

Missing Data Prediction and Forecasting for Water Quantity Data

Prakhar Gupta¹ and R.Srinivasan²⁺

¹ Department of Chemical Engineering & Technology, Institute of Technology, BHU, Varanasi, India

² Nalco Technology Center, Pune, India

Abstract. In industrial applications, especially in water treatment plants, it is necessary to obtain flow data (quantity and quality) for a system over a broad range of time. In most cases it is not possible to obtain the parameter of interest in a closed form. Generation of the data over the time period within the known range is not possible or may be extremely time-consuming. Numerous methods are available for interpolation or extrapolation to determine the unknown data or missing data within or outside a range of known data points. In this paper two developed methodologies were analyzed, one to predict the missing values within a data range and another to forecast the seasonal data outside the data range for a raw water quantity data. A proven method like Two-Directional Exponential Smoothing (TES) is applied for predicting the missing values for a raw stream flow data and a seasonal data was forecasted using Exponentially Weighted Moving Average (EWMA) by using the known data values of previous two seasons. Both the methods predicted the data within and outside the period range of water quantity data with good results.

Keywords: Missing Data, Two-Directional Exponential Smoothing (TES), Exponentially Weighted Moving Average (EWMA), Forecasting.

1. Introduction

Predicting the missing values within a time period and forecasting data for future periods is of common interest for many industrial applications. Replacing missing data with time series within the range of known data is crucial for more accurate design proposals and performance evaluation. Especially, in water treatment plants it is required to replace the missing values of the water quality and quantity data to gain knowledge in the system and also to manage the water resources effectively. The water quantity and quality data are defined as time series variables which are recorded at successive time intervals. To use existing operational data as an input to a process simulation model, the missing data should be replaced. To provide a more accurate design proposal and system performance, a reasonable and reliable prediction of missing data is highly needed to determine the correct variability of water treatment plant data. Most common methods available for predicting the missing values in time series are replacing all missing values for a given variable with mean, median or other location statistics [1], SAS program using a time-series model to predict missing values [2,3], Average Nearest Observation (ANO)[4], and Two-Directional Exponential Smoothing (TES) [5].

Also, forecasting and predicting the future data can be very helpful in preparing in advance for any unexpected values. The forecast of seasonal high or low output is crucial in taking care of the installed capacity systems and therefore trigger alarms as needed according to the values. Popular methods like Auto Regressive Integrated Moving Average (ARIMA) [6], Exponentially Weighted Moving Average Method (EWMA) [7], Thomas-Fiering method [8] are also used for forecasting data from a known data set. Comprehensive methods for forecasting based on the exponentially weighted moving average are available [7] for series with trend, non-seasonal, and seasonal series.

⁺ Corresponding author. Tel.: +91 20 39394089; Fax: +91 20 39394381
E-mail address: sramanathan@nalco.com

In this paper a proven method like Two-Directional Exponential Smoothing (TES) is applied for predicting the missing values with time series in a data set. A sample data was used to test the applicability of the method by intentionally deleting some of the known values. Also, the raw sample data set was extended to forecast the values using Exponentially Weighted Moving Average (EWMA).

2. Methodology & Results

The methodology for predicting the unknown values within and outside a known range by applying established methods like Two-Directional Exponential Smoothing (TES) and Exponentially Weighted Moving Average (EWMA) respectively are discussed in the following sections. For application of these methods, sample raw data of stream flow over a period of 5 years was used [9]. Here, only the stream flow rate data was considered for prediction and forecasting, assuming that the water flow data was used by an industrial plant nearby for water management. A test data set was created by removing some data points for predicting the stream flow within the known range and the sample test data set for two known seasons was used for predicting outside the range.

2.1 Missing Data Prediction

For predicting the missing data of stream flow, Two-Directional exponential smoothing (TES) [5] was applied. Test data of stream flow rate by removing some data points over a period of five year is shown in Figure 1.

