ICIWRM – 2000, Proceedings of International Conference on Integrated Water Resources Management for Sustainable Development, 19 – 21 December, 2000, New Delhi, India

Group based estimation methods for missing values in hydrologic data sets

U. S. PANU
Department of Civil Engineering, Lakehead University, Thunder Bay, Ontario, P7B-5E1, CANADA
W.C. LENNOX
Department of Civil Engineering, University of Waterloo, Waterloo, N2L-3E1, CANADA

Abstract
The concepts of seasonal groups and neural networks and their characteristics are the focus of this paper in estimating missing values in monthly streamflows. The group approach recognizes the utility of associative and distributive properties of data points at local and global levels across the data series. At the local level, the associative properties are identified and used in the formation of groups; whereas, at the global level the distributive properties across the data series are recognized and used in the formation of group-clusters. The formation of groups and group-clusters enhances the extraction and utilization of information content of the data set and, thus enhances the development of effective data infilling methods and techniques. Efficacy of the approach for data infilling in monthly streamflow time series has been demonstrated with reasonable degree of success through applications to five rivers across Canada.

INTRODUCTION

Data sets of various hydrologic variables are at times not only short, but also often have gaps because of missing observations. Time series methods, among others, do not tolerate missing observations, and thus numerous data infilling techniques have evolved in various scientific disciplines to deal with incomplete data sets. One could envision two basic problems in dealing with inadequate hydrologic data series. In the first case, the time series are of adequate time length but suffer from the presence of data gaps. In this case, data infilling has been referred to as data augmentation. In the second case, the historic time span of the data series is inadequate, and thus efforts are made to extend the historic time span to a desired one. This latter case of data infilling has commonly been referred to as data extension. Towards modeling of hydrologic data series for use in the design of water resources systems, it is imperative that all relevant characteristics of hydrologic time series be considered. For example, the monthly streamflow time series exhibit the presence of heterogeneous and nonlinear relationships among data points [Panu and Unny (1980), Unny et al. (1981), Panu and Afza (1993), Goodier and Panu (1994), Panu and Ku (1997), and Panu et al. (2000)]. This paper develops hydrologic data infilling techniques capable of encompassing nonlinear characteristics of hydrologic data and also of overcoming the problems that may arise when the importance of heterogeneous relationships among data points is ignored. In such considerations, groups instead of individual observations form the basis of development.

Groups are considered to possess certain characteristics that distinguish them uniquely as separate entities. In other words, the variables (or elements) that form a group represent a certain uniqueness of the group. Time dependent observations in a time series, which form groups collectively, reflect the unique characteristics of such groups. In this paper, for example, hydrologic observations corresponding to seasonal periods of dry and wet are considered to form groups [Elshorbagy et al. (2000), Khalil et al. (1998), Unny et al. (1981), Panu et al. (1978) and others]. On the other hand, the term grouping refers to the formation of cluster in-groups.

The problem of missing values has been addressed not only in different ways but also from different perspectives. The definitions of 'missing data' and the mathematical formulations used to express various infilling processes are no less diversified than the variety of different techniques used. The methods and techniques of estimation of missing values can be classified into several ways, such as those related to univariate versus multivariate, regression versus non-regression (i.e., time series analysis), or any other basis of classification. However, it is perhaps prudent and desirable to classify the methods and techniques on the basis of single-valued versus group-valued. The single-valued approach entails all those methods and techniques in which each datum is treated as an individual observation. The group-valued approach encompasses all those methods and techniques in which individual observations are considered to be integral members of specific sequences or segments [Panu et al. (1978)]. The majority of data infilling methods and techniques are based on the single-valued approach except for the group-valued approach in Panu (1991), Panu and Afza (1993), Panu and Afza (1997), Khalil et al. (1998), and Elshorbagy et al. (1999). A brief but relevant assessment of existing methods and techniques of data infilling has been provided recently by Panu et al. (2000).

