مجله علمی - پژوهشی
علوم و مهندسی آبخیزداری ایران
Iran-Watershed Management
Science & Engineering

Vol. 1, No. 1, Spring 2007
سال اول-شماره ۱-بهار ۱۳۸۶

Technical Paper

Application of Geostatistics in Time Series: Mashhad Annual Rainfall

Bijan Ghahraman[1] and Firooz Ahmadi[2]

Abstract

There are numerous methods for data filling in hydrology. Most, however, are based on correlations with nearby stations in a general scheme of regionalization. These methods, though robust, fail to function when and where all the near stations are missed-data too. The Mashhad synoptic station has annual rainfall data over a 50 year period from 1951 to 2000 and historic rainfall data from 1893 to 1940, just before World War II. This time series has about 15 years of missed data which can not be filled by usual methods. So, the techniques of geostatistics and kriging were adopted to this long-term time series as an alternative. The data showed a poor correlation at every time lag, showing that, while all the semi-variogram models performed nearly equal, there was a high correlation among each of the others. The results included with polynomial regression fits to different moving average orders, nailed at some reasonable estimates for missed rainfall values.

Keywords: Regionalization, Data filling, Applied statistics, Polynomial regression.

Introduction

The estimation of missed data of climatic or hydrological parameters is one of the applications of statistics in hydrology. The reason we proceed for estimation lies in the fact that delaying a project is not warranted, even under long data-short series. On the other hand, one can not ignore the marked role of data in hydraulic designs. There are plenty of methods for such estimations in hydrology. Texts and common / recent or just literature fully describe these methods.

Today it is quite possible to make an unbiased estimate of any missed parameter through a method called kriging, after D.G. Krige [1], which shows a geostatistical structure. In addition to presenting an estimate over a geographical location, a kriging estimator specify the error of estimate of the parameter at hand, which holds marked superiority over the common estimation methods. Therefore, one may categorize the applications of geostatistics as interpolation, averaging, and network design. In interpolation it is possible to estimate any missed parameter over a regular, or even irregular data points [2]. In averaging, it is possible to average a specific parameter over a region under study, e.g. rainfall over a watershed [3]. In network design, it is possible to find the best locations for monitoring a specific parameter over a region [4,5].

Szentimery [6] has focused on the mathematical background of spatial interpolation methods, especially those of geostatistics, which are currently in use in climatology. More recently, geostatistical space–time models are being used increasingly for addressing environmental problems such as monitoring acid deposition or global warming, and forecasting precipitation or stream flow [7]. In this field

1-Associate Professor of Irrigation, Ferdowsi University of Mashhad, College of Agriculture, Mashhad 91775, Iran, bijangh@ferdowsi.um.ac.ir,
2-Former Graduate Student of Irrigation, Ferdowsi University of Mashhad, College of Agriculture, Mashhad 91775, Iran

of research, Perry and Hollis [8] generated monthly and annual 5km x 5km grided datasets covering the UK for the 1961-2000 period for 36 climatic parameters. Time series approaches may also be generalized to a continuous spatial domain and maps of a specific parameter (e.g. precipitation levels as reported by Johnson et al., [9] may be constructed at any arbitrary location via interpolation of time series model parameters. Kyriakidis et al. [10] presented a framework for stochastic spatio-temporal modeling of daily precipitation as a shorter scale parameter in a hindcast mode. Observed precipitation were modeled levels in space, and time as a joint realization of a collection of space indexed the time series, one for each spatial location.

The above literature supports the use of kriging in space and also in space-time environments. Both of the above techniques fail to operate in the case of a general data-lack in a region. This research has focused on the possibility of a geostatistics application on long term annual rainfall for Mashhad in Iran, where no other regionalization method exists to fill the missed data.

