membership values, and the membership values are updated with respects to the distances (d_{ck}) during the process of minimizing function F.

If the optimal clusters are ordered by the value of the cluster center, then both artificial intelligence techniques are identical. The output neuron in the Kohonen network is similar to a cluster in fuzzy c-means, and vice-versa. Thus, for both approaches similar procedures can be employed to detect shifts and trends, as detailed in Section 3.3.

3. 3 Applicability of the Kohonen Network and Fuzzy c-means

The purpose of clustering techniques is to classify the individuals of a data sample with respect to some given features. Consider a data sample made of individuals coming from two different populations as would be the case if there were a shift. If this data sample were used to calibrate a Kohonen network or fuzzy c-means clusters, it would be expected from the calibration process that the individuals from the first population activate a particular region of neurons on the map or specific clusters, while the individuals from the other population activate other neurons or clusters. On the Kohonen network map, the centroids of the regions affected by each population can be calculated the same way centroids can be determined on a topographic map, and the distance between centroids thus becomes an indicator of the magnitude of the shift. Similarly, if the cluster in fuzzy c-means can be ordered to form a map, then centroids and distances can also be determined. Employed as such, the Kohonen network and fuzzy c-means approach perform exactly the same function as the Student's and Mann-Whitney tests whose goals are also to provide a measure of the distance that differentiates two populations.

For all techniques, tests and artificial intelligence techniques, if the location of the shift in the data set is not known, the strategy is to verify all potential locations, and the one that produces the largest distance is assumed to be the actual location of the shift. Also, it must be noted that if there is no shift at all in the data set, then all techniques should produce distances equal to zero at all potential locations.

With trends, it is assumed that features in the data sample constantly vary over time. This is the equivalent of having shifts at all possible locations on the data sample. Under this circumstance, all techniques employed for the detection of shifts would provide distances that are greater than zero for all potential locations. One may also judge the presence of trends with the Kohonen network and fuzzy c-means by verifying how the individuals of the data sample are grouped in each neuron and cluster. Indeed, if either technique were calibrated for data that fall on a curve, then the chosen weights would assign each neuron or cluster to only a specific part of that curve. Similarly, with a data set that represents the record of a trend over time, each neuron or cluster should be activated by the data coming from a specific period of time. The Mann-Kendall and Spearman tests evaluate how each individual in the data set ranks with respect to the other individuals, which is a task equivalent to checking how the individuals are grouped within neurons and clusters in the Kohonen network and fuzzy c-means, respectively.

A minor disadvantage of the use of artificial intelligence techniques for the detection of shifts or trends is that there is no decision criterion such as those with the conventional statistical tests to say, with some degree of confidence, whether or not there is a shift or a trend. However such criteria are only indicative, and the results in this paper show that such decision criteria are somewhat unreliable for either the detection of shifts or trends or the detection of no-shift or no-trend.

The advantage of these artificial intelligence techniques is that, unlike the statistical tests presented here, multivariate input cases can be addressed. Indeed, only one type of inputs, such as a sequence of annual streamflow peaks or of annual streamflow means, can be evaluated with the statistical tests. With the Kohonen network and fuzzy c-means, more than one type of input can be used. For example, both flood peaks and flood volumes are normally needed in flood routing analysis leading to the design of spillway gates on a dam. Normally, a hydrograph with a representative shape, as defined by its peak and volume, is chosen from the streamflow records and used in the routing analysis. Of course, shifts and trends can affect the shape of the hydrographs in the records, and could be noticed only if both the peaks and volumes are considered together. The domains of hydrology and water resources include many other multivariate input cases that could be handled only with multivariate clustering techniques such as those presented here to detect of shifts and trends.

4. APPLICATION CONTEXT

Monte-Carlo simulations are used to generate synthetic data to determine the value of all statistical tests and artificial intelligence techniques. The behavior of each method is evaluated with respect to (1) the length of the data samples, (2) the mean and the coefficient of variation of the data samples, and (3) the amplitude of the shifts or trends imposed on the data samples. These characteristics are set to represent those of natural streamflow records found in Canada. More precisely, these characteristics can represent annual streamflow peak and mean sequences, or seasonal streamflow peak and mean sequences.

