A blended model for estimating of missing precipitation data (Case study of Tehran - Mehrabad station)

Sohrab Hajjam[1*], Nosratolah Yusefi[2] and Parviz Irannejad[1]

1-Institute of Geophysics, University of Tehran

2- Islamic Azad University, Takestan Branch

(Received 11 Oct. 2005, Accepted 21 Dec. 2006)

Abstract

Meteorological stations usually contain some missing data for different reasons.There are several traditional methods for completing data, among them bivariate and multivariate linear and non-linear correlation analysis, double mass curve, ratio and difference methods, moving average and probability density functions are commonly used.

In this paper a blended model comprising the bivariate exponential distribution and the first-order Markov chain is introduced for estmating of missing precipitation data. In this method, the day having the missing precipitation record is marked as either wet or dry using the first-order Markov chain and randomly generated numbers. If the Markov chain model marks the day as wet, then a bivariate exponential distribution is used for estimating the magnitute of the missing precipitation datum. Application of the model to the precipitation data from Tehran Mehrabad station shows a good correlation between the statistics of the predicted precipitation data with observed ones.

Key words: Tehran, Iran, bivariate exponential distribution, Markov chain, random number.

[*] Corresponding author: Tel: +98 9123813845

Introduction

Meteorological datasets may lack records for some months or years. Destruction of a meteorological station by adverse natural events such as storms and floods, misfunctioning of meteorological instruments for some time and missing to record the data by the observers are among the possible reasons for missing data. In such conditions there is a need to complete the data before any analysis could be performed. Many different traditional methods such as the double-mass curve, cummulative curve, ratio to neighbouring stations, running averages, gradual trend in means, linear and non-linear single and multi-variable correlation analysis have been introduced for estimating of missing data.

Karl and Knight (1998) and Brunetti et al. (2001) fitted the Gamma Distribution Function (GDF) to daily precipitation data for each month to estimate the parameters of the GDF and use the best fit for calculating the missing values. Based on historical precipitation datasets with missing data from a station, Todorovic and Woolheiser (1975), Katz (1977) and Woolheiser (1992) used a first order Markov chain and randomly generated numbers to define the wet days and the dry days. Woolheiser and Pegram (1979), Wilson et al. (1992) and Hanson et al. (1994) applied the bivariate exponential distribution to fill the missing precipitation data.

In the present study, we introduce a blended model based on the bivariate exponential distribution and the Markov chain with randomly generated numbers for estimating of missing data and evaluate its results for the Tehran Mehrabad station.

Materials and Methods

Bivariate Exponential Distribution and Markov Chains

In this section, the bivariate exponential distribution and the Markov Chain are described. By definng a wet day as a day during which precipitation occurs, we will have:

$X_t(k) = |0$, if day t is dry at position k

$|1$, if day t is wet at position k (1)

where k the position of the station and t is time. The magnitude of precipitation during day t at the station k is:

$$Y_t(k) = r_t(k) \cdot X_t(k). \quad (2)$$

Combining Equations (1) and (2) results in:

$$Y_t(k) = \begin{cases} 0 & , \quad X_t(k) = 0 \\ r_t(k) & , \quad X_t(k) \neq 0 \end{cases} \quad (3)$$

The conditional probability in a first-order Markov chain is:

$$\begin{aligned} \Pr\{X_t(k) = 1 | X_{t-1}(k) = 0\} &= P_{01}(k) \\ \Pr\{X_t(k) = 1 | X_{t-1}(k) = 1\} &= P_{11}(k) \end{aligned} \quad (4)$$

For a Markov chain with limited number of possible states, we can produce the square P matrix, which in general equals the transfer probablity matrix:

$$P=\begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1j} & \cdots & P_{1m} \\ P_{21} & P_{22} & \cdots & P_{2j} & \cdots & P_{2m} \\ \vdots & \vdots & & \vdots & & \vdots \\ P_{i1} & P_{i2} & \cdots & P_{ij} & \cdots & P_{im} \\ \vdots & \vdots & & \vdots & & \vdots \\ P_{m1} & P_{m2} & \cdots & P_{mj} & \cdots & P_{mm} \end{bmatrix} \quad (5)$$

where m is the number of elements of the state vector.