Materials and Methods

Mashhad a city in the center of Khorasan Razavi Province, is located in the northeast part of Iran at a latitude of $36^{o}17^{'}$, longitude of $59^{o}38^{'}$, and an altitude of 946 MSL. The Mashhad synoptic station has annual rainfall data over a 50 year period from 1951 to 2000, prepared from The Islamic Republic of Iran Meteorological Organization (IRIMIO). There is historic rainfall data from 1983 to 1940, just before World War II when the British embassy had priority over the politics of Iran. However, over this period of 107 years, there is a scatter (1894, 1895, 1905, 1918, 1919, and 1929) and a continuous 10 year period of (1941-1950) missed data. Figure 1 shows the Mashhad annual rainfall in Mashhad. The historic data were reported in inches, so the recent data was changed from millimeters to inches in order to have a uniform data set.

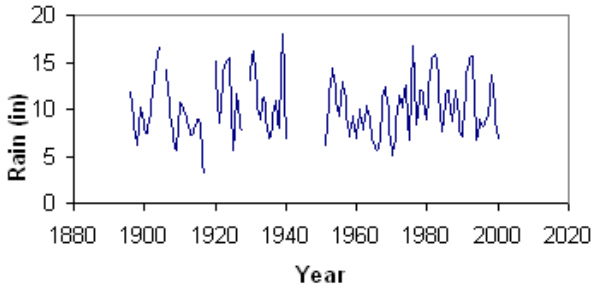

Fig (1) Mashhad annual rainfall time

The normality test of data is performed by Anderson-Darling, Joiner-Ryon, and Kolmogrov-Smirnov tests [11]. Considering the 10-year gap from 1941-1950 in the data, the rainfall time series is divided into two periods, 1893-1941 and 1950-2000. Some standard hypothesis tests were done on these two series. The Equality test of variances [12], parametric tests of the rank-sum test and two-sample test, and the non-parametric test of Kruskal-Wallis [13] were adopted for hypothesis testing for the equality of means. To test whether there is any similarity between the two groups we also traced the linear trend in the form of y=a+b.x (x=number of year- starting from 1- as an independent variable, and y=rainfall as a dependent variable). An ANOVA analysis was carried out to compare the two trend lines corresponding to the two halves [12].

In the kriging system the estimate of a variable value, $Z^{*}(x_o)$, at a specified location x_o and its corresponding variance, $VAR(Z^{*}(x_o))$, (minimum estimation error) is computed as follows:

$$Z^{*}(x_o) = \sum_{i=1}^{n} [\alpha_i Z(x)]_i \qquad (1)$$

$$VAR(Z^{*}(x_o)) = \mu + \sum_{i=1}^{n} [\alpha_i \cdot \Gamma_{io}] \qquad (2)$$

where $Z(x_i)$ is the value of the parameter under investigation at location x_i, and Γ_{io} is the semi-variogram between points at (i) and (o). The optimal weights (α_i) and Lagrangian multiplier (μ) are found by matrix algebra [1]. The semi-variogram $\Gamma(h)$, a measure of spatial dependency, is an essential part of the spatial model and can be computed as:

$$\Gamma(h) = \frac{1}{2} N(h) \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \qquad (3)$$

where h is the lag, N(h) is the number of paired points with lag h, and $Z(x_i+h)$ is the value of the parameter under investigation at location (x_i+h). The sample semi-variogram consists simply of ordered sets of discrete values and are subject to error. However, there are some well-known semi-variogram models of stationary (Gaussian, Exponential, and Spherical models), and intrinsic non-stationary models (Power, and Linear models) to fit the raw values. The Comprehensive definitions of these models are presented elsewhere [1] and are not dealt with here.

The appropriateness of a semi-variogram model can be tested by standardized residuals [1]. Assume that the sample consists of n point measurements of $z(x_i)$. Drop one measurement, $z(x_k)$, then using the other measurements and the assumed variogram, estimate the value z at location x_i and its mean square estimation error, $\hat{z}_k$ and σ_k^2, respectively. The corresponding standardized residual can then be computed from:

$$e(k) = \frac{z(x_k) - \hat{z}_k}{\sigma_k} \qquad k = 1,..., n \qquad (4)$$

The same procedure is repeated for all measurements. Then the following two statistics are determined:

$$S_1 = \frac{1}{n}\sum_{k=1}^{n} e_k \qquad (5)$$

$$S_2 = \frac{1}{n}\sum_{k=1}^{n} e_k^2 \qquad (6)$$

If the model is consistent with the data, the first number must be near zero, while the second must be near one.