SOME OBSERVATIONS ON METHODS & TECHNIQUES OF DATA INFILLING

The individual observations forming a data set collectively define the inherent characteristics of the data set. Such characteristics can be described and enhanced through considerations of groups of individual values. The utility of such groups in streamflow synthesis, water demand forecasting, and data infilling have been demonstrated in water resources. With the exception of group-valued data approach (i.e., pattern recognition based methods), the existing techniques implicitly or explicitly invoke the assumption of linear relationships among variables. Tong (1983) described the drawbacks of linear models and pointed out their inadequacy in the prediction of sudden bursts of streamflows with large amplitudes at random time intervals. Tiao and Tsay (1989) also described some of the difficulties that may occur with linear relationships in multivariate models. Due to these difficulties, many nonlinear statistical models have been developed [Granger and Newbold (1986), and Tong (1990)].

In this paper, the internal structure (i.e., the intra-structure) of groups is explored to describe this structure through the use of ANN (artificial neural networks), PR (pattern recognition), and AR (auto regressive) techniques. The interrelationship (i.e., the inter-structure) among various groups is explored in an effort to develop a methodology for infilling one or more of either full or partial missing segments. Recent advances in ANN along these lines are encouraging [Tong (1983), Granger and Newbold (1986), Tiao and Tsay (1989), Tong (1990), Chakraborty et al. (1992), Khalil et al. (1998), and Elshorbagy et al. (1998)].

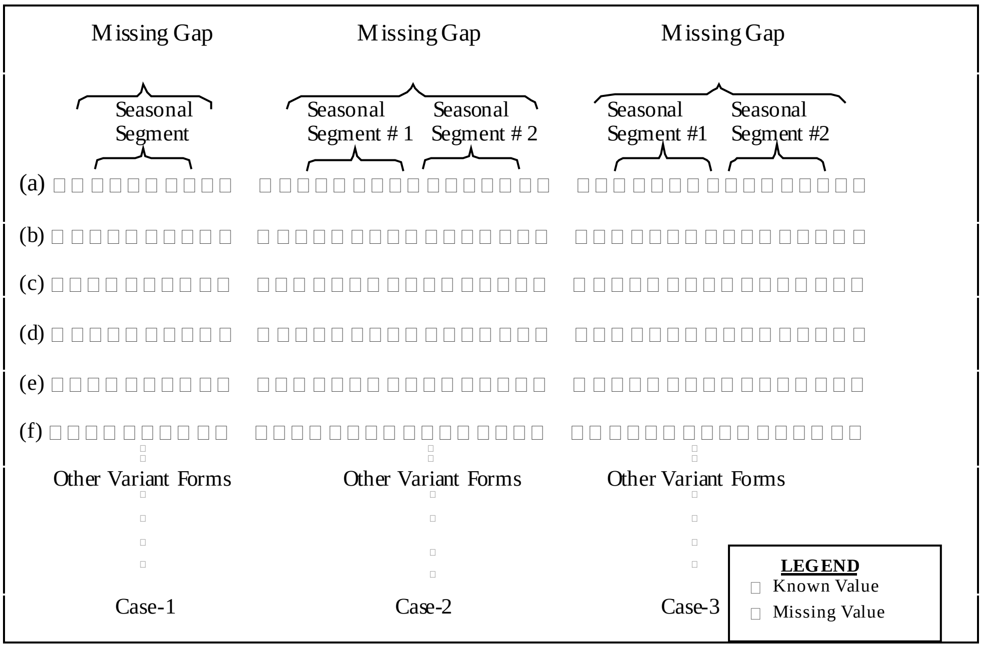

Figure 1. Various considerations in gap infilling based on group concepts.