A total of 60,000 univariate data samples are created randomly, half of them corrupted with shifts (one per set) and the other half corrupted with trends, either continuously upward or downward for the whole length of each sample. For both shifts and trends, there is an equal number of 20-, 50- or 100-individual data samples, while the mean and the coefficient of variation vary equally between 1 and 20,000 and between 0.05 and 0.3, respectively. With data samples created for shifts, the amplitude of the shift is equally either 0%, 1%, 3%, 5%, 10%, 15% or 30%, while the location of the shift can be anywhere on the set except within the first and last five individuals. With data samples created for trends, the trend is equally either downward or upward and its amplitude is equally either 0%/yr, 0.02%/yr, 0.06%/yr, 0.1%/yr, 0.2%/yr, 0.3%/yr, or 0.5%/yr.

As indicated in Section 2.2, the Mann-Kendall test is used on pre-whitened samples if the lag-one auto-correlation value of the sample is higher than 0.1. The number of output neurons in the Kohonen network and the number of clusters in the fuzzy c-means approach must be set in advance, and of course these numbers cannot exceed the number of individuals of the samples. A total of 4, 10 and 20 neurons or clusters are employed for the 20-, 50- and 100-individual samples, respectively. The neurons are structured in line on the output layer, while the clusters are ordered with respect to the value of their cluster center.

5. RESULTS

5. 1 Validity of Statistical Tests

Figures 1 and 2 illustrate the relatively poor efficiency of the conventional statistical tests with the types of data used here. The decision criteria for all of these test results are based on a significance level of 10%. Figure 1 shows the success rate of the criterion for the Mann-Whitney test to detect a shift when there is actually one (M_S in the figure) and to detect no-shift when there is actually none (M_NS), as well as the success rate of the criterion for the Student's test to detect a shift (S_S) or no-shift (S_NS). Figure 2 presents similar results for trends, that is, the Mann-Kendall test detecting an actual trend (M_T) and no-trend (M_NT), and the Spearman test detecting an actual trend (S_T) and no-trend (S_NT).

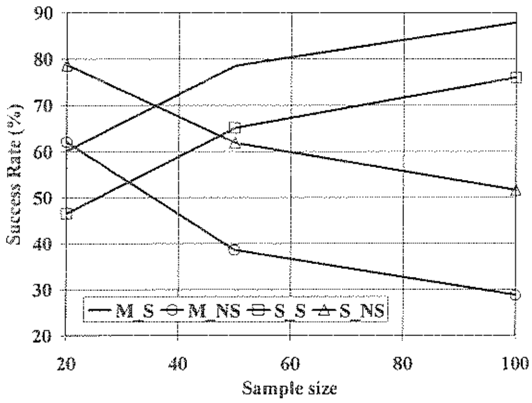

Figure 1. Success rates of the detection of shifts and the detection of no-shift with the Mann-Whitney and Student's tests.

As a whole, these success rates can be considered low in most cases. Only the success rates for the detection of no-trend by the Mann-Kendall and Spearman tests are above 90% at all times. This demonstrates that the use of decision criteria for the statistical tests for the detection of shifts or trends and the detection of no-shift or no-trend is somewhat unreliable.

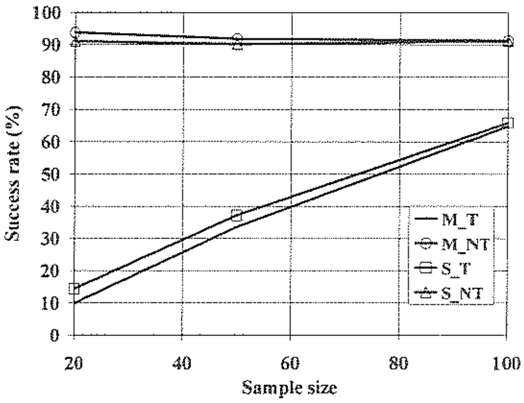

Figure 2. Success rates of the detection of trends and the detection of no-trend with the Mann-Whitney and Student's tests.