Polynomial fit. The similarities between an undulating rainfall time series and a polynomial may tempt a curious person to fit a polynomial regression on such series. On an annual time scale, however, rainfall behaves so erratically that it can hardly follow a smooth polynomial. Moving average is a technique commonly adopted for smoothing such wild series. By this, wet and dry years are found in a more sensible manner. We have managed different moving averages, up to 11 years, for such smoothening. Which order of the polynomial is more fitted to the data is another issue. It is known that lower orders of a polynomial are less flexible for covering optimal humps and depressions. Higher orders of polynomials, on the other hand, try "sharply" to follow such points, which is a serious danger in the case of sparse data series. This may result in abnormally high or low estimates for the missed values. To resolve this, we roughly bounded any prediction between $1.2\times y_{max}$ and $0.8\times y_{min}$, where y_{max} and y_{min} are the maximum and minimum values in the time series, respectively. The maximum order of a polynomial corresponding to these criteria was then selected.

Results and Discussion

1. Statistical considerations

There may be a suspicious thought that the data are not homogeneous before and after the Second World War. Therefore, we divided the data into two main groups of 1893-1940, and 1951-2000. Although the second time series is complete, the first one is not. At this stage we ignored the missed data and considered this series complete. The main statistical characteristics of these 2 series are presented in Table 1. There are no marked differences between these two series. Due to nearly equal means and standard deviations of the series, all data was pooled together and the normality test was conduted [14] over it. Fig. 2 shows an output portrayal for The Ryan-Joiner normality test [11], which supports that the rainfall time series is normal. The other two Anderson-Darling tests, and Darling tests, the Kolmogrov-Smirnov tests were essentially the same, and therefore are not shown here. Based on the normality hypothesis, it followed that at a 5% level of significance, there is no reason to reject the null hypothesis of $H_0: \sigma_1^2=\sigma_2^2$ against the alternative hypothesis of $H_1: \sigma_1^2\neq\sigma_2^2$. Comparing the means of the two groups by both the rank-sum test as a non-parametric test, and the two-sample t test as a parametric test, [13], could not reject that the two groups are identically distributed with the same means. The Kruskal-Wallis test for multi-groups (k=2) [13] also confirmed the above findings.

Table 2 presents the statistical features of three lines corresponding to different time spans. These statistical features do not illustrate many differences. As statistical tests could not reject the equality of the variances, the equality of the two trend lines was tested [12]. The results are abbreviated in an ANOVA table (Table 3). Based on the results of Table 3, (a) the hypothesis of the equality of the two slopes can not be rejected at α=5% level of significance, (b) there is no reason to reject the equality of intercepts, (c) there is no linear association of time on rainfall (correlation coefficients do not differ from zero).