DEFINITION OF DATA GAPS

Towards defining a data gap, Panu (1991) stated that "an exact description of a data gap satisfying the varying nature of its interpretations by various researchers is difficult, because the significance and the complexity of a data gap varies according to the time-scale usage of a time series. Furthermore, a data gap of variable duration may occur at one or more locations within a time series. Such a variability in the duration of a data gap increases as the time-scale usage of such series is altered from a yearly to a monthly basis." He further stated that "in traditional data infilling procedures, a single data value is estimated at a point in time and therefore, the variability in the duration of a data gap is inconsequential. However, it is noted that the error of estimation increases after the very first infilled value, because the infilled value or values, in turn, are treated as known values for the estimation of subsequent missing data values. On the other hand, in the group-valued approach, a seasonal segment (i.e., a pattern) of finite duration is projected onto the data gap. As a result, the estimation error over the duration of a pattern is time invariant". In addition, he implied that "the estimation error can further be reduced by infilling of data gaps by patterns (i.e., groups) based on the following considerations. (i) *Duration of the data gap is less than the duration of the patterns.* Such a case is simple to deal with because one or more data values are known either at the beginning or at the end of the data gap. (ii) *Duration of the data gap is equal to the duration of the patterns.* Such a case is not simple, at least for the scenario where all missing data values in the gap exactly match the duration of the patterns. (iii) *Duration of the data gap is greater than the duration of the patterns.* Such a case is complex, at least for the scenario in which a sequence of data values equal to the duration of patterns is missing either at the beginning, or in the middle, or at the end of the data gap". For the above cases, Panu (1991) describes some scenarios in each case are shown in Figure 1.

STOCHASTIC DATA INFILLING BASED ON GROUP CONCEPTS

The problem of the infilling of the missing data is viewed as the process in which one or more of the full (or partial) segments (i.e., groups) are either sequentially or intermittently missing. For the estimation of missing data values, it is envisioned that intra- and interrelationships among groups can be developed, either based on the information contained in the data set with missing values or from information available in the form of other relevant data sets. In the hydrologic context, such data sets may pertain to nearby streamflows or to data sets of precipitation or temperature at the site with missing data values. A methodology utilizing the information contained in groups of data and the concepts of ANN is presented for the infilling of missing segments of hydrologic time series. In this sense, the methodology differs from the traditional approaches in stochastic hydrology. Although in this paper the method is applied to monthly streamflow data, it can easily be extended to weekly and daily hydrologic time series. Monthly hydrologic time series are often short and exhibit a nonlinear multi-variable nature; thus it is difficult for linear models (e.g. multiple regression, seasonal ARIMA, and pattern recognition (PR) based techniques) to infill accurately the data gaps. Alternate methods capable of dealing with this complexity are the focus of considerable research. In identifying and studying the intricate nature of hydrologic data, neural networks coupled with seasonal grouping offers one such alternative to linear statistical methods [Khalil et al. (2000), and Elshorbagy et al. (2000)].

Development of Stochastic Data Infilling Models

Based on the consideration of group-valued data approach and the suitability of structural composition of ANN, models are proposed for cases involving the following: (1) only single data series with gaps, and (2) data series with gaps along with the availability of one (or more) concurrent but complete data series from neighboring stations. Two types of models, namely, the autovariate series (i.e., the series with gaps) and the bivariate series (i.e., the series with gaps and another series with complete but concurrent records) are proposed. In the autovariate series case, the relationships among groups within the data series with missing values are utilized. In the bivariate series case, the relationships among groups of two (or more) concurrently occurring data series (one series with missing data values and another series with complete data values) are utilized. In the bivariate series case, complete data series may be available at an upstream/downstream location, or at a tributary of the same river, or at a tributary of an adjacent river. The concurrent complete data series need not be of streamflows but could be of rainfall, temperature, evaporation, etc. In this paper, concurrent and complete data series considered are only of streamflows. Parameters of the neural network (e.g., number of hidden layer(s), number of nodes in the hidden layer(s), and number of epochs used to stop the network) were experimentally determined [Khalil et al. (1998)] by choosing values that minimize the mean squared error (mse) between the actual and the infilled data.