5. 2 Analysis of Shifts

With all techniques, it is assumed that the location, among all potential sites, for which the techniques produce the greatest distance between centroids is considered as the actual location of the shift. Based on that assumption, all techniques behave similarly, and detect the location of the shift exactly only 20%, 16% and 14% of the time for the 20-, 50- and 100-individual samples, respectively. If one is content to know the location of the shift within ±5 time-steps, then the success rates increase to 85%, 50% and 40%, respectively. The success rate decreases as the sample size increases, because the number of potential locations of shifts increases with the sample size, thus increasing the potential to choose an erroneous location.

All techniques are insensitive to the mean of the sample. This is because they all consider the differences between values of the individuals in the sample so that the size of the mean has no effect. The performance of all techniques decreases as the coefficient of variation increases. That is, it is easier to detect a shift when the individuals in the sample are relatively close to a mean value than if the individuals are not. As expected, the performance of all techniques increases as the amplitude of the shift increases.

The entire set of data samples can be divided into four subsets based on the results of the statistical tests: the subset of samples corrupted with shifts and identified by the statistical tests as corrupted (C:C), the subset of corrupted samples identified as uncorrupted by the tests (C:NC), the subset of uncorrupted samples identified as uncorrupted (NC:NC), and the subset of uncorrupted samples and identified as corrupted (NC:C). In this study the properties of the distances between centroids obtained with the artificial intelligence techniques are examined graphically for each of these data subsets. Figure 3 is a typical graph used of these results. It shows the mean of the maximum distance for each subset with respect to the mean of the sample, as derived from the Kohonen network for samples with 100 individuals

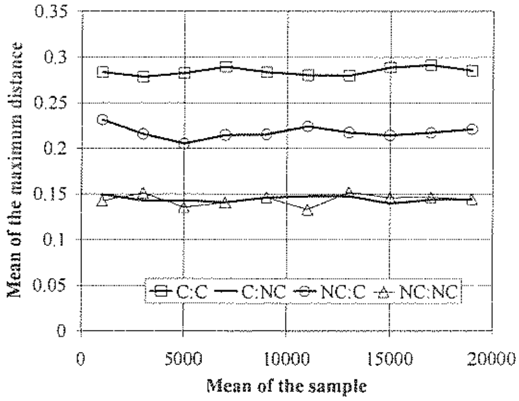

Figure 3. Mean of the maximum distance with respect to the mean of the sample, as derived from the Kohonen network with 100-individual samples.

This example shows that subsets C:NC and NC:NC have identical mean maximum distance, and this means that the Kohonen network cannot correctly diagnose shifts for samples of subset C:NC because the distance properties of this subset are similar to those of subset NC:NC. The mean maximum distance for subset NC:C is different from that of any other subset. This means that, unlike the statistical tests, the Kohonen network can potentially diagnose no-shift for samples of subset NC:C. The accuracy of this diagnosis is of course sensitive to the standard deviation of the maximum distance for each subset.

Graphs such as Figure 3 have been produced to compare the distance properties (maximum values, means and standard deviations) obtained with the artificial intelligence techniques with respect to the mean, coefficient of variation and size of the samples, as well as the amplitude of the shift. The conclusions are that: (1) subset C:C has distinct properties that clearly differentiate it from the other subsets; (2) like the statistical tests, the artificial intelligence techniques cannot distinguish subsets NC:NC and C:NC; (3) unlike the statistical tests, the properties of subset C:NC are distinct from those of the other subsets, making it possible to correctly diagnose a no-shift situation.