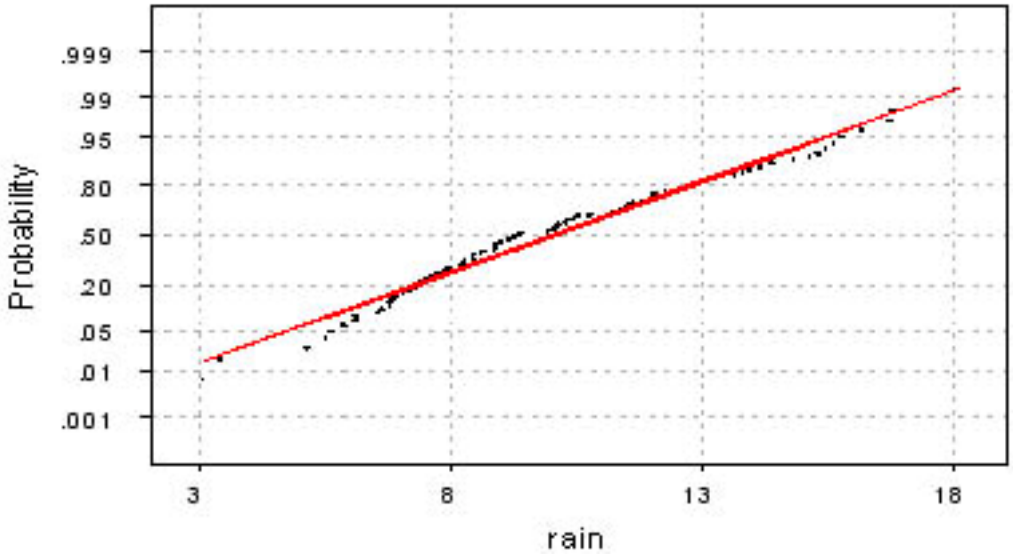

Fig. 2. Ryan-Joiner normality test for Mashhad annual rainfall (in).

Table (1) The main characteristics of the 2 time series of Mashhad rainfall

Series number	N	$\overline{X}$ (in)	SD (in)	min(in)	max (in)	Skewness
1 (1893-1940)	42	10.0017	3.5254	3.12	18.07	0.4040
2 (1951-2000)	50	10.1492	2. 8686	5.15	16.81	0.4261
Total	92	10.0818	3. 1679	3.12	18.07	0.3968
Completed	108	10.1323	2.9506	3.12	18.07	0.3690

Table (2) Statistical features of the 2 trend lines corresponding to two halves of time series*

Parameter	First half (1893-1940)	Second half (1951-2000)	Total span
a (in)	8.8525	9.4188	9.5728
b	0.0452	0.0286	0.0089
r	0.1781	0.1456	0.0912
σ^2 (in^2)	12.3351	8.2223	8.5440
n	42*	50	108

* missed data not included

Table (3) ANOVA table for comparing the 2 trend lines corresponding to 2 halves of time series

Source	Sum of square	Degrees of freedom	Mean square	F-value	
				Measured	Critical
Overall	23.48	1	23.483	2.3269	3.96
Differences in position	0.44	1	0.4429	0.0439	3.96
Differences in slope	1.23	1	1.2266	0.1215	3.96
Residual	888.11	88	10.092		
Total	913.26	91			

2. Estimation methods

a. kriging

Rainfall values did not resemble a meaningful time trend. Therefore, simple point kriging versus universal kriging is a good alternative. Fig. 3 depicts the semi-variogram of the raw data. As the lag proceeds, the semi variogram is computed from fewer paired points. Therefore, an active lag, usually, is taken at most 50% of the total lag. Fig. 3 is prepared for the maximum lag, however. There is no difference between the active and maximum lags on the trends of the semi variogram data points. Different theoretical models are fitted to the semi variogram data of Fig. 3. However, none of the models resulted in a fair fit. The outputs of some of the common models are plotted in Fig. 3 for a rapid comparison with the scatters of raw data. Based on this figure, one may conclude that not only is the nugget variance very high, but also there is a weak dependency of data to each other at any time lag. This weak dependency causes all theoretical semi-variogram models to appear nearly the same. Therefore, outputs of all of the semi-variogram models showed high correlations with each other (Table 4). As a result, we used an average for the kriged values corresponding to every piece of missed data amongst the 5 models. A comparison of Mashhad annual rainfall missed-value estimations by different methods is made in Table 5.

