PAPER • **OPEN ACCESS**

Pre-processing Streamflow Data through Singular Spectrum Analysis with Fuzzy C-Means Clustering

To cite this article: Najah Nasir *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **864** 012085

IOP Conf. Series: Materials Science and Engineering **864** (2020) 012085 doi:10.1088/1757-899X/864/1/012085

Pre-processing Streamflow Data through Singular Spectrum Analysis with Fuzzy C-Means Clustering

Najah Nasir[1], **Ruhaidah Samsudin**[1] **and Ani Shabri**[2]

[1] Fakulti Kejuruteraan, Universiti Teknologi Malaysia, 81310, Johor, Malaysia
[2] Fakulti Sains, Universiti Teknologi Malaysia, 81310, Johor, Malaysia

E-mail: najah6@live.utm.my

Abstract. One approach to improve water resource management is by making use of streamflow forecasts. In this study, eigenvector pairs were clustered by employing fuzzy c-means (FCM) during the grouping stage as an enhancement to the singular spectrum analysis (SSA) technique for data pre-processing. The FCM-SSA pre-processed streamflow data was then supplied to an auto-regressive integrated moving average (ARIMA) model for forecasting. The Department of Irrigation and Drainage Malaysia provided the monthly streamflow records of Sungai Muda (Jambatan Syed Omar) and Sungai Muda (Jeniang) for this research, wherein each was split into training (90%) and testing (10%) sets. The R software was the platform for building every FCM-SSA-ARIMA, SSA-ARIMA and ARIMA model, while the root mean squared errors and mean absolute errors were used to compare the performance between those models. The proposed FCM-SSA-ARIMA was discovered to be capable of surpassing the SSA-ARIMA and ARIMA models.

1. Introduction

Streamflow can be defined as the movement of water in a stream channel. It is among the central components of a hydrological cycle. In the planning and management aspect, the forecasting of streamflow is essential as it allows efficient operations of water resource systems. Hence, the application of streamflow data for forecasting is gradually becoming an active area of research in time series modelling. Streamflow forecasts can be generated with the presence of chronological hydrographs, i.e. the data for rate of flow over time of a stream.

Hydrologic models are often basic and theoretical depictions of the components in the hydrological cycle. These are mainly applied to comprehend the hydrologic processes and for making predictions. Fundamentally, hydrologic models can be classified into two types; process-based models and stochastic models. In process-based models, the model will attempt to represent the real-world physical activities. On the other hand, stochastic models are usually black box in nature, whereby it heavily relies on data and applies statistical concepts to relate the input data to the model output. The techniques frequently applied by researchers in this area include regressions, transfer functions and neural networks. Recently in the hydrologic modelling research, more focus has been given to the interpretation of hydrologic systems behaviour using a more universal approach. This is crucial to produce a more reliable prediction so that important issues in water resource management is possible to overcome.

Time series data are data arranged in a series of particular time intervals. Hence, time series analysis is an analysis of time series data to obtain significant statistical interpretations of the series [1]. The estimation of general autocovariance structures and identification of possible parametric models in time

series modelling can be attained by using linear models such as auto-regressive or moving average models [2]. In the area of time series forecasting, the most commonly used model is the auto-regressive integrated moving average (ARIMA) model. A time series can be made stationary by either differencing the data or subtracting the estimated trend and seasonal components from the data [3]. Additionally, the ARIMA model can be regarded as a noise filter which separates a signal from its noise where the signal can then be extrapolated into the future to acquire forecasts.

Data pre-processing is a crucial step for stochastic modelling as it relies greatly on data to approximate the input-output relationships. Most often than not, real-world data are usually found incomplete, filled with noises or inconsistent. A data can be considered as incomplete when they lack certain attributes [4]. Meanwhile, a data is categorised as noisy when it contains outliers or errors, and a data is said to be inconsistent if they contain discrepancies [4]. Undeniably, having poor quality data can disrupt the results of data analyses. Since the quality of the data serves as the basis for quality decision, if quality data are non-existent, there may be no quality data mining results [4] and consequently no reliable database or data warehouses.

Singular spectrum analysis (SSA) is among the notable and well-established techniques in time series analysis. SSA is applicable to various areas for solving countless practical problems such as economics, mathematics, and physics, as it consists of a combination of elements from classical time series, dynamical systems, multivariate geometry, multivariate statistics, and signal processing [5]. Additionally, complex time series with prominent structure could also be solved through the application of SSA [6]. In terms of SSA as a data pre-processor, the efficiency of SSA in reconstructing the time series after identifying, extracting and eliminating its noise components [8] has been proven to significantly improve the performance of a forecasting model [7].

2. Fuzzy C-Means Clustering

Let X be a set of N data objects represented by p-dimensional feature vectors $X_i = (x_i, x_{2i}, \dots, x_{pi}) \in R^p$ where $X = \{X_1, X_2, \dots, X_N\}$. A set of N feature vectors is then represented as a $p \times N$ data matrix. A fuzzy clustering algorithm partitions the data X into C fuzzy clusters, forming a fuzzy partition in X. A fuzzy partition can be conveniently represented as a matrix U, whose elements $u_{ji} \in [0,1]$ represent the membership degree of X in cluster j. Hence, the jth row of U contains values of the jth membership function in the fuzzy partition. The FCM algorithm was introduced by [10], which is based on minimization of the following objective function, with respect to U, a fuzzy C-partition of the data set, and to V, a set of C prototypes,

$$J(X; U, V) = \sum_{j=1}^{C} \sum_{i=1}^{N} u_{ji}^{m} d^2(X_i, V_j) \quad (1)$$

where C and N satisfies $2 \leq C < N$, $m \in (1, \infty)$ is a fuzzy index which determines the fuzziness of the clusters, $d^2(X_i, V_j)$ is any inner product metric (distance between X_i and V_j), and $V = (V_1, V_2, \dots, V_C), V_j \in R^p$ is a C-tuple of cluster prototypes which have to be determined. In order to avoid the trivial solution, U must satisfy the constraints $\sum_{j=1}^{C} u_{ji} = 1, \forall_i$ and $0 < \sum_{i=1}^{N} u_{ji} < N, \forall_j$. Fuzzy clustering is carried out through an iterative optimization of equation (1) according to [11].

3. Singular Spectrum Analysis

The main idea in singular spectrum analysis (SSA) is decomposing an original time series into a number of reconstructed series characterizing the trend, periodic or noise components of the series [9]. There are generally four stages in the process of SSA, that is embedding, singular value decomposition (SVD), grouping and diagonal averaging. These stages can also be combined to form two phases which are decomposition and reconstruction, such that the first two stages are in the decomposition phase while the latter two are in the reconstruction phase. A brief SSA procedure is provided in figure 1 below, mainly following the discussion by [6].

Stage 1: Embedding	1. Let $Y_T = (y_1, \dots, y_t)$ be a one-dimensional time series. 2. Transfer Y_T into the multi-dimensional series $X_1, \dots, X_K$ with $X_i = (y_i, \dots, y_{i+L-1})' \in R^L$, where $K = T - L + 1$ and $2 \le L \le T$. 3. A Hankel trajectory matrix $X = [X_1, \dots, X_K] = (x_{i,j})_{i,j=1}^{L,K}$ is then produced, such that all its diagonal elements are equal.
Stage 2: Singular value decomposition	4. Let (U_i, U_j) be the inner product of vectors U_i and U_j. 5. Let $\|U_i\|$ be the norm of U_i. 6. Order the eigenvalues of matrix XX' in decreasingly by $\lambda_1, \dots, \lambda_L$. 7. Order the orthonormal system of eigenvectors corresponding to the eigenvalues of matrix XX' in decreasingly by $U_1, \dots, U_L$, that is, $(U_i, U_j) = 0$ for $i \neq j$ and $\|U_i\| = 1$. 8. Let X_i be the elementary matrices. 9. Let U_i be the left eigenvectors of the trajectory matrix. 10. Let V_i be the right eigenvectors of the trajectory matrix. 11. Write $d = \max(i, such\ that\ \lambda_i > 0) = rank\ X$. 12. Write $V_i = X' U_i (\sqrt{\lambda_i})^{-1}$. 13. Write the SVD of the trajectory matrix $X = X_1 + \cdots + X_d$, such that $X_i = \sqrt{\lambda_i} U_i V_i'$ $(i = 1, \dots, d)$. 14. The i-th eigentriple for matrix X is then $(\sqrt{\lambda_i}, U_i, V_i)$, its singular values are $\sqrt{\lambda_i}$ $(i = 1, \dots, d)$ and its matrix spectrum is the set $\{\sqrt{\lambda_i}\}$.
Stage 3: Grouping	15. Divide the elementary matrices X_i into groups and sum them within the group. 16. Write $X_I = X_{i_1} + \cdots + X_{i_p}$, such that $I = \{i_1, \dots, i_p\}$. 17. Select the eigentriple grouping $X = X_{I_1} + \cdots + X_{I_m}$, in which the indices $K = 1, \dots, m$ are separated into disjoint subsets $I_1, \dots, I_m$.
Step 4: Diagonal averaging	18. Transfer every matrix I into an additive component of the original series Y_T. 19. Diagonal average or Hankelise matrix Z by averaging all z_{ij} of Z to get the k-th term of the resulting series, wherein $i + j = k + 2$, resulting in $\mathcal{H}Z$. 20. Diagonal average all $X = X_{I_1} + \cdots + X_{I_m}$ to get $X = \tilde{X}_{I_1} + \cdots + \tilde{X}_{I_m}$, in which $\tilde{X}_{I_1} = \mathcal{H}X$.

Figure 1. Procedures for each stage of singular spectrum analysis.

4. Data

This research utilized the monthly streamflow records available from the Department of Drainage and Irrigation Malaysia, specifically the Sungai Muda (Jeniang) and Sungai Muda (Syed Omar) data sets. Both records have varying total observations wherein there are 600 for Sungai Muda (Jeniang) while the Sungai Muda (Syed Omar) has 432 observations. This difference is due to the time lengths of the records in which the Sungai Muda (Jeniang) data set is from 1960 to 2009 (50 years), whereas the Sungai Muda (Syed Omar) data set is only for 36 years (1974-2009). For each research data, 90% was separated for model training while the remaining 10% was for model testing, in which 540 observations constitute the training set of Sungai Muda (Jeniang) and the other 60 is for testing, while Sungai Muda (Syed Omar) was divided into 389 and 43 observations as its training and testing sets. Figure 2 displays the plots of monthly streamflow for Sungai Muda (Jeniang) and Sungai Muda (Syed Omar).

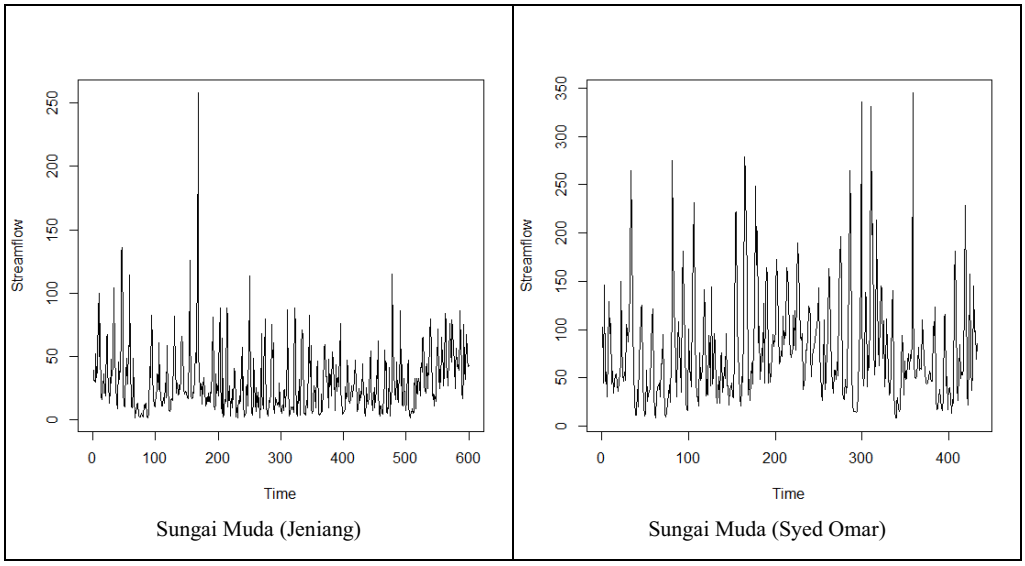

Figure 2. Graphs of streamflow per month for the respective rivers.

5. Results

The SSA technique was performed in this research as a preprocessor for the data sets before they are being supplied to the ARIMA model for forecasting. The FCM clustering method was also utilized to make groupings of the eigenvector pairs during the implementation of SSA besides simply using the SSA technique. The models developed from these processes, i.e. the SSA-ARIMA and FCM-SSA-ARIMA, were compared with the basic ARIMA model to assess their performance. These models were built using appropriate packages available from the CRAN repository in the R software environment. The forecasting errors were calculated by computing the root mean squared error (RMSE) and mean absolute error (MAE), as defined respectively by the equation (2) and (3) below. In the equations, the term n represents the number of data, a is the actual output and p is the predicted output.

$$\text{RMSE} = (n^{-1}\sum_{i=1}^{n}(a_i - p_i)^2)^{1/2} \quad (2)$$

$$\text{MAE} = n^{-1}\sum_{i=1}^{n}|a_i - p_i| \quad (3)$$

To simplify the process of building the ARIMA model from every training set, the auto.arima() function was employed through the 'forecast' package. Subsequently, these models were input with the testing sets into the Arima() function to produce the streamflow forecasts of each data set. The performance of the ARIMA models were also automatically generated with the forecast results. The process of constructing the SSA-ARIMA and FCM-SSA-ARIMA models differ with building a simple ARIMA model only by the implementation of SSA and FCM-SSA before developing the ARIMA model itself. In turn, FCM-SSA is only an extension of the basic SSA by which FCM clustering was used to group suitable eigenvector pairs during execution of the SSA technique. The SSA technique was performed through the usage of relevant functions available in the 'Rssa' package. The decomposition stage of SSA was accomplishable by ssa() whereas reconstruct() was for reconstructing the data series. The other parameter to be supplied to ssa() apart from a data series is the window length L, which for this research was 12, obtained through the formula $(n/2) - p$. The window length L should be less than half of the number of data n, and divisible by the series period p as stated by [12]. The FCM clustering of eigenvectors was done using fcm() from the 'ppclust' package, which was executed between the ssa() and reconstruct() processes. The w-correlation matrix of the eigenvector pairs were yielded from the ssa() function and then became the input for fcm() to generate a list of eigenvector clusters to be used by the reconstruct() function together with the original data series for reconstructing the data into a new

series. Each of the newly reconstructed series were after that modelled and forecasted with ARIMA and the model performance computed. Finally, summation of all models' performances per specific data set was done to get a single RMSE and MAE value.

The performances of ARIMA, SSA-ARIMA and FCM-SSA-ARIMA were measured for comparison to each other. For Sungai Muda (Jeniang), the RMSE of its ARIMA is 18.48, SSA-ARIMA is 15.32 and FCM-SSA-ARIMA is 12.31. The MAE calculated for ARIMA, SSA-ARIMA and FCM-SSA-ARIMA after forecasting Sungai Muda (Jeniang) are 14.40, 12.66 and 9.91 respectively. The ARIMA model gave an RMSE value of 102.08 and SSA-ARIMA returned 42.68, while FCM-SSA-ARIMA yielded 29.51 when applying Sungai Muda (Syed Omar). The MAE value of ARIMA, SSA-ARIMA and FCM-SSA-ARIMA for Sungai Muda (Syed Omar) were discovered to be 83.34, 32.39 and 21.40 respectively. The values calculated for each model can be viewed in table 1.

Table 1. Streamflow forecasting errors computed for individual models.

		ARIMA	SSA-ARIMA	FCM-SSA-ARIMA
Sungai Muda (Jeniang)	RMSE	18.48	15.32	12.31
	MAE	14.40	12.66	9.91
Sungai Muda (Syed Omar)	RMSE	102.08	42.68	29.51
	MAE	83.34	32.39	21.40

6. Conclusions

Streamflow forecasts enable administrators to better manage water resources. Apart from using other statistical tools, forecasting streamflow could aid in the decision-making process by these administrators regarding irrigation strategies, drought control and flood preparations, among others. This research was able to demonstrate that SSA is suitable to be used as a data pre-processing technique that can improve the performance of ARIMA. Besides that, FCM was also successfully combined with SSA and implemented as another data-preprocessor to be compared with SSA. It was shown that FCM-SSA can further improve the forecasting ability of ARIMA.

References

[1] Shumway R H and Stoffer D S 2010 *Time Series Analysis and Its Applications with R Examples* (New York: Springer Science+Business Media) section 1.1 pp 1-2
[2] Breidt J 2005 *J. Am. Stat. Assoc.* **100** 348-9
[3] Brockwell P J and Davis R A 2016 A general approach to time series modeling *Introduction to Time Series and Forecasting* (New York: Springer-Verlag) section 1.3.3 pp 14-5
[4] Kamber M, Han J and Pei J 2012 Data preprocessing *Data Mining: Concepts and Techniques* (Amsterdam: Elsevier) chapter 3 pp 83-124
[5] Zhigljavsky A A 2010 *Stat. Interface* **3** 255-8
[6] Golyandina N, Nekrutkin V and Zhigljavsky A A 2001 *Analysis of Time Series Structure: SSA and Related Techniques* (Chapman and Hall / CRC)
[7] Gao Y, Qu C and Zhang K 2016 *Energies* **9** 757-85
[8] Hassani H and Mahmoudvand R 2013 *Int. J. Energy Stat.* **1** 55-83
[9] Hyndman R J and Athanasopoulos G 2014 Stationarity and differencing *Forecasting: Principles and Practice* (Melbourne: OTexts) section 8/1 pp 213-21
[10] Bezdek J C 1981 *Pattern Recognition with Fuzzy Objective Function Algorithms* (New York: Springer Science+Business Media)
[11] Gath I and Geva A B 1989 *IEEE Pattern Anal.* **11** 773-780
[12] Golyandina N, Korobeynikov A, Shlemov A and Usevich K 2013 *Preprint* arXiv:1309.5050