Environ Monit Assess (2018) 190: 569
https://doi.org/10.1007/s10661-018-6913-y

The effect of simple imputations based on four variants of PCA methods on the quantiles of annual rainfall data

Loucif Benahmed · Larbi Houichi

Received: 9 May 2018 / Accepted: 8 August 2018 / Published online: 4 September 2018

Abstract Hydrology-related studies often require complete datasets. However, missing data is an unavoidable reality. In this regard, the imputed data could fulfill the same role as the observed ones, while they are uncertain and just estimated. The aim of this study is to compare the performance of four simple imputation variants derived from the principal component analysis (PCA) for imputing annual total rainfall series obtained from stations located in northeast Algeria. On the other hand, the study focuses on the effects on quantiles of annual rainfall data due to imputations by the former methods. The four variants are probabilistic PCA, expectation maximization PCA, regularized PCA, and singular value decomposition PCA. Annual rainfall data from 30 stations for the period ranging from 1935 to 2004 (69 years) are used to generate and impute gaps for four different percentages of missing values (PMV), namely, 10, 20, 30, and 40%. Based on some well-known statistical indices, the results show that the regularized PCA and expectation maximization PCA variants perform better than the other imputation methods considered in this study and result in very good to acceptable predicted quantiles, such as the following: correlation coefficient is equal to 0.97 with 10% of percentage of missing values and 0.66 with 40%; the relative error between observed and predicted quantiles is equal to 4.74% with 10% of percentage of missing values and 3.82% with 40%.

Keywords Algeria · Rainfall · Missing data · Simple imputation · PCA methods

L. Benahmed (✉) · L. Houichi
Hydraulics Department, Faculty of Technology, University of Bejaia, Targa Ouzemmour , 06000 Bejaia, Algeria
e-mail: loucifbenahmed@yahoo.fr

L. Houichi
e-mail: houichilarbi@yahoo.fr

Introduction

Rainfall is the oldest and most commonly recorded climate variable and is a very valuable indicator for studying climate change, water resource management, irrigation scheduling, flood prevention, and the construction of hydraulic structures (Tabari and Talaee 2011; Tabari et al. 2012; Kebede et al. 2014; Nkiaka et al. 2016; Melanie and Maria 2018). In addition, in order to adequately equip these studies, the correct estimation of hydrological events uses frequency analysis to predict the rainfall that corresponds to certain return times T (quantiles) such as floods and low flows (Karlsson et al. 2016).

According to several articles, many hydrological applications rely on knowledge of these events. Unfortunately, rainfall data remain limited in both time and space, which does not always yield reliable estimates (Cantat 2004). These studies should be based on the series free of missing data and heterogeneity (Bigot 2002; Faizah et al. 2016). Since there is no perfectly reliable and continuous dataset, some uncertainty will

remain (Cantat 2004). But for series with gaps, how and what is the reliability of the reconstituted series?

Missing data is a common problem in most areas of scientific research and remains major in Hydrology and Climatology Science. They may result from different human and material sources. These errors are critical because they affect the continuity of precipitation data and ultimately influence the results of hydrological models using precipitation as inputs (Lee and Kang 2015). This problem appears to be more widespread in the developing countries than in the developed countries, particularly Algeria due to various causes such as (i) frequent failures in measuring equipment, (ii) the total closure of some rain stations, and (iii) the gaps on a daily or monthly scale, therefore lead to gaps on an annual scale. Therefore, the evaluation of missing data is an important task for designing hydrological models (Dastorani et al. 2010; Ouarda et al. 2008).

Rubin (1976) defined the missing data according to three failure mechanisms: data missing completely at random (MCAR) when the probability that an instance (case) has a missing value for a variable does not depend on the known values or missing data. Data missing at random (MAR) when the probability of an instance with a missing value for a variable may depend on the known values but not the value of the missing data itself. Data are missing not at random (MNAR) when the probability that an instance has a missing value for a variable may depend on the value of that variable (Little and Rubin 2002).