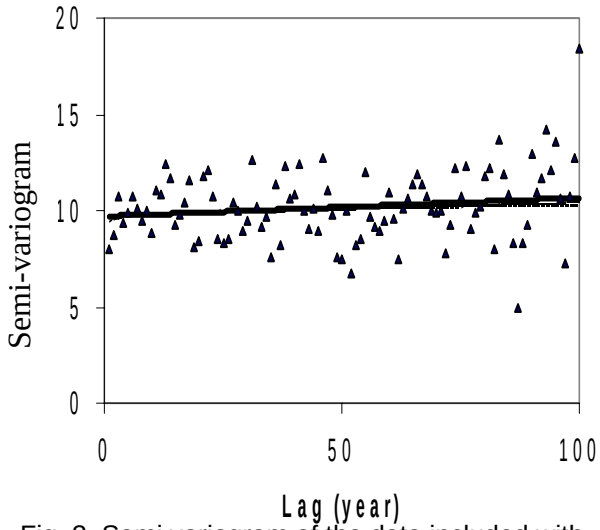

Fig. 3. Semi variogram of the data included with fitted theoretical models (triangle: raw data, thick line: linear model, dashed line: power model, and dashed-dotted line: exponential model)

Table (4) Correlation of different series of estimated Mashhad annual rainfall from some common semi-variogram models

Semi-variogram model	Spherical	Exponential	Gaussian	Linear	Power
Spherical	1	0.9945	0.9938	0.9969	0.9962
Exponential	0.9945	1	0.9767	0.9831	0.9816
Gaussian	0.9938	0.9767	1	0.9994	0.9997
Linear	0.9969	0.9831	0.9994	1	1
Power	0.9962	0.9816	0.9997	1	1

Table 5. Comparison of Mashhad annual rainfall missed-value estimations (inch) by different methods

Year	Method Linear trend	Krging	Degree of moving average for Polynomial fit+						Time series*
			1	3	5	7	9	11	
1894	9.59	8.11	8.38	15.05	10.36	18.83	12.07	3.41	10.07
1895	9.60	8.11	12.05	13.22	8.68	10.81	16.17	8.52	12.88
1905	9.69	12.07	13.43	7.34	7.35	10.62	5.09	8.52	14.09
1918	9.80	8.83	9.82	12.51	19.36	3.39	11.99	18.25	10.07
1919	9.81	9.79	10.41	10.32	8.00	3.67	11.16	9.34	9.26
1929	9.90	11.23	10.64	17.13	5.40	7.40	5.28	15.72	8.05
1941	10.01	10.39	9.46	17.72	10.96	8.93	3.23	13.51	10.16
1942	10.02	10.39	8.93	6.08	10.56	12.94	7.51	15.68	10.16
1943	10.03	10.39	8.41	10.86	7.84	11.18	11.97	10.63	10.16
1944	10.04	10.83	7.95	18.21	17.68	5.52	15.86	8.16	10.16
1945	10.04	10.05	7.62	6.47	6.37	10.25	12.98	10.37	10.16
1946	10.05	11.27	7.46	11.13	10.47	14.47	10.68	6.25	10.16
1947	10.06	9.78	7.51	18.35	10.04	11.97	14.16	7.46	10.16
1948	10.07	10.38	7.76	6.48	7.29	10.08	11.38	9.77	10.16
1949	10.08	10.38	8.20	11.02	17.10	13.98	8.22	7.07	10.16
1950	10.09	10.38	8.76	18.12	5.77	12.03	1.90	16.91	10.16

+ selected polynomial orders for 1, 3, 5, 7, 9, and 11 moving averages are 24, 16, 8, 8, 7, and 5, respectively.

* after Khalili and Bazrafshan [14]

b. Polynomial fit

The best moving average order is not known. While Adopting a low order causes greater variation of rainfall data, a high order for the moving average resulted in losing more data. Fig. 4 is a portrayal showing more sparse time series as the moving average order progress. On the other hand, a low polynomial order causes rigidity in fitting the data, yet a high polynomial order may be responsible for sharp fitting the humps and depressions. While Fig. 5 shows such a sharp fit of the higher order polynomial of 10 to humps and depressions, especially for a 9-year moving average sparse series. The results of polynomial fits are also included in Table 5.

c. Selecting the optimum values

Table 5 portrays some differences amongst the different methods of estimation. Yet the coefficient of variations corresponding to every year of missed-data, on average, is 0.326 (Table 6). The statistical features of the completed series are included in Table 1. The Skewness coefficient of the completed series is lower by around 7%, while the other parameters are in the limit of the raw data.