In the first-order bi-conditional Markov chain model, we can define symbols P_{01}, P_{10}, P_{00} and P_{11} as:

$P_{01}(k)$: the probability of the occurance of a wet day following a dry day,

$P_{10}(k)$: the probability of the occurance of a dry day following a wet day,

$P_{00}(k)$: the probability of the occurance of a dry day following a dry day, and

$P_{11}(k)$: the probability of the occurance of a wet day following a wet day.

The above-mentioned definitions lead to:

$$P_{10}(k)=1-P_{11}(k), P_{00}(k)=1-P_{01}(k) \quad (6)$$

The probabilities are calculated as:

$$P_{11}=\frac{n_{11}}{n_{11}+n_{10}} \qquad P_{01}=\frac{n_{01}}{n_{01}+n_{00}} \quad (7)$$

$$P_{10}=\frac{n_{10}}{n_{11}+n_{10}} \qquad P_{00}=\frac{n_{00}}{n_{01}+n_{00}}$$

where n_{11} is the conditional frequency of a wet day following a wet day, n_{10} is the conditional frequency of dry day following a wet day, n_{00} the conditional frequency of a dry day following a dry day, and n_{01} is the conditional frequency of a wet day following a dry day. Using the calculated values, we can produce the probability transfer function of the Markov chain as:

$$P=\begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix} \quad (8)$$

The critical probability, $P_c(k)$, is then defined as:

$$P_c(k)=\begin{cases} P_{01}(k) & , \quad X_{t-1}(k)=0 \\ P_{11}(k) & , \quad X_{t-1}(k)=1 \end{cases} \quad (9)$$

To decide whether day t is with or without precipitation, we produce a random number, $U_t(k)$, between [0, 1] and determine $X_t(k)$:

$$X_t(k)=\begin{cases} 1 & if \quad U_t(k)\le (P_c(k)) \\ 0 & otherwise \end{cases} \quad (10)$$

If $X_t(k) = 0$, then the day is dry and $Y_t(k) = 0$. If $X_t(k) = 1$, then the day is wet and the bivariate exponential distribution is used to estimate the magnitude of precipitation as follows.

The median value of the daily precipitation records of the wet days (M_d) is determined and the mean values of the upper 50% (β_1) and the lower 50% (β_2) are calculated. The probability density function is:

$$f(r(k))=\frac{\alpha(k)}{\beta_1(k)}\exp\left(\frac{-r(k)}{\beta_1(k)}\right)+\frac{1-\alpha(k)}{\beta_2(k)}\exp\left(\frac{-r(k)}{\beta_2(k)}\right) \quad (11)$$

with:

$\beta_1(k)\ge\beta_2(k)>0$ and $0<\alpha(k)\le 1..$

It gives the cumulative distribution function as:

$$F(r)=\int_0^{M_d}\left(\frac{\alpha(k)}{\beta_1(k)}\exp\left(\frac{-r(k)}{\beta_2(k)}\right)+\frac{1-\alpha(k)}{\beta_2(k)}\exp\left(\frac{-r(k)}{\beta_2(k)}\right)\right)dx \quad (12)$$

where $\alpha(k)$ is the parameter value of the exponential function at k th position and M_d is the median.

We calculate $r_t(k)$ as:

$$r_t(k) = r_{\min} - \beta \ln(\upsilon(k)) \quad (13)$$

where $\upsilon(k)$ is a random number within the interval [0, 1] and $r_{\min}$ is minimum observed precipitation. To define β, we generate another random number, q, in the [0, 1] range and compare it with $\alpha(k)$; $\beta = \beta_1$ if $q < \alpha(k)$ and $\beta = \beta_2$ if $q > \alpha(k)$. Finally, by using $r_t(k)$ from Equation (13) the value of the missing data of the daily precipitaion is estimated from Equation (2).

Results & Disscussion

Application of the Blended Model

Daily precipitation data in April from the Tehran Mehrabad station for the 29-year period of 1958-1986 (inclusive) are used to evaluate the performance of the blended model. The data are used to construct the transfer probability matrix of the Markov chain and to calculate median and values of α, β_1 and β_2.

The April precipitation dataset for the study period is complete with no missing values. We artificially set the values for the five days of April 2-6 of a single year as missing and used the methodology outlined in Section 2 to estimate the values for each day. The procedure was repeated 29 times, each time deleting and estimating data for the five days of one year.

For the total 145 days set as missing, we found 33 days as wet and 112 days as dry, quite close to the observed 35 wet days and 110 dry days. For 95 days, the estimated wet condition was the same as that observed one. For 26 cases, the model defined the day as wet while it was dry and for 24 cases, the model estimated the day as dry while it was actually wet.

The estimated daily precipitation values are compared with those observed in Figure 1. The scatter diagram shows estimated daily precipitation using the blended model for days 2 to 6 of April during the period 1958-1986 versus those observed at the Tehran Mehrabad station are presented in Figure 2. The total precipitation, mean and the standard deviation of the estimated daily precipitation for the second to sixth of April were 124.3 mm, 0.9 mm and 2.7 mm, respectively. The corresponding observed values were 116.2 mm, 0.8 mm and 2.5. The results show that the blended model is quite successful in estimating the statistical characteristics of the missing daily precipitation in Tehran, It is not, however, successful in predicting the wetness condition of exact dates.

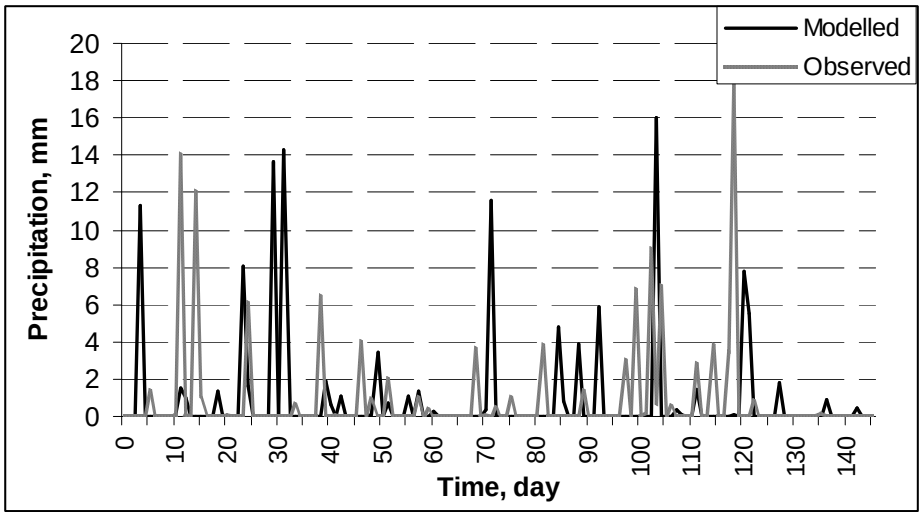

Figure 1: Observed and estimated precipitation using the blended model for days 2 to 6 April in the period 1958-1986 at the Tehran Mehrabad station. The five first days are for 1958 and the last five days are for 1986

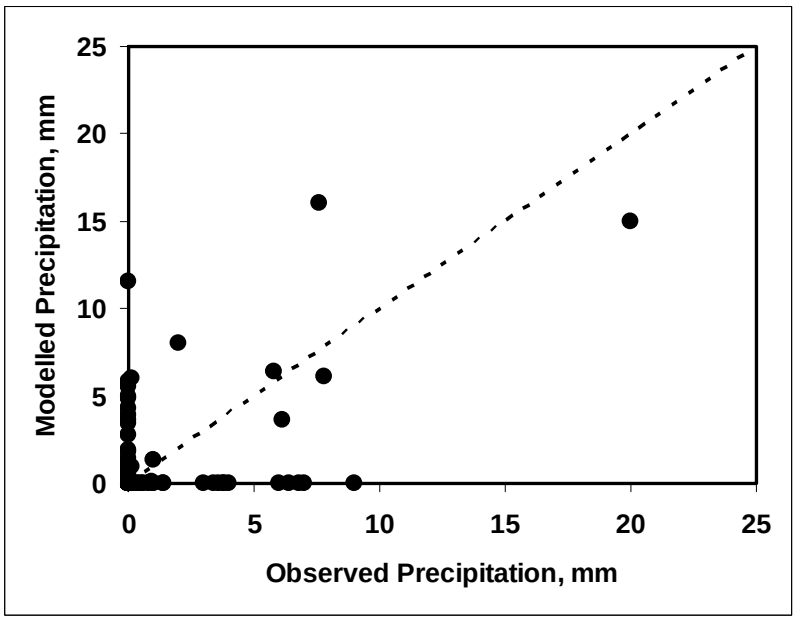

Figure 2: Scatter diagram showing estimated daily precipitation using the blended model for days 2 to 6 of April during the period 1958-1986 versus those observed at the Tehran Mehrabad station. The dashed diagonal is the 1:1 line

Conclusion

Based on the first-order Markov chain and the bivariate exponential distribution, a blended model was introduced for estimating missing precipitation data. Unlike most of other methods of estimating missing data, the blended model does not require information from neighbouring

stations. This implies that the blend model is an appropriate tool for the estimation of missing data in scarce data regions, where a reliable station is too far that the assumption of consistency between neighbouring stations could be fulfilled.

The blended model was used to estimate daily precipitation at the Tehran Mehrabad station during April 2 to April 6, 1958-1987. Observations showed that of the 145 days studied, 110 days were dry and 35 days were wet. The model predicted 112 days as dry, but for only 86 days the dates of the predicted and obsreved dry days matched; on 26 predicted dry days the real condition were wet. The dates of predicted and observed wet days matched on only nine days; the model predicted 24 days as wet, while they were really dry. The mean and the standard deviation of the predicted daily precipitation were, respectively, 0.9 mm and 2.7 mm, very close to 0.8 mm and 2.5 mm observed. Comparing the modelled total number of dry and wet days and the mean and variations of daily precipitation with those observed reveals that the introduced model captures the statistical features of the daily precipitation data very well. However, the model is not a strong tool for predicting the amount of precipitation on a specific day. The model needs to be tested in different climatic conditions and its results must be compared with those from different traditional methods of filling meteorological data gaps before a firm conclusion about its efficiency can be made. This is the subject of a subsequent paper.

References

1- Brunetti, M., Maugeri. T., Nanni. N. (2001). Changes in total precipitation, rainy days and extreme events in northeastern Italy. Int. J. Climatol. 21, 861-871.

2- Karl, T.R., Knight, R. (1998). Secular trends of precipitation amount frequency and intensity in the United states. Bull. Amer. Meteor. Soc. 79, 231-241.

3- Hanson, C.L., Cumming, K.A.., Woolhiser, D.A.., Richardson, C.W. (1994). Microcomputer program for daily weather simulartions in the contiguous United States. USDA/ARS, ARS-114,38pp.

4- Katz, R.W. (1977). Precipitation as a chain- dependent process. J. Appl. Meteorology, 16, 671-676.

5- Todorovic, P., Woolhiser, D.A. (1975). A stochastic model of n-day precipitation. J. Appl. Meteorology, 14, 17-24.

6- Wilson, L.L., Lettenmaier, D.P., Skyllingstad, E. (1992). A hierarchical stochastic model of large-scale atmospheric circulation patterns and multiple station daily precipitation. Journal of Geophysical Research, D3,2791-2809.

7- Woolhiser, D.A. (1992). Modeling daily precipitation-progress and problems. Statistics in the Environmental and Earth Sciences. John Wiley, NewYork, pp. 71-89.

8- Woolhiser, D.A, Pegram, G.S. (1979). Maximum likelihood estimation of Fourier coefficients to describe seasonal variation of parameters in stochastic daily precipitation model. J. App. Meteorology, 18, 34-42.