Missing data may affect the properties of statistical estimators such as means, variances, or percentages, which lead to a loss of power and disastrous misleading conclusions especially for the prediction of extreme events and quantiles (El Methni 2013). A variety of techniques have been proposed to replace missing values with statistical prediction; this process is usually called "imputation of missing data" (Little and Rubin 2002; Audigier et al. 2015).

Various techniques have been used to estimate missing data mainly simple imputation and multiple imputations (Presti et al. 2010; Audigier et al. 2016).

The first solutions provided by researchers to simply manage the problem of missing data were to use simple imputation methods (Audigier et al. 2015). The problem of simple imputation methods is that no distinction is made between the observed data and the imputed data. In 1977, Donald Rubin has proposed idea of multiple imputation technique. The first theoretical work on multiple imputations was subsequently launched in 1987 (Little and Rubin 1987). Since 2005, the scientific community (Van Buuren 2012) has accepted multiple imputation, and the number of publications on this subject grows exponentially. Today, multiple imputation methods are numerous. They differ in particular from the imputation models they use (Sattari et al. 2017). Today, most published articles focus on developing new imputation methods (Brock et al. 2008; Luengo et al. 2012). But, few studies deal with the effect of the rainfall series' imputation methods on the quantiles.

In this study, we have compared and evaluated four different variants of simple imputation based on principal component analysis (PCA): probabilistic PCA (PPCA), expectation maximization PCA (EMPCA), regularized PCA (RPCA), and singular value decomposition PCA (SVDPCA), according to four evaluation criteria: root mean square error (RMSE), mean absolute error (MAE), quadratic error (EQR), and correlation coefficient (CC). The objective here is not to apply a statistical method on an incomplete table but to evaluate the properties of the four simple imputation methods based on the principal components analysis. Therefore, we have focused on the quality and effect of prediction of missing data and quantiles in data processing.

Study area and data

Study area

The study area is the whole northern extent of Algeria, which is approximately between 34° N and 38° N in latitude and between 2° W and 8° E longitude. Spread over 15 watersheds (Fig. 1) characterized by different climates. The northern zone of Algeria is characterized by a Mediterranean climate with a cold and rainy winter and a hot and dry summer. The annual rainfall is on average 436 mm in the west (Tlemcen), 648 mm in the center (Dar el Beida), 512 mm in the east (Constantine), and 1000 mm for the coast (Jijel).

Data

Annual rainfall series from 30 stations were obtained from the National Meteorological Office (NMO) and National Water Resource Agency (NWRH), and a record length of 69 years (1936/1937–2004/2005) was considered. This period is the maximum common time period