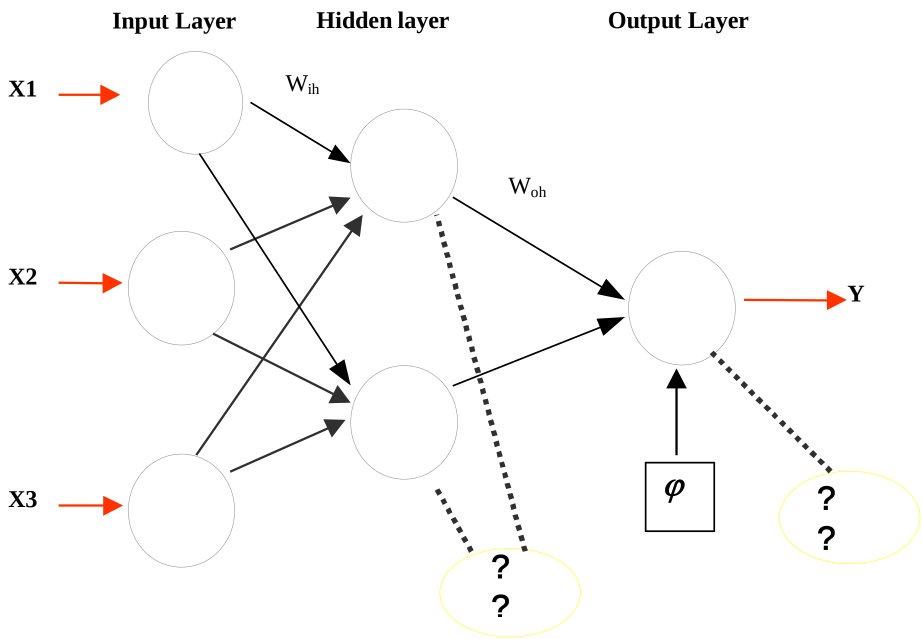

Figure 2. Architecture of a simple neural network.

The basic elements of neural networks are neurons, which are extremely simple processing elements. The identification of neurons (Figure 2) involves three basic elements, as follows: (1) the input vectors, ‘xi’, each of which ischaracterized by a weight of strength ‘wih’(The weight, ‘wih’, can either be positive or negative); (2) an adder for summing the input signals that have been weighted by respective synapses of the neurons with a linear combiner, and (3) an activation function ‘φ’ for amplitude limitation of the output of the neuron to some finite value and also a threshold ‘θ’ that has the effect of lowering the input of the activation function

Multi-layer-Feed-Forward Autovariate Series Model (M-ASM)

Using the group-valued data approach, the autovariate lag-one series model of the type suggested by Panu (1991) and developed by Afza and Panu (1992) is proposed to infill the missing seasonal gap(s) for a seasonal streamflow data set. In this paper, a Markovian relationship between the seasonal segment prior to the missing seasonal segment and the missing seasonal segment is assumed. The governing equation for this type of model can be expressed as follows.

$$Qi = f(Qi\text{-}1) \quad (1)$$

Where, Qi is a vector of monthly streamflows at season i, while Qi-1 is a vector of monthly streamflows at season i-1. The M-ASM is formulated as a fully connected back propagation network, involving input, hidden, and output layers (Figure 2). The number of nodes in the input and in the hidden layers depends on the number of months that define a season in the data set. The network is based on a NN(3,7,3) configuration. A nonlinear logistic (sigmoid) function is used for activation to map the nonlinearity of streamflow data.

Multi-layer Feed-Forward Bivariate Series Model (M-BSM)

The bivariate series model deals with infilling the missing values of the subject river (i.e., the river with missing data values) and uses the prior information obtained from a similar data series from nearby rivers (also called the base rivers). The streamflows from base rivers exhibit synchronous seasonality and are cross-correlated to each other. A lag-zero bivariate series model of the type suggested by Panu (1991) and developed by Afza and Panu (1992) is proposed to infill the missing seasonal gap(s) in a seasonal streamflow data set. In this model, the data series at base stations can be any other nearby streamflow data, any precipitation data or any information that strongly affects the streamflows at the subject river. This model can be expressed as follows:

$$Qsi = f(Q1bi , Q2bi ,\ldots Qnbi ,Pi) \qquad (2)$$

Where, Qsi is a vector of monthly streamflows of the subject river in season i, Qjbi is a vector of the monthly streamflows of the base rivers, j = 1,2,...,n, and Pi is a vector of the precipitation data in season i collected on the watershed of the river with data gaps.

In the case of M-BSM, the network consists of an input layer, hidden layer(s), and an output layer. Similar to the M-ASM model, the number of nodes in the input and the hidden layers depends on the number of months that define a season in the data sets. One can develop neural networks based on NN(3,7,3). As noted earlier, a nonlinear logistic (sigmoid) function is used as an activation function.

Evaluation Basis of Stochastic Data Infilling Models

The model performance indicators based on statistical considerations as described (Panu et al. (2000) and Khalil et al. (1998)), the proposed ANN-based models, namely the autovariate series model (M-ASM) and the bivariate series model (M-BSM), are compared to evaluate the best reliable approach for the selection of a suitable data infilling model. To evaluate the suitability of the proposed ANN-based models, it is necessary to compare the estimation capabilities of ANN with those of existing models. This comparison is usually made by testing all the models of interest on the same data set. The bivariate multi-dimensional regression (MR) and the pattern recognition (PR) models are used for comparative analysis.

APPLICATIONS TO MONTHLY STREAMFLOWS

In applying ANN-based models to watersheds, the issues addressed were (1) the effect of seasonal cycles on the estimation accuracy, and (2) the effects of ANN-based models on the quality of infilled data using seasonal groups in case of (a) the autovariate series, and (b) the bivariate series. Seasonal groups and their characteristics were investigated in the context of neural networks for the infilling of missing data values in monthly streamflows. The efficacy of the proposed methodology was tested on five different watersheds across Canada. The methods of selection and preparation of pertinent data sets are described (Panu et al. (2000), and Khalil et al. (1998)).

In this paper, all the selected streamflow data sets were complete and thus exhibit no data gaps. However, for testing of the various models, a variety of data gaps were randomly created in each streamflow data set. To evaluate the overall performance of any proposed model, the data duration equal to one season in any year was assumed to be missing and then was subsequently infilled by the model. Such an operation for any model was successively repeated for a streamflow data in each year until all seasons over the years were assumed to be missing and then succeedingly infilled. The first season in the data set for the autovariate series model (M-ASM) was not estimated because the model is formulated to simulate only for the forward shift (e.g., season i-1 to season i).

For the training and testing phases of the neural networks (or the phases of calibration and verification of parametric models), the data set for each site or a site-pair were divided into two equal halves. The first half of the data set was used in the training phase, while the other half was used in the testing phase of the models developed in this paper.

The seasonality of streamflow data sets was determined based on correlation and spectral analyses. The resulting correlograms and periodograms [Khalil et al. (1998), (1999)] exhibit a strong indication of the presence of twelve-month seasonality and also six-month seasonality (i.e., two seasons of six-month duration) in all of the five-paired sites of streamflow data sets. The presence of three and four seasons respectively, each of four-month and three-month duration, is apparent from the correlograms [Khalil et al. (1998)] of the Osilinka, Graham, and Nagagami rivers. Further, the existence of four seasons of three-month duration is also supported by the periodograms, especially for the previously mentioned three rivers.

A weak indication of the presence of four seasons of three-month duration is indicated in the English River [Panu et al. (2000), and Khalil et al. (1998)]. This is probably due to the storage capacity of such large watersheds. An experiment to infill the missing data using two seasons of six-month length was also conducted for this river. The results of this experiment were then compared to PR-based models presented by Goodier and Panu (1993, 1994). By an iterative method, the starting and ending months of the four seasons, each of three-month duration, were determined to be November to January, February to April, May to July, and August to October in all the rivers except the Nagagami River. The seasons thus defined were respectively referred to as the semi-dry, dry, wet, and semi-wet seasons. The seasons for the Nagagami River were identified to be January to March, April to June, July to September, and October to December and thus were accordingly classified as the dry, wet, semi-wet, and semi-dry seasons [Khalil et al. (1998)].

Multivariate normality for each of the four seasonal segments in all streamflow data sets was tested because of the underlying requirement of multivariate normal data sets by the multi-dimensional regression and pattern recognition models. It was determined that streamflow data normality was best achieved through the use of a natural logarithm transform. Based on the results of the sample χ^2-statistic [SPSS Computer Package (1995)], and the Mahalanobis distance [Gnanadesikan (1977)]; the seasonal segments in all rivers were found to follow the assumption of multivariate normality [Khalil et al. (1998)].

PERFORMANCE ASSESSMENT OF STOCHASTIC DATA INFILLING MODELS

The proposed ANN-based (M-ASM and M-BSM) models were first assessed among themselves, and subsequently were also assessed in comparison with the existing PR- and MR-based models. For this purpose, the models were applied to various data sets involving autovariate series and bivariate series.

Comparison between of the ANN-based models (M-ASM and M-BSM)

Test results pertaining to the M-ASM and M-BSM models are indicative of acceptable performance of these models. It was observed that M-BSM (i.e., bivariate case) performs, as expected, much better than the M-ASM (i.e., autovariate case) in all rivers except the English River. Based on *rme* considerations, the M-BSM models were found, on average, to be 36% better than the M-ASM models. On a seasonal basis, the enhanced performance of the M-BSM models over the M-ASM models was found to vary from 21% to 90 % [Khalil et al. (1998)]. The superior performance by the M-BSM models over the M-ASM models was found not only across the spectrum of seasonal considerations but also across geographical variations that were represented through the use of different streamflow data sets. In general, the values of cross-correlation between participating rivers in the M-BSM models were higher than the values of lag-one auto-correlation for rivers in the M-ASM models. A higher value of auto- or cross-correlation coefficient respectively indicates the relative suitability of auto- or cross-series models to estimate missing values. Similar observations on the relative suitability of the bivariate series models were apparent in all the five rivers. These observations are consistent with those of Afza and Panu (1992) and Goodier and Panu (1994) for the bivariate series models based on the concepts of groups and pattern recognition.

Comparison between the ANN-based and MR-based Models

The efficacy of the ANN based models (M-ASM and M-BSM) was examined in relation to the existing multi-dimensional regression (MR) models. In general, the bivariate series models (B-ASM) were found to be better than the autovariate series models (M-ASM), as discussed below.

Comparison between the M-ASM and MR Models

At various levels of the seasonal comparison, the values of *rme* and *noise* terms in the case of the M-ASM model were smaller except in the dry season in all the rivers. The values of the *rme* at the various seasonal levels range between 0.289 and 0.419 for the M-ASM model, and between 2.84 and 8.67 for the MR model. In other words, the M-ASM model was relatively more capable of handling the nonlinear characteristics of monthly streamflows than the multi-dimensional regression (MR) model. Due to small flow variation during the dry season, it is noted that the MR model was found to be better in estimating the missing values than the M-ASM. In an overall assessment, the M-ASM model appears to be a promising estimator of the missing values for the wet, semi-wet, and semi-dry seasons in autovariate series.

Comparison between the M-BSM and MR Models

The results indicate that the M-BSM model was slightly more capable of handling the nonlinear characteristics of monthly streamflows. Similar to the M-ASM and MR models during the dry season, the MR model was again found to be better in estimating the missing data than the M-BSM model. In case of the M-BSM model, the values of rme are smaller for all seasons except for the dry season. For the bivariate series, it is noted that the performance of both the M-BSM and the MR models was superior to the corresponding M-ASM and MR models in the autovariate series. Further, the performance of the M-BSM model was even better than the MR model in the bivariate series at all seasonal levels and in different streamflow data sets. In the dry season, the slight advantage of the MR model over the M-BSM model was insignificant. In other words, the M-BSM model appears to be a promising estimator of missing values for the wet, the semi-wet, and the semi-dry seasons in bivariate series.

It is further noted that the MR models exhibit better estimation capabilities for the missing values in rivers that have upstream/downstream relationships (English River and Halfway River). This may be due to the presence of a linear relationship between the upstream and downstream sites. However, the ANN-based models show relatively good estimation in the wet season for all sites except the English River. In other words, the ANN-based models possess a higher capability for mapping the streamflow characteristics during the moderate to high flow seasons.

Comparison among ANN-based, PR-based, and MR-based Models

Results of PR-based models [Goodier and Panu (1994)] for the two seasonal segments (i.e., two seasons) of six-month duration were available in the case of the English River at Sioux Lookout. It was indicated earlier for this river that the correlogram and the periodogram did not exhibit strong evidence of the existence of seasonality of four seasons of three months duration. Three types of models (ANN-based, PR-based, and MR-based models) were applied in both the cases of the autovariate and the bivariate series. An ANN-based model of NN(6,7,6) was developed and used for comparative analyses.

In the case of the autovariate series, the results of the ANN- and PR-based models were found to be comparable. In general, the ANN-based models performed better during all seasons. For example, the values of rme were respectively 0.32 and 0.45. However, for specific cases of semi-dry, wet, and semi-wet seasons, the PR models gave the best results. It is again noted that the MR-based models were the best estimators during the dry season.

The values of rme were found to improve slightly in the MR-based model when two seasons of six months duration instead of four seasons of three months duration were used in the modeling of the English River. In summary, the ANN-based models perform better in cases of both the autovariate and bivariate series.

CONCLUSIONS AND RECOMMENDATIONS

Neural networks appear to be good candidates for infilling of the partial or full missing seasonal groups in monthly streamflow time series. For most rivers, except the English River and the Halfway River, the ANN-based models provided relatively more accurate estimates of the values in the data gaps. For both types of models (M-ASM and M-BSM), the average percent improvement in terms of rme during the testing phase was 53.4%, thus indicating that the ANN-based models are better for data infilling in seasonal groups than the multi-dimensional regression (MR) models. The MR-based models for the autovariate series have been found to possess relatively poor estimation ability for streamflow data infilling. This poor estimation is quite pronounced for the small watersheds (i.e. the Graham, the Halfway, and the Osilinka). However, the MR based-models did provide slightly better estimates for missing values in larger watersheds (i.e. the English and the Nagagami). In these rivers, a slightly better performance by the MR models can be related to the fact that the value of auto-correlation at the third (month) lag and beyond was low. Further, the larger watersheds appear more stable, and therefore less prone to the effects of extreme events on the seasonal variations in streamflows.

In comparative analyses among the ANN-, PR-, and MR-based models, the ANN-based models were found to provide relatively better estimates of missing values during the testing phase. The average percent improvement in the ANN-based models compared to the PR- and MR-based models was found to be respectively 33.6% and 54.8%. The average percent improvement in the PR-based models over the MR-based models was found to be 45.1%. In other words, the ANN-based models, in general, are more accurate for data infilling in seasonal groups than the other existing models, and the PR-based models are more accurate than the MR-based models.

On the basis of results and discussion presented in this paper, the proposed group-based methodology in general and the ANN-based methods in particular appear to be promising for infilling of missing values in streamflow time series. Towards the development of effective and efficient data infilling procedures, additional investigations involving groups and artificial neural networks may entail the use of prior information either from more sites forming the base-set or from the mixed records of precipitation, temperature, and/or snowmelt data.

References

Afza, N., and Panu, U.S. (1992), Infilling Missing Monthly Streamflow Data for Rivers with Seasonal Runoff, Civil Eng. Tech. Report, No. CE-92-3, Lakehead University, Ontario.

Chakraborty, K., Mehrotra, K., Mohan, C., and Ranka, S. (1992), Forecasting the Behavior of Multivariate Time Series Using Neural Networks, Neural Networks, 5, 961–970.

Elshorbagy, A., Simonovic, S. P. and Panu, U. S. (1998), Performance Evaluation of Artificial Neural Networks for Runoff Prediction. ASCE Journal of Hydrologic Engineering (in press).

Elshorbagy, A., Panu, U. S. and Simonovic, S. P. (1999), Investigations into Group Based Data Infilling Techniques. CSCE Annual Conference, Regina, Canada.

Elshorbagy, A., Panu, U. S. and Simonovic, S. P. (2000), Group-based Estimation of Missing Hydrological Data. Part-1 Approach and General Methodology, Hydrological Sciences Journal (Accepted for Publication).

Gnanadesikan, R. (1977), Methods for Statistical Data Analysis of Multivariate Observation, John Wiley, New York

Goodier, C. and Panu, U.S. (1993), Applications of a Multivariate Approach for Infilling of Missing Monthly Streamflows, Civil Engineering Technical Report No. CE-93-3, Lakehead Univ., Thunder Bay, Ontario.

Goodier, C. and Panu, U.S. (1994), Infilling Missing Monthly Streamflow Data Using a Multivariate Approach, Stochastic and Statistical Methods in Hydrology & Environmental Engineering, Kluwer Academic Publishers, 3, pp.191-202.

Granger, C. W., and Newbold, P. (1986), Forecasting Economic Time Series. Orlando, Academic Press.

Johnson, R. A., and Wichern, D. W. (1988), Applied Multivariate Statistical Analysis, Prentice Hall, N.Y.

Khalil, M., Panu, U. S. Panu, and Lennox, W.C. (1998), "Infilling of Missing Streamflow Values Based on Concepts of Groups and Neural Networks, Civil Engineering Technical Report No CE-98-2, Lakehead University, Thunder Bay, Ontario, Canada.

Khalil, M., Panu, U. S., and Lennox, W. (2000), "Group and Neural Networks Based Streamflow Data Infilling Procedures", Journal of Hydrology (Accepted for Publication).

Panu, U.S., Khalil, M., and A. Elshorbagy (2000), Chapter 12: Streamflow Data Infilling Techniques Based on Concepts of Groups and Neural Networks in Artificial Neural Networks in Hydrology, Govindraju, R.S. and A.R. Rao (Editors), Kluwer Academic Publishers, The Netherlands, pp. 235-258.

Panu, U. S. (1991), Application of Some Entropic Measures in Hydrologic Data Infilling Procedures. In Entropy and Energy Dissipation in Water Resources. Singh, V. P. and Fiorentino, M., eds., Kleuwer Academic Publishers, the Netherlands, 175-192.

Panu, U.S., Unny, T.E. and Ragade, R.K., 1978. (1978), A Feature Prediction Model in Synthetic Hydrology Based on Concepts of Pattern Recognition. WRR, 14(2), 335-344.

Panu, U. S. and Unny, T. E. (1980a), Extension and Application of Feature Prediction Model for Synthesis of Hydrologic Records. WRR, 16(1), 77-96.

Panu, U.S. and Unny, T.E. (1980b), Stochastic Synthesis of Hydrologic Data Based on Concepts of Pattern Recognition, 1-. General Methodology of the Approach, Journal of Hydrology, 46, 5-34.

Panu, U.S. and Unny, T.E. (1980c), Stochastic Synthesis of Hydrologic Data Based on Concepts of Pattern Recognition, 2-. Application to Natural Watersheds, Journal of Hydrology, 46, 197-217.

Panu, U.S. and Unny, T.E. (1980d), Stochastic Synthesis of Hydrologic Data Based on Concepts of Pattern Recognition, 3-. Performance Evaluation of the Methodology, Journal of Hydrology, 46, 219-237.

Panu, U. S., and Afza, N. (1993), Entropic Evaluation of Streamflow Data Infilling Procedures, Proceedings of Stochastic and Statistical Methods in Hydrology and Environmental Engineering, pp. 410–412.

Panu, U. S. and A. Ku (1997), Forecasting Monthly Streamflow Patterns for Reservoir Operations. CSCE Annual Conference, 3, 159-168.

Panu, U. S. and Afza, N. (1998), Development of Feature Infilling Procedures for Hydrologic Data Series. Journal of Hydrology, (to appear).

SPSS (1995), Base Systems User's Guide (Part-II), SPSS Inc., Chicago, Illinois, USA

Tiao, G. C., and Tsay, R. S. (1989), Model Specification in Multivariate Time Series, Journal of the Royal Statistical Society, B 51, 157–213.

Tong, H. (1983), Threshold Models in Nonlinear Time Series Analysis, Lecture Notes in Statistics, 21, New York: Springer-Verlag.

Tong, H. (1990), Nonlinear Time Series: A Dynamical System Approach, Oxford: Oxford University Press.

Unny, T.E., Panu, U.S., McInnes, C.D., and Wong, A.K.C. (1981), Pattern Analysis and Synthesis of Time Dependent Hydrologic Data. Advances in Hydrosciences, Academic Press, 12, 222-244.