5. 3 Analysis of Trends

In brief, all techniques behave similarly. They are all insensitive to the mean of the sample because they all consider the differences between values in the sample and therefore the mean does not matter. The success rate decreases as the coefficient of variation increases and it increases as the amplitude of the trend increases. The success rate increases as the sample size increases because the trend is defined by a rate of increase or decrease of the sample mean per year, which means that the overall trend is more significant when the sample size is large and therefore is easier to detect.

With the artificial intelligence techniques, the properties of the distances have also been analyzed graphically. In this analysis, a case is corrupted if a trend is present in the data sample. The conclusions are identical to those provided above for shifts, that is: (1) the properties of subset C:C are clearly different from those of the other subsets; (2) subsets C:NC and NC:NC have similar properties and therefore cannot be differentiated; (3) subset NC:C have distinct properties, making it possible to correctly diagnose a no-trend situation.

In this application, it is known *a priori* which samples are corrupted with a shift and which ones are corrupted with a trend. When this information is not known, it must be guessed using these techniques, and it appears that, in their present form, the Kohonen network and fuzzy c-means are not accurate enough to make the distinction between a shift or trend when either one of them is present in the data sample. Indeed, the distance and clustering properties are quite similar when a shift or a trend is present. The fact is that statistical tests are not immune from this default either. When a shift or a trend is present, it is highly possible that both statistical tests for shifts and trends respond positively at the same time. However, unlike statistical tests, the Kohonen network and fuzzy c-means offer room for improvement since they can accommodate multiple types of inputs. In further developments, the task is therefore to examine the advantages of this feature.

6. CONCLUDING REMARKS

The results presented in this paper show that the statistical tests that have been used commonly for the detection of shifts or trends are not totally reliable under any circumstance. They can be considered as performing poorly when the coefficient of variation of the data sample is high and when the amplitude of the shift or the trend is small. For the purpose of studying inflow regimes in watersheds, such tests can likely be used on annual inflow means since the coefficient of variation for such data is usually low. It is debatable whether it would be prudent to use these tests on annual inflow peaks since the coefficient of variation of such data is sometimes very high. As for the amplitude of the shift or the trend, one can never know this in advance. It would be highly desirable to have a technique that is able to reliably detect shifts and trends of low amplitude. An overall 5% change of the mean in the inflow regime can already make the difference between a profitable and non-profitable water resources project, and no method performs very well for this kind of amplitude.

The artificial intelligence techniques developed for this paper for the detection of shifts and trends replicate the behavior of the usual statistical tests, and therefore produce similar results to these tests. However, their capacity to accommodate multiple types of inputs may offer some advantage.

7. REFERENCES

Conover, W.J., *Practical Nonparametric Statistics, 2nd Edition*, John Wiley & Sons, 493 pp.,New York, 1980.

Hall, M.J. and A.W. Minns, The classification of hydrologically homogeneous regions, *Hydrological Sciences Journal*, 44(5), 693-704, 1999.

Kohonen, T., The self-organizing map, *Proceedings of the IEEE*, 78(9), 1464-1480, 1990.

Lettenmaier, D.P., E.F. Wood and J.R. Wallis, Hydro-climatological trends in the continental United States, 1948-88, *Journal of Climate*, 7(4), 586-607, 1994.

Lins, H.F. and J.R. Slack, Streamflow trends in the United States, *Geophysical Research Letters*, 26(2), 227-230, 1999.

Ross, T.J., *Fuzzy Logic with Engineering Applications*, McGraw-Hill, 600 pp., New York, 1995.

Salas, J.D., Analysis and modeling of hydrologic time series, *Handbook of Hydrology*, D.R. Maidment (ed.), McGraw-Hill, Chapter 19, New York, 1993.

Yulianti, J.S. and D.H. Burn, Investigating links between climatic warning and low streamflow in the Prairies region of Canada, *Canadian Water Resources Journal*, 23(1), 45-60,1998.

Zhang, X., K.D. Harvey, W.D. Hogg and T.R. Yuzyk, Trends in Canadian streamflow, *Water Resources Research*, 37(4), 987-998, 2001.