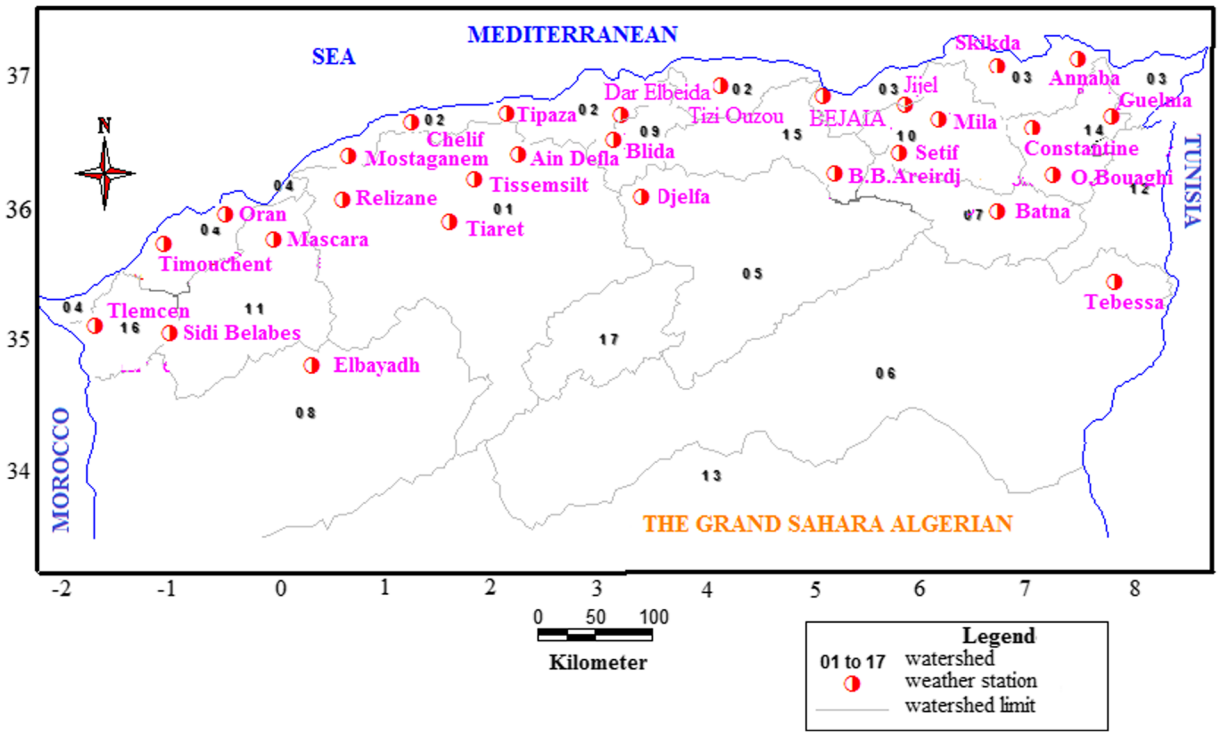

Fig. 1 The stations in the study area

of precipitation data recorded. The information about the stations are presented in Tables 1 and 2. The geographical locations of the stations are shown in Fig. 1.

Methods

The data of 30 rainfall stations for the 1935–2004 periods (69 years) were used to generate and impute deficiencies according to missing completely at random (MCAR) hypothesis using the package missMDA of the free R software (Josse and Husson 2016).

The R software provides a powerful and comprehensive system for analyzing data, used in conjunction with the R-commander (a graphical user interface, commonly known as Rcmdr); it also provides one that is easy and intuitive to use (Suzuki and Shimodaira 2006).

Table 1 Ranges of variables considered in study

Variable	Abbreviation	Range
Average annual rainfall (mm)	AvAR	224.2 to 1013.8
Standard deviation of annual rainfall (mm)	SdAR	78.6 to 275.5
Maximum annual rainfall (mm)	MaxAR	459.4 to 1696.8
Minimum annual rainfall (mm)	MinAR	49.7 to 499.4
Latitude	Lat	33.58° N to 36.88° N
Longitude	Lon	2.61° W to 8.13° E
Altitude (m)	Alt	1.76 to 1347

Gap generation and principle of the analysis

First of all, gaps were generated with the Library "miss Forest" (Stekhoven and Bühlmann 2011) ProdNA algorithm missing completely at random "MCAR" at different percentages 10, 20, 30, and 40% from observed data, so-called reference data. From the original datasets (without missing values), we introduced in the data a varying percentage of missing values (from 10 to 40%) generated MCAR assumption. These simulated missing values were imputed using four methods and four evaluation criteria: RMSE, MAE, EQR, and the CC were measured, and difference between the replaced values and the original true values was evaluated.

Table 2 Geographic characteristics of the selected rainfall stations in Northern Algeria

No.	Station	Latitude (° N)	Longitude (° E/W)	Altitude (m a.s.l.)	No.	Station	Latitude (° N)	Longitude (° E/W)	Altitude (m a.s.l.)
1	Blida	36.47	2.83 E	256	16	Constantine	36.28	6.61 E	694
2	Alger (Dar El Beida)	36.68	3.25 E	25	17	Bordj Bou Arreiridj	36.06	4.76 E	930
3	Djelfa	34.33	3.25 E	1144	18	Mila	36.433	6.267 E	649
4	Tipaza	36.59	2.45 E	12	19	Chelif	36.21	1.33 E	143
5	Tizi Ouzou	36.70	4.05 E	189	20	Tlemcen	35.01	1.46 W	247
6	Ain Defla	36.30	2.23 E	721	21	Tiaret	35.35	1.43 E	1127
7	Oum El Bouaghi	35.86	7.11 E	891	22	Sidi Belabess	35.20	2.61 W	476
8	Batna	35.567	6.167 E	993	23	Mostaganem	35.88	0.11 E	138
9	Bejaia	36.72	5.07 E	1.76	24	Mascara	35.21	0.15 E	513
10	Tebessa	35.41	8.13 E	813	25	Oran	35.63	0.60 W	90
11	Jijel	36.80	5.78 E	2	26	El Bayadh	33.66	1.00 E	1347
12	Setif	36.18	5.41 E	1038	27	Tissemsilt	35.60	1.833 E	881
13	Skikda	36.88	6.95 E	7	28	Naama	33.58	0.43 W	1149
14	Annaba	36.83	7.81 E	4	29	AinTimouchent	35.3	1.35 W	70
15	Guelma	36.46	7.46 E	228	30	Relizane	35.73	0.55 E	75