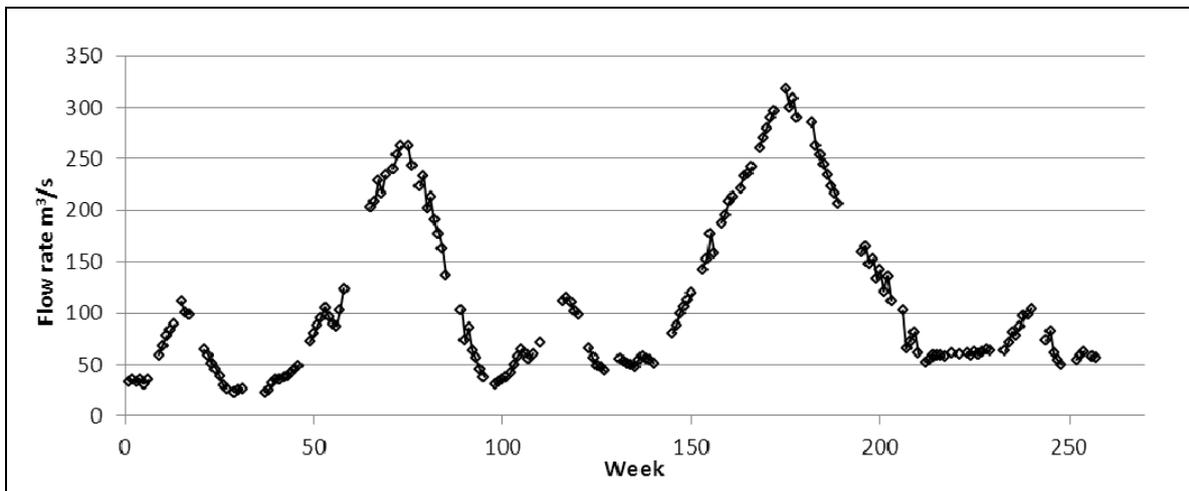


Figure 1. Raw data of stream flow rate [9] with missing values for a 5 year period

TES method was developed to replace missing data. TES method depends on a suitable Exponential Smoothing (ES) method and was developed by using Holt's linear trend algorithm method. The TES method estimates missing data points based on the autocorrelations of the time series to account for the fact that the missing values occur at non-random times. The TES method is designed to represent both forward and backward autocorrelations in the time series, can decrease the difference above caused by different directions. The first step in TES method is to generate the full data set of data using Average nearest observation (ANO) method [4]. The ANO method will replace the missing values with the average of the nearest previous and the following observation i.e., the values are estimated by a weighted average of the nearest observations with higher weight given to the closer observation. Once the data set is generated using the ANO method, the missing values are predicted using a suitable Exponential Smoothing method, Holt's linear trend method, in the forward and reverse direction. An ES method could generate different values depending on the direction of the time series. The Holt's Linear Trend algorithm can be represented as

$$a_t = \delta Y_t + (1 - \delta)(a_{t-1} + b_{t-1}) \quad (1)$$

$$b_t = \gamma(a_t - a_{t-1}) + (1 - \gamma)b_{t-1} \quad (2)$$

where Y_t is the actual value at time t , a_t and a_{t-1} are intercepts (smoothed levels) at time t and $t-1$ respectively, b_t and b_{t-1} are the slopes (smoothed trends) at time t and $t-1$ respectively, δ and γ are smoothing constants that are between 0 and 1 [1,10].

The smoothing constants, δ is used to smooth the new actual and trend-adjusted previously smoothed level and γ is used to smooth or average the trend, which eliminates some of the random error reflected in the unsmoothed trend. The smoothing constants determine the weight given to most recent past observations and therefore control the rate of smoothing or averaging. Values near 1 give weightage to more recent data and near 0 distribute the weights to consider data from the more distant past data.

The averaged forward and backward ES estimates were used for predicting the missing data points. The TES method is a combination time series method and represented for missing values as

$$TES_t = (ES_{forward,t} + ES_{backward,t}) / 2 \tag{3}$$

Figure 2 shows the flow diagram of TES method. This TES method is applied to the test data (Figure 1). Figure 3 shows the raw stream flow rate data (with missing values) and the predicted missing stream flow values (by TES). For estimating the ES forward and ES backward, the constants δ and γ were chosen as 0.7 and 0.9 by trial and error method. The replaced stream flow values determined by the TES method are relatively close to the original values.

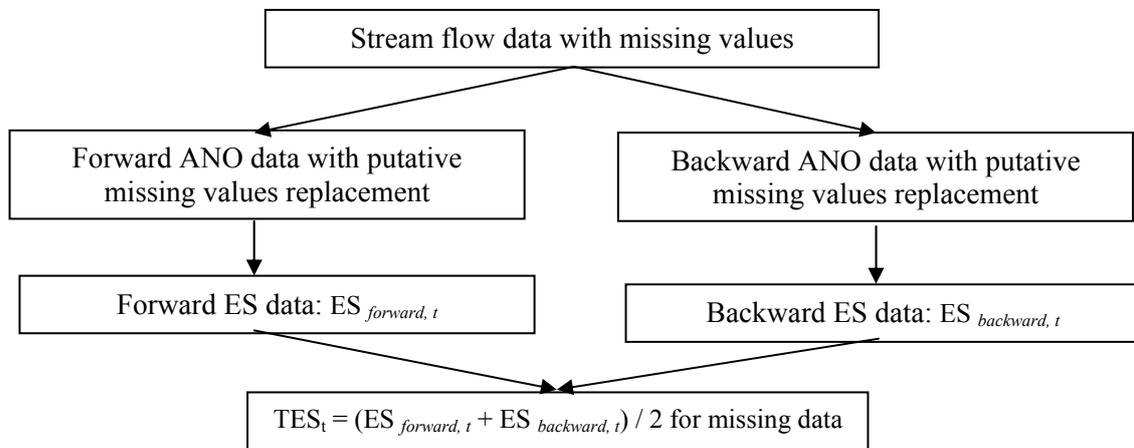


Figure 2. Flow diagram of TES method

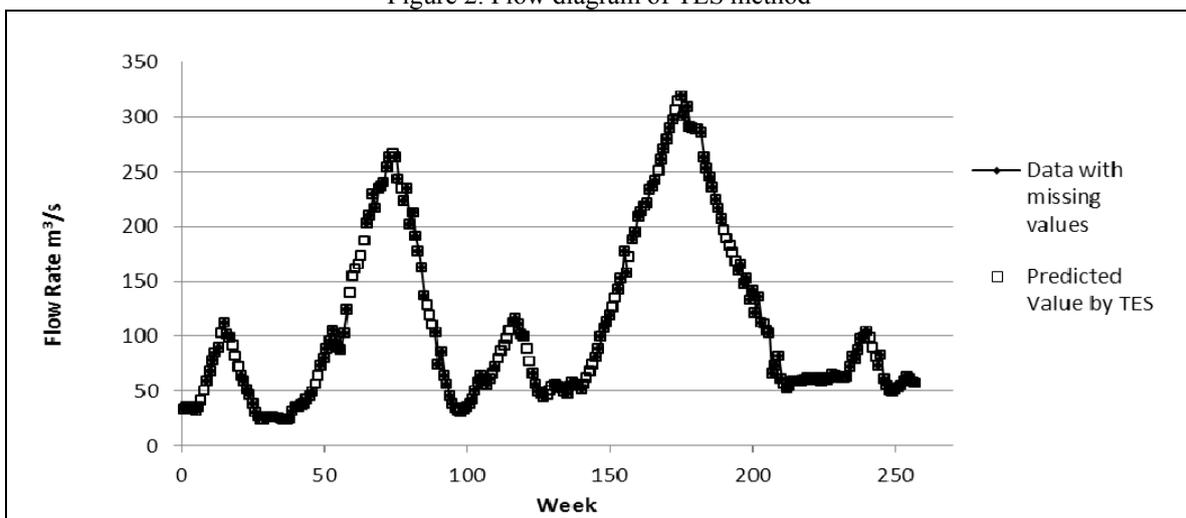


Figure 3. Comparison between the predicted values by TES method and raw data with missing values

2.2 Data Forecasting

The forecasting ratio seasonal method [7] developed for predicting the sales rate was used for predicting the water quantity stream flow data. The forecasting ratio seasonal method can be used to predict the data for next season when the previous seasons share the comparable behaviour. It follows the exponentially weighted moving average method of smoothing the random fluctuations which is extremely easy to compute

with minimum historical data required. The predicted or the forecasted value can be estimated from the following extrapolation equation:

$$ES_{t+T} = \frac{S_t'}{P_{t+T-N}} \quad (4)$$

where S_t' is the smoothed seasonally adjusted rate in period t , P_t is the periodical adjustment ratio for the ' t 'th period, N is the number of the periods in one seasonal and T is the number of forecast periods. The smoothed seasonally adjusted rate in the period t , is represented as

$$S_t' = AP_t S_t + (1-A)S_{t-1}' \quad (5)$$

where S_t is the value in period t and A is a constant that ranges between 0 to 1, which determines how fast the exponential weights decline over the past consecutive periods. The current seasonal adjustment ratio is obtained by combining the current ratio of data and the data with the seasonal adjustment rate from a season ago.

$$P_t = B \frac{S_t'}{S_t} + (1-B)P_{t-N} \quad (6)$$

where the constant B , determines how fast the exponential weights decline over the past season i.e., one period drawn from each season. By solving Equation (5) and (6), the explicit analytic expression for the new seasonal ratio [7] can be given as

$$S_t' = \left[\frac{A(1-B)}{1-AB} \right] P_{t-N} S_t + \left[\frac{1-A}{1-AB} \right] S_{t-1}' \quad (7)$$

$$P_t = \left[\frac{1-B}{1-AB} \right] P_{t-N} + \left[\frac{B(1-A)}{1-AB} \right] \frac{S_{t-1}'}{S_t} \quad (8)$$

The above methodology is used to predict the seasonal data for the third season by using the known values of previous two seasons. The test data published for stream flow rate was considered for seasonal variation of 32 weeks for each season (high and low peaks) out of 257 weeks [9]. The main objective is to predict the third season of 32 weeks data considering the 64 weeks data. Here the sample data was carefully chosen to be in variation with seasonal changes from the available data of 257 weeks. Figure 4 shows the raw test data (two seasons of 64 weeks) and the forecasted data for third season (32 weeks). Here the constants A and B were chosen as 0.1 and 1 by trial and error method. The forecasted data for these constant values seems to be more similar to the immediate previous season i.e., the second season. However, if the constant values are change to 0.2 and 0.1, as shown in Figure 5, the forecasted data is towards the first season. The significance of the constants in the equation determines the data correlations to the immediate or previous season's data.

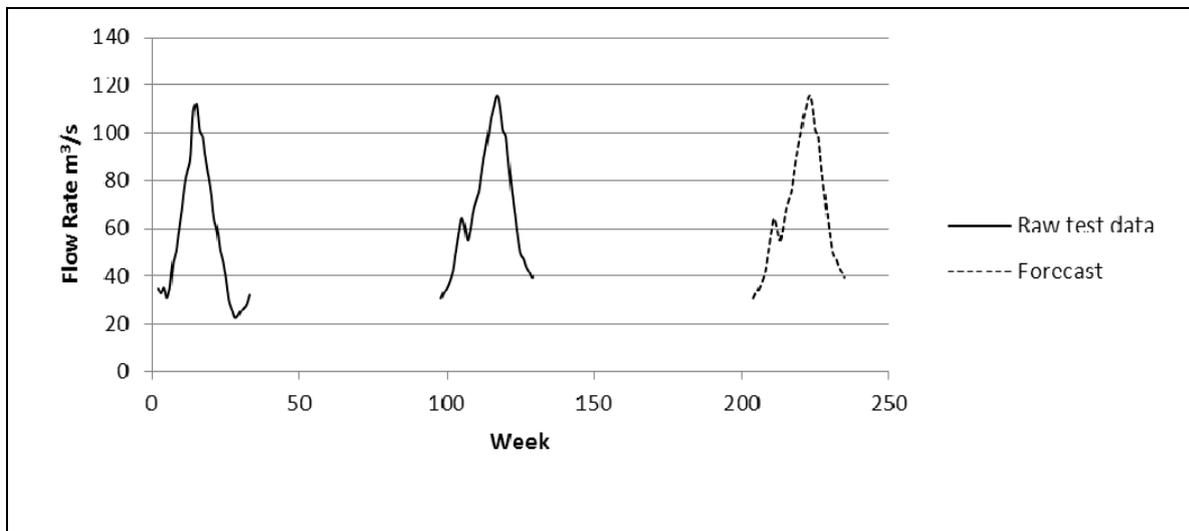


Fig.4. Comparison of forecasted data for one season with the raw test data [9] of two seasons

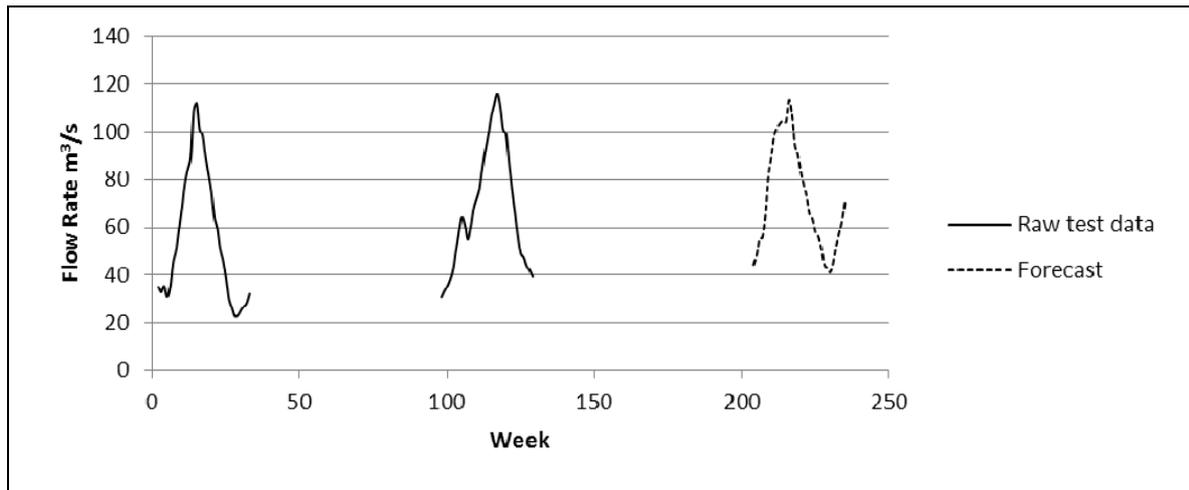


Fig 5. Comparison of forecasted data for one season with the raw test data [9] of two seasons

3. Conclusions

In this paper, a Two-Directional exponential smoothing (TES) method was applied for predicting the missing values with time series and exponentially weighted moving average (EWMA) was applied for forecasting the water stream flow data with seasonal variation. Both the methods predicted the data within and outside the period range with good results.

4. Acknowledgements

This work was carried out at the Nalco Technology Center, Pune, as a part of an industry-university collaborative research internship project. The authors would like to acknowledge Dr. Hari Reddy, Director, Nalco Technology Center, Pune, India and Dr. A.K. Verma, Head, Department of Chemical Engineering & Technology, Institute of Technology, BHU, Varanasi, India for providing their support to carry out this project.

5. References

- [1] DeLurgio, S. A.. *Forecasting Principles and Applications*, McGraw-Hill, New York (1998)
- [2] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, M. Kolehmainen. Methods of imputation of missing values in air quality data sets. *Atmos . Environ.* (2004), 38, 2895-2907.
- [3] T. Scheider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Clim.* 14(5) (2001), 853-871
- [4] J. Huo. Application of statistical methods and process models for the design and analysis of activated sludge wastewater treatment plants. PhD dissertation (2005), The Univ. of Tennessee , Knoxville, Tenn.
- [5] J. Huo, C. D. Cox, W. L. Seaver, R.B. Robinson, Y. Jiang. Application of Two-Directional Time Series Models to Replace Missing Data. *J. of Environmental Engineering*. ASCE (April 2010)
- [6] G.E.P. Box, G.M. Jenkins , *Time Series Analysis Forecasting and Control*. Holden-Day, San Francisco, 1976
- [7] C. C. Holt, Forecasting Seasonals and trends by exponentially weighted moving averages, *International J. of Forecasting* 20 (2004) 5-10
- [8] R.T. Clarke. *Mathematical Models in Hydrology*, FAO of United Nations, Rome, 1984
- [9] A. Kurunc, K. Yurekli, O. Cevik. Performance of two stochastic approaches for forecasting water quality and stream flow from Yeşilirmak River, Turkey. *Environmental Modeling & Software* 20 (2005) 1195 -1200
- [10] SAS Institute Inc. SAS onlineDoc, version 8, SAS Institute, Cary, N.C. (1999)