We compared our results with a time series model (Khalili and Bazrafshan, [14]). Based on the resolution of the figures of this reference, rainfall values corresponding to the data-missed years were determined (Table 5). Our results, on average, differed from those of Khalili and Bazrafshan [14] by 0.38%. However, a constant rainfall value of 10.16" for 10 consecutive years from 1941-1950 seems rather strange. This may be a clue to the mal-functioning of time series modeling for the rainfall time series.

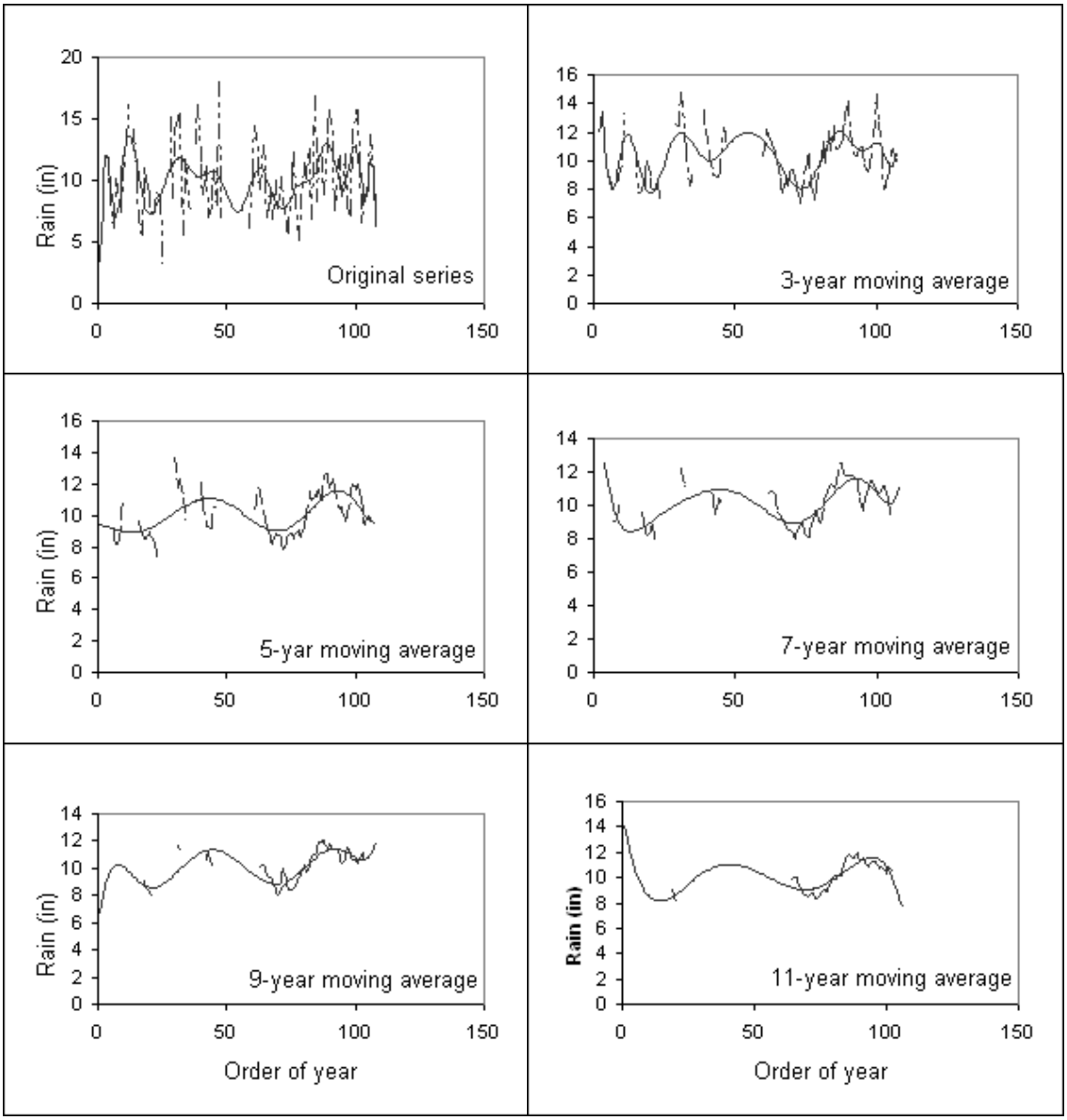