m a.s.l. meters above sea level

Imputation

Four PCA simple imputation methods were selected to cover techniques widely applied in the literature and representative of various statistical strategies.

Expectation maximization PCA

EM is a general algorithmic approach to manage latent variable models (including mixtures) popular in large part because it is typically highly scalable and easy to implement (Lin 2010).

Probabilistic PCA

PPCA combines an EM approach for PCA with a probabilistic model. The EM approach is based on the assumption that the latent variables as well as the noise are normally distributed. In standard PCA data, which is far from the training set but close to the principal subspace, may have the same reconstruction error. PPCA defines a likelihood function such that the likelihood for data far from the training set is much lower, even if they are close to the principal subspace. This allows to improve the estimation accuracy. PPCA is tolerant to amounts of missing values between 10 to 15%. If more data is missing, the algorithm is likely not to converge to a reasonable solution (Stacklies and Redestig 2017).

Regularized PCA

Regularized PCA is based on the regularized iterative algorithm, which allows to obtain a point estimate of the parameters and to overcome the major problem of the unfit (Josse et al. 2012).

Singular value decomposition PCA

This implements the SVD impute algorithm as proposed by Troyanskaya et al. (2001). The idea behind the algorithm is to estimate the missing values as a linear combination of the k most significant eigengenes. The algorithm works iteratively until the change in the estimated solution falls below a certain threshold. Each step, the eigengenes of the current estimate are calculated and used to determine a new estimate. An optimal linear combination is found by regressing an incomplete variable against the k most significant eigengenes. If the value at position j is missing, the value of the eigengenes is not used when determining the regression coefficients. SVD impute seems to be tolerant to relatively high amount of missing data (> 10%).

Results and discussion

Performance of the estimation methods

In this study, the comparison was made on the pluviometric series of real data for 10, 20, 30, and 40% gaps. The performances of the estimation methods used are compared and assessed using four measures of performance. The RMSE, MAE, EQR, and CC as criteria to choose the best method of imputation, which have been selected to cover techniques widely applied in the literature and representative of various statistical strategies (Boke 2017). The error measures the difference between the estimated values (predicted) and their corresponding observed values. The four error indices are given according to the following expressions:

$$RMSE = \left[\sum_{i=1}^{n} \frac{(PanObs - PanPred)^2}{n}\right]^{0.5} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |PanPred - PanObs| \quad (2)$$

$$\sum EQR^2 = \sum (PanPObs - PanPred)^2 \quad (3)$$

$$CC = \frac{\sum_{i=1}^{n}\left(PanObs - \overline{Pan}\right)\left(PanPred - \overline{Pan}\right)}{\sqrt{\sum_{i=1}^{n}\left(PanObs - \overline{Pan}\right)^2 \sum_{i=1}^{n}\left(PanPred - \overline{Pan}\right)^2}} \quad (4)$$

where (*PanObs*) *is* the amount of precipitation observed. (*Pan*Pred) is the expected predicted value of precipitation (in this case, it is the imputed value of precipitation). ($\overline{Pan}$) is the means of precipitation and *n* is the number of neighboring station.

The results of the performance of the estimation methods are shown numerically and graphically. Table 3 and Fig. 2 show, respectively, numerical and graphical assessment of simple imputation methods for various percentages of missing values using as criteria: RMSE, MAE, EQR, and CC.

