

PERFORMANCE OF HYDROLOGICAL DATA INFILLING TECHNIQUES USING ENTROPY APPROACH: EXPECTATION MAXIMIZATION ALGORITHMS.

M. Ilunga¹ And D. Stephenson²

¹ *PhD Candidate, Civil Engineering, Wits University, South Africa. Tel: +27-11-7717152
Fax: +27-11-3391762, e-mail: masengoi@civil.wits.ac.za*

² *Prof. Hydraulics, Civil Engineering, Wits University, South Africa. Tel: +27-11-7177154
Fax: +27-11-4032062, e-mail: Steph@civil.wits.ac.za*

Abstract

Hydrological data (e.g. rainfall, river flow) are used in water resources and management for planning reservoir size and operation. However, very often in developing countries, hydrological time series have gaps or are incomplete, or are not of good quality or are not of sufficient length. In this paper, statistical techniques namely standard Expectation-Maximization (EM) algorithm and the EM-Constrained (ECM) algorithm are evaluated using entropy approach, i.e. mutual information (T)/directional information transfer index (DIT) for filling in the hydrological data, i.e. river flows. Empirical comparisons of the predictive accuracy, in terms of DIT/T between the different techniques are then made. Another model accuracy criterion such as Root Mean Square Error of predictions (RMSEp) is also used to crosscheck the results. A preliminary case study (i.e. Diepkloof gauge on Wonderboomspruit River and Moltero gauge on Stormbergspruit River) is made to demonstrate the performance of these two techniques.

Introduction

For planning, management and effective control of water resources systems, a considerable amount of data on hydrologic variables such as rainfall, streamflow, etc. are required. Very often in developing countries, hydrological data sequences at a given network have gaps or are incomplete, or are not of good quality or are not of sufficient length. This can severely affect the reliability of the design, e.g. hydropower plant, construction of water storage, etc.

Traditionally, multivariate time series methods are the mostly used for information transfer between two or more points; e.g. linear regression (Chow, 1964; Makhuvha et al., 1997). In most regression methods, the parameters are derived from the concurrent data between two or more sites and are then used for infilling the gaps. These methods can suffer where information is little. However, there are techniques (within the regression methods), which cope with the missing data, i.e. the Expectation-Maximization (EM) techniques. The EM technique and its extensions have been used intensively for infilling records for the last two decades (e.g. Makhuvha et al., 1997). The measurements of information content or transfer (i.e. variance, cross-correlation) could also suffer where information could be little. This led some researchers to quantify conveniently the information transfer between hydrological variables by using entropy approach (Amorocho and Espildora, 1973; Singh and Florentino, 1992; Hamancioglu and Yevjevich, 1987). In this respect the theoretic entropy was found to be more appropriate where information is limited or incomplete; i.e. developing countries (Singh, 1998). The mutual information was used as information transfer

among rivers (rainfall) points (Amorocho and Espildora, 1973). Later, the directional information index appeared as the generalization of the mutual information (Yang and Burn, 1994).

The literature of EM techniques in hydrology is very sparse; specifically the ECM has not been applied in this discipline. This paper will mainly use the two concepts *accepting cases of little information (missing data)*, i.e. entropy, EM and ECM algorithms for infilling the missing values (i.e. streamflow) in a given hydrological network. EM and ECM performance is evaluated using entropy approach, i.e. mutual information (T). In this paper, the directional information transfer index (DIT) as defined by Yang and Burn (1994) is extended to model performance evaluation. The dependency between gauging stations is determined from the same approach as well. A procedure for infilling hydrological data is proposed and a preliminary test is done on a case study. Algorithm and technique can be used interchangeably for the rest of this paper.

Hydrological information transfer

Generally the majority of the current techniques are based on the classic correlation and regression theory, which basically constitutes a means of transferring information in space and time (Krstanovic and Singh, 1992; Harmancioglu and Yevjevich, 1987). In particular linear regressions have widely been used (Chow, 1964; Makhuvha et al., 1997 and many others). However, the correlation coefficient cannot take care of arbitrarily relation between coordinates and classes (Battiti, 1994). Moreover, the measurements of information content or transfer (i.e. variance, cross-correlation) could also suffer where information could be little (Singh, 1998), e.g. case of developing countries.

Since 70's hydrologists tried to find another way of measuring information by theoretic entropy (a term borrowed from communication, see Shannon and Weaver, 1949). Thus the concept has been applied in water resources (Singh and Florentino, 1992), in rainfall network design (Krstanovic and Singh, 1992), in river flow network design (Yang and Burn, 1994) and many other water related fields. Entropy is considered as a measure of the degree of uncertainty of random variable hydrologic processes (Amorocho and Espildora, 1973). Since the reduction of the uncertainty by means of making observations is equal to the amount of information gained, the entropy criterion indirectly measures the information content of a given series of data (Harmancioglu and Yevjevich, 1987). It arises that the distribution of the variable can be unknown a prior although some of its properties may be known, e.g. mean, variance, normality condition. These proprieties (information) enable to determine the distribution of the variable, which maximizes the entropy function. In this way the distribution is consistent with the available information, but retains maximum uncertainty within the feasible domain and thus ensure the least bias; that is the principle of maximum entropy (POME) introduced by Jaynes in 1968. The entropy approach is a versatile tool where information is little (Singh, 1998), e.g. case of developing countries. If two random variables X and Y are considered, then the information inferred by X about Y is represented by the mutual information, e.g. $T(X,Y)$ and its high value suggests a strong dependency between the two variables. Under the assumption that the marginal distributions of X and Y, e.g. $H(X)$; $H(Y)$, are normal (their joint distribution is normal too) and using the POME, it can be

shown that their mutual information is a function of their correlation coefficient (Krstanovic and Singh, 1992). The entropy computations for normal distributions are generally simpler than for other distributions (Amorocho and Espildora, 1973; Krstanovic and Singh 1992).

The mutual information as opposed to the correlation coefficient, can measure arbitrarily elements relation between variables and it does not depend on the transformation acting on the different variables (Battiti, 1994). Amorocho and Espildora (1973), Singh and Fiorentino (1992) suggested that the mutual information (between the observed and the simulated ones) could be used as a criterion in the selection of hydrological models; e.g. rainfall-runoff. The higher the value of T , the better is the model. Later the directional information transfer index (DIT) appeared as a generalization of the mutual information Yang and Burn (1994). In a station pair, DIT is physically the fraction of information transferred from one site to another (Yang and Burn, 1994). Yang and Burn (1994) defined DIT for dependency between gauging stations pairs. In this paper, it is suggested that since T is used for model performance assessment, its generalization (i.e. DIT) can be extended to model performance evaluation as well.

Streamflow data infilling techniques

Introduction

Traditionally, multivariate time series methods are the mostly used *for information transfer* between two or more points. At sites where one or more proxy data are available, one may develop regression models between two or more stations or two more variables at the same site (Chow, 1964). In most regression methods, the parameters are derived from the concurrent data between two or more sites and are then used for infilling the gaps, for example simple regression and its extensions (Alley and Burn, 1982). For design-oriented studies maintaining the statistical properties is the overriding objective while for accuracy of water balance equations the dominant objective is the minimization of the difference between the observed and the estimated values.

Within the regression methods, the standard EM technique and its modifications have been intensively used for infilling records for the last two decades. However the literature of EM techniques is very sparse in hydrology and water related fields. The only studies dedicated to hydrology and found in the literature are by Kuczera (1987) and Makhuvha et al. (1997).

The standard EM (in the context of linear regression)

The standard EM algorithm was first introduced formally by Dempster et al. (1977). Some modifications of the standard EM exist (e.g. Makhuvha et al., 1997, etc). The EM algorithm is an iterative procedure (with two steps, i.e. E and M) where the E-step (expectation) adjusts the values of the sufficient statistics, given the incomplete data and the current values of the parameters (mean, variance, etc). The M step (maximization) solves the likelihood equations using the adjusted values of the sufficient statistics in a sample of complete data. The repeated applications of the E and M steps lead ultimately to the maximum likelihood. The missing values are assumed to be missing at random (MAR), thus the mechanism of missing values does not depend on the missing values.

In the bivariate case (i.e. station pairs) with data missing from one site only, the assumption made is that observation pairs (y_{i1}, y_{i2}) from a bivariate normal distribution are independently and identically distributed as $N(\mathbf{m}, \Sigma)$, where \mathbf{m}, Σ are the mean vector and the covariance matrix. If $((y_{i1}, y_{i2}); i=1, 2, \dots, m)$ is m complete bivariate observation and $(y_{i2}, i = m + 1, \dots, n)$ is $n-m$ univariate observation, therefore as there are no missing values for site 2, its values will remain unchanged throughout.

ECM algorithm (in the context of linear regression)

No literature is available for ECM and its applications in hydrology. From Meng and Rubin (1993), the ECM algorithm performs the following:

Perform E-step as above, and then the first CM-step is carried out by finding the regression coefficient \mathbf{b} of y_{i1} on y_{i2} given the values of sufficient statistics computed in step E and the current values of the parameters. Given the computed value of \mathbf{b} , the second CM-step is performed by computing the conditional maximum likelihood estimate of mean, variance and covariance.

Proposed procedure for filling in the missing values

Distributions of hydrological variables are assumed to be normal, since the theory of entropy and EM techniques developed so far is much easier for these distributions. Missing data are assumed to be missing at random. The imputation method (Buck's) is used here to start EM and ECM. The procedure for filling in the missing values is summarized as follows:

1. Check dependency within each time series. If satisfied, reject the time series. Otherwise, continue.
2. Transform data to follow normality assumption *if they are not*.
3. Define station pairs and compute T (DIT). Control station satisfies $DIT > \text{Threshold 1}$, otherwise target station.
4. Apply EM and ECM to the station pairs and *estimate missing values*.
5. For each technique, compute T (DIT) between observed and simulated data. Compute the root mean square error of the predictions RMSEP to see whether DIT is reasonable. Select the technique for which $DIT > \text{threshold 2}$.
6. In case step 2 has been done, thus untransform the data only if inverse transformation is unbiased. Otherwise terminate.

A preliminary test was done with monthly-naturalized streamflow data of two rivers, namely Wonderboomspruit (D1H001) and Stombergspruit (D1H004) in South Africa (Middleton et al., 1994). The geographical location of these rivers is given in figure 1 and summarized in Table 1.

Table 1. Naturalized streamflow data.

Gauge	Name	River	Latitude	Longitude	Area (km ²)	Period of records	%Missing
D1h001	Diepkloof	Wonderboomspruit	31°00'03	26° 21 11	2397	1924-53	0
D1h004	Moltero	Stormbergspruit	31° 24'00	26° 22' 17	348	1924-53	0

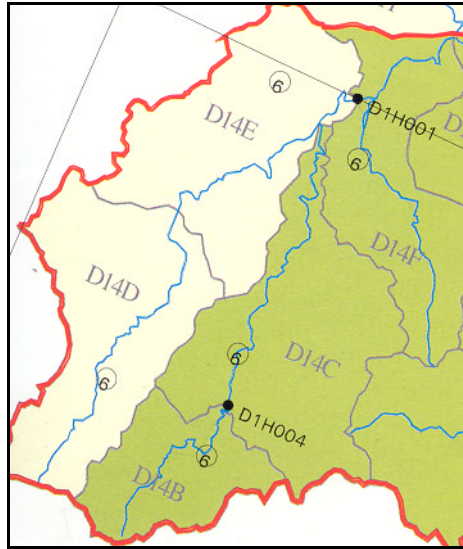


Figure 1 Geographical location of gauges D1H001 & D1H004

Results and discussion

The selected streamflow data set was complete and thus exhibits no gaps. However, for testing of the different infilling techniques some gaps (e.g. 9 % of missing) were created on the data set. The raw data were found to be approximately normal as the Chi-square test at 95% confidence interval was not rejected. Tables 2 and 3 give the summary of the results. Considering only the last digit of each gauge, $T(1,4) / (4,1)$ is information inferred by gauge 1 / 4 about 4 / 1 and $DIT14 / 41$ being the fraction of information transferred from 1 / 4 to 4 / 1. From table 1 ($DIT14 > \text{Threshold } 1$), gauge 1 can predict better information 4 about 1 than 4 does. Therefore gauge 1 is regarded as control gauge and 4 as the target gauge. In other words the uncertainty of gauge is so higher (see H4) that it needs to be decreased by making inferences from gauge 1. From the high values of DIT/T , table 2 shows that both techniques, i.e. EM and ECM perform likely equally and can be used to fill in the gaps for this specific data set. The ECM shares all the appealing convergence properties of the EM (Meng and Rubin, 1993). The calculation for $RMSEp$ (0.87) also confirmed the results from entropy computations. Note that if gauge 4 were to be considered control gauge, $RMSEp$ would be 4.5. It is also interesting to note that the ratio of variance of the infilled series to the observed was 0.999 for both techniques. Thus the infilled series could be used in design-oriented studies (e.g. reservoir) or for evaluation of the severity and duration of hydrologic extremes.

Table 2. Stations dependency based on T (DIT) values (threshold 1= 0.15)

H1 (Napier)	H4 (Napier)	T14 (Napier)	T41(Napier)	DIT14	DIT41
1.940	3.730	0.449	0.449	0.236	0.120

Table 3. Model efficiency (threshold 2= 0.5)

Algorithm	DIT	T(Napier)	RMSEp (Mm3/month)	Iterations
EM	0.995	1.839	0.874	4
ECM	0.994	1.838	0.874	3

Conclusion

This study applied for the first time the ECM technique to missing data in hydrology problem and compares it with standard EM technique using entropy approach. The result is crosschecked through RMSEp. A new methodology for data infilling has been proposed. The results showed that the EM performs as well as ECM does on that specific station pair. At this stage, for 9% missing values both EM and ECM can be used to fill in the gaps (RMSEp = 0.87) or the infilled series could be used for design oriented studies (see ratio of variance 0.999) such as reservoir design, reservoir operating rules. It is suggested to investigate the validity of the two techniques on other regimes of data (e.g. peaks, mean annual, etc) and consider a wider range of missing values, say, 15%, 20%, etc.

REFERENCES

- Alley, W.M, Burns, A.W. (1983) Mixed extension of monthly streamflow records, *Journal of Hydraulic Engineering*, 109(10), 1272-1284.
- Amorocho, J. and Espildora, B. (1973) Entropy in the assessment of uncertainty in hydrologic systems and models, *WaterResources Research*, 9(6), p.1511-1522.
- Battiti, R. (1994) Using mutual information for selecting features in supervised neural network learning. *IEEE Transactions on Neural Networks*, 5(4), 537-550. *Proceedings of the Second International Hydrological Symposium*, p. 17-28.
- Chapman, G. (1985) Entropy as measure of hydrologic data uncertainty and model performance, *Journal of Hydrology*, 93, 111-126.
- Chow, V. T. (1964) Handbook of applied hydrology. Compendium of Water Resources Technology. McGraw-Hill, Inc.
- Dempster, A. P., Laird, N. M., Rubin, D. B. (1977) Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39, 1-27.
- Harmancioglu, N. and Yevjevich, V. (1987) Transfer of hydrologic information among rivers points. *Journal of Hydrology*, 93, 103-118.
- Harmancioglu, N.B, Alpaslan and Singh, V.P. (1994) Assessment of the entropy principle as applied to, water quality monitoring network design. *Stochastic and Statistical Methods in Hydrology and Environmental Engineering*, 3, 135-148.
- Hirsch, M.R., (1982) Comparison of four streamflow methods extension techniques, *Water Resources Research*, 18(4), 1081-1088.
- Krstanovic, P.F. and Singh, V.P. (1992) Evaluation of rainfall network using entropy: II. Applications, *Water Resources Management* 6, 295-314.
- Makhuva, T., Pregram, G., Sparks R., Zucchini, W. (1997) Patching rainfall data using regression methods. 1. Best subset selection, EM and pseudo-EM methods: theory, *Journal of Hydrology*, 198, 289-307.
- Meng, X.-L., and Rubin, D. B. (1993) Maximum likelihood estimation via the ECM algorithm: A general framework, *Biometrika*, 80, 267-278.
- Middleton, B.J., Pitman, W.V. and Midgley, D. C. (1994) Surface Water Resources of South Africa, 1990. WRC Report no. 298/3.1/94, First Edition.
- Shannon, C. E and Weaver W. (1949) The Mathematical Theory of Communication. University of Illinois Press Urbana, Chicago, London.
- Singh, P. (1998) Entropy as a decision tool in environmental and water resources, *Journal of Hydrology, Indian Association of Hydrologists*, 21(1-4), 1-12.
- Singh, P. and Florentino, M. (1992) A historical perspective of entropy applications in water resources, Kluwer Academic Publishers, Dordrecht, The Netherlands.
- Singh, V.P., and Krstanovic, P. F. (1987) A stochastic model for sediment yield using the principle of maximum entropy, *Water Resources Research*, 23(5), 781-793.
- Yang, Y. and Burn, H. (1994) Entropy approach to data collection network design, *Journal of Hydrology*, 94, 307-324.