Fig. 4. Measured annual rainfall (in) at different orders (1 for 1886) estimated by polynomial regression (thick line). Thin line is due to actual data.

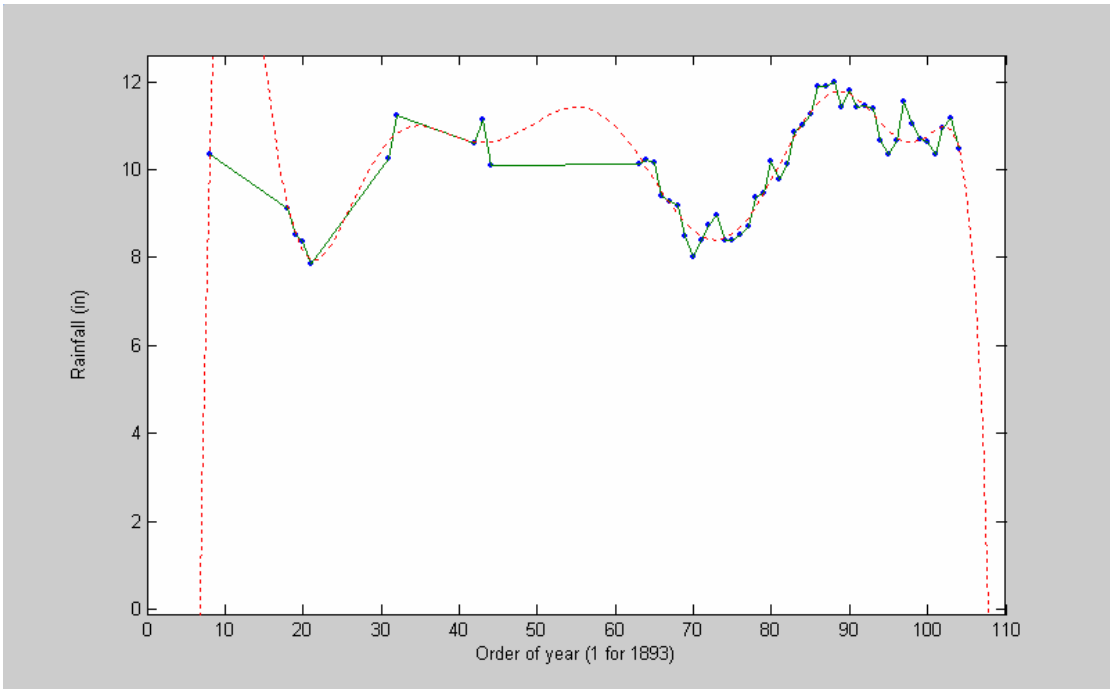

Fig. 5. Danger of high polynomial order to sparse time series (polynomial order 10 for 9-year moving average) (dashed line). Historic data are connected by solid line (ignoring missed data).

Table 6. Simple statistical inferences on different methods of Mashhad annual rainfall missed-value estimations (Table 5).

Year	average (inch)	CV
1894	11.8	0.461
1895	11.10	0.242
1905	8.97	0.302
1918	12.04	0.408
1919	8.43	0.346
1929	10.62	0.394
1941	10.53	0.390
1942	10.26	0.294
1943	10.16	0.137
1944	11.78	0.410
1945	9.27	0.245
1946	10.22	0.244
1947	11.17	0.326
1948	9.15	0.190
1949	10.76	0.310
1950	10.50	0.510
Mean	------	0.326

Conclusion

Having a complete time series is important in design applications. Different methods were utilized to estimate rainfall values for the 16 years of missed values. Geostatistical and kriging techniques were applied to The Mashhad long-term annual rainfall time series. All the semi-variogram models were poor. As a result, the estimated rainfalls were distributed among the mean. Further

investigations are needed to reach a firm conclusion on the applicability of geostatistical methods for data filling purposes. We adopted the polynomial regression fits of different moving average orders and also different polynomial orders. Coefficient variation amongst different methods, on average for 16 years of missed-data, was around 0.3. Also the deviation among our results and a literature-reported time series method, was completely negligible. Therefore, different methods verify each others. As none of the methods are crucial, it may be hypothesized that the results are satisfactory.

References

1-Kitanidis, P.K. 1993. Geostatistics. In: D.R. Maidment (ed.), Handbook of Hydrology. McGraw Hill Book Company, Chapter 20.

2-Kitanidis, P.K. and S. Kuo-Fen. 1996. Geostatistical interpolation of chemical concentration. Advances in Water Resour., 19(6):369-378.

3-Bastin, G., B. Lorent, C. Duque and C.M. Gevers. 1984. Optimal estimation of the average areal rainfall and optimal selection of rain gage locations. Water Resour. Res., 20(4):463-470.

4-Ghahraman, B. and A.R. Sepaskhah. 2001. Autographic rain-gage network design for Iran by Kriging. Iran. J. Sci. Tech., 25(B4):653-660.

5-Ghahraman, B., S.M. Hosseini and H.R. Asgari. 2003. Use of geostatistics in evaluation of groundwater quality monitoring networks. Amirkabir, 14(55H):971-981, (in Persian).

6-Szentimrey, T. 2001. Statistical problems connected with the spatial interpolation of climatological time series. COST719 Meeting, Funchal, Madeira, 17-18 October, Working group 2 (The use of geographical information systems in climatology and meteorology), Paper 02.01.03. URL: www.knmi.nl/samenw/cost719/documents/Szentimrey.pdf

7-Kyriakidis, P.C. and A.G. Journel. 1999. Geostatistical space-time models: A review. Mathematical Geology, 31(6):651-684.

8-Perry, M. and D. Hollis, 2004. The generation of monthly girded datasets for a range of climatic variables over the United Kingdom. Manuscript 18, Version 2.0, 18/03/2004, Group Head, Development Resourcing and Technology, Met. Office, United Kingdom. URL: www.met-office.gov.uk/research/hadleycentre/obsdata/ukcip/monthly_gridding_methods_v2.pdf.

9-Johnson, G.L., C. Daly, G.H. Taylor and C.L. Hanson. 2000. Spatial variability and interpolation of stochastic weather simulation model parameters. J. Applied Meteorol., 39, 778–796.

10-Kyriakidis, P.C., N.L. Miller, and J. Kim. 2004. A spatial time series framework for modeling daily precipitation at regional scales. J. Hydrol., 297(1-4): 236-255.

11-Minitab, 2005. Minitab statistical user guide. Brandon Court, United Kingdom.

12-Holder, R.L. 1985. Multiple Regression in Hydrology. Institute of Hydrology, Wallingford, England, 147p.

13-Hirsch, R.M., D.R. Hesel, T.A. Cohn and E.J. Gilory. 1993. Statistical analysis of hydrologic data. In: D.R. Maidment (ed.), Handbook of Hydrology. McGraw Hill Book Company, Chapter 17.

14-Khalili, A. and J. Bazrafshan, 2004. A trend analysis of annual, seasonal and monthly precipitation over Iran during the 116 years. Biyaban, 9(1):25-33 (in Persian).

15-Salas, J.D., J.W. Delleur, V. Yevjevich and W.L. Lane, 1980. Applied modeling of hydrologic time series. Water Resources Publications. Littleton, Colorado, USA, 484p.