For each percentages of missing data (from 10 to 20%), the performance of each estimation of four methods (PPCA, EM, regularized, and SVD) tends to decrease for RMSE, EQR, and MAE values, resulting in

Table 3 Comparison of estimation methods based on RMSE, CC, MAE, EQR, and number of principal component (NCP) used with four different percentages of missing values after imputation

Percentage of missing values	Method	Number of gaps	RMSE	CC	MAE	EQR	NPC
10	EM	207	142.98	0.82	103.00	0.07	4
	PPCA		147.48	0.81	109.19	0.08	1
	Regularized		138.11	0.83	99.51	0.07	4
	SVD		149.09	0.81	109.95	0.08	1
20	EM	414	135.98	0.84	98.60	0.06	4
	PPCA		148.32	0.80	106.79	0.07	1
	Regularized		128.33	0.86	92.56	0.05	4
	SVD		150.11	0.80	107.64	0.07	1
30	EM	621	151.28	0.82	108.31	0.08	4
	PPCA		151.26	0.81	105.39	0.08	1
	Regularized		141.03	0.84	98.80	0.07	4
	SVD		154.15	0.80	106.12	0.08	1
40	EM	828	165.74	0.78	120.47	0.09	4
	PPCA		147.64	0.82	109.60	0.07	1
	Regularized		147.43	0.82	107.78	0.07	4
	SVD		168.41	0.76	119.42	0.09	1

the increment in CC coefficient. While, for each percentage from 20 to 40%, the performance of each estimation of four methods (PPCA, EM, regularized, and SVD) tends to increase for RMSE, EQR, and MAE values, resulting in the decrease in CC coefficient. The regularized method is found to be the best for four estimation methods used, and the EM method is the second best based on their values of the four error

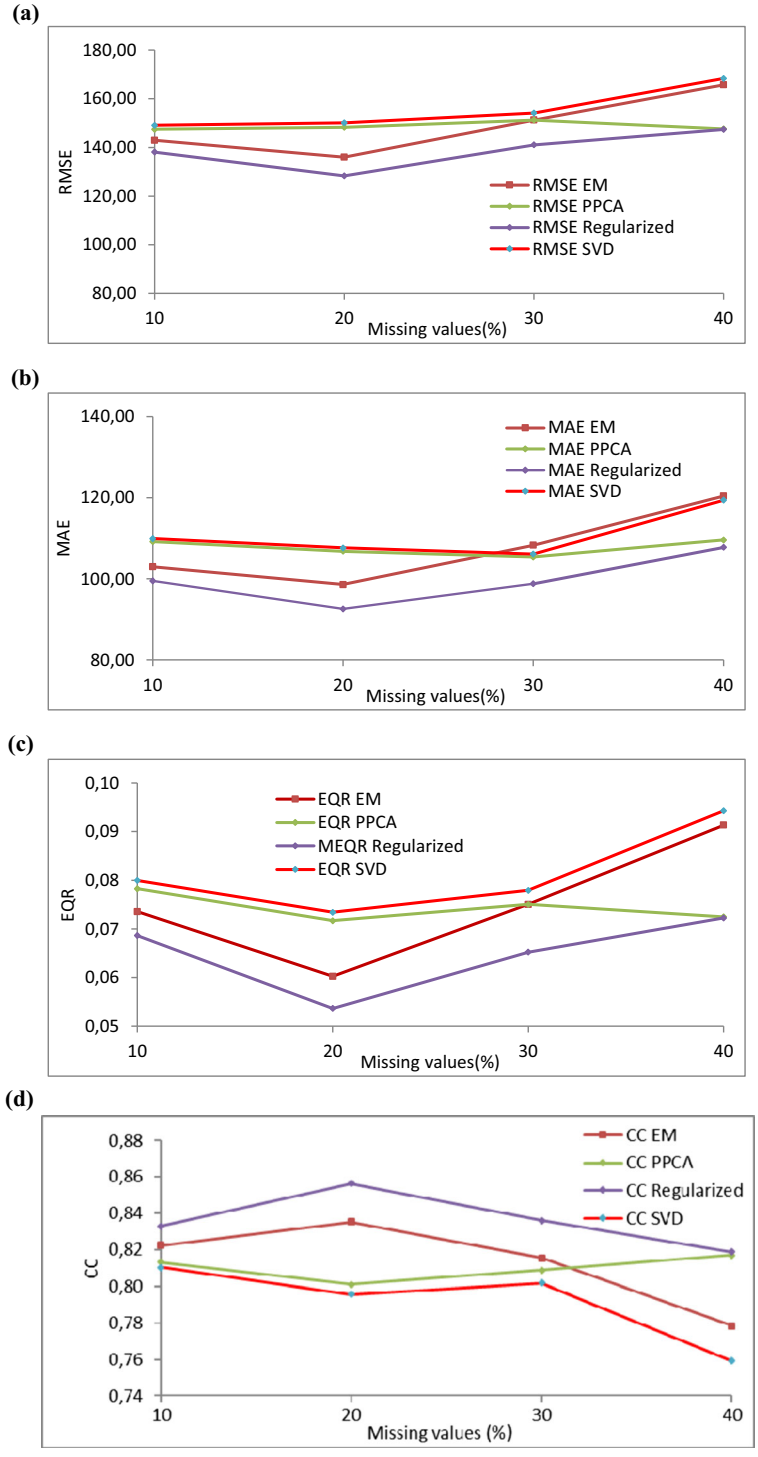

Fig. 2 Assessment of simple imputation methods for various percentages of missing values using four measures of performance criteria. **a** RMSE. **b** MAE. **c** EQR. **d** Correlation coefficient (CC)

indices of 10 to 40%. The lowest performances are given by the SVD and PPCA methods.

Influence of the imputations on the quantiles

According to the above performance, the regularized variant proves to be the best for imputation; nevertheless, other estimates after the filling of a rainfall series are necessary to predict hydrological events using frequency analysis.

In this context, is it always the best valid imputation method for quantile estimation?

In order to answer this pertinent question, we have been interested in the estimation of quantiles.

To avoid calculation for all stations (30 stations), we preferred to proceed to a hierarchical classification by Ward method based on the results of a principal component analysis (Brito et al. 2016). The FactoMiner package of Free Software R was used for this purpose (Lê et al. 2008).

The classification of individuals (stations) into four classes is based on the use of the mean rains of the 12 months of the year over the period of 69 years as active variables (the values of the latter are not mentioned here). Geographic coordinates (latitude, longitude) as well as altitude and interannual monthly totals are taken as additional variables (Fig. 3).

Each of the four classes is represented by a station called "Paragon" (Lê et al. 2008). The paragon is an individual (station) which characterizes on average all the characteristics of its corresponding class.

For this purpose, all the analysis will be done only on the four synoptic stations representative of their classes.

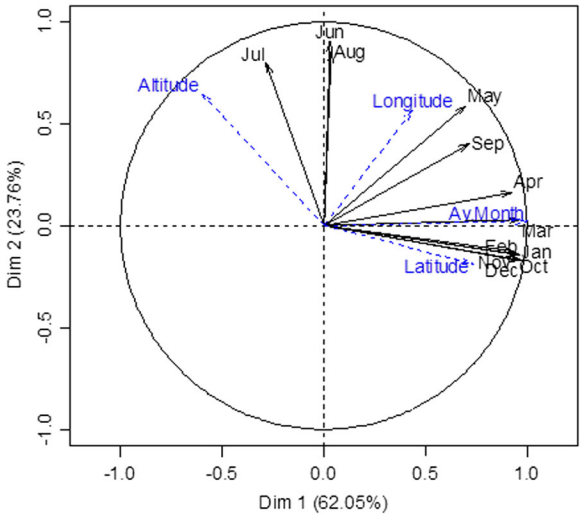

Fig. 3 PCA circle of correlations

Classification of rainfall stations

After a PCA and classification of rainfall stations according to the criteria altitude, attitude, and mean rains of the 12 months, we allowed to have four clusters.

Clusters 1 and 4, respectively, contain 11 and 3 stations, on the other hand, clusters 2 and 3 each contain 8 stations, respectively, illustrated in Fig. 4 and Table 4.

Each cluster is carried by a synoptic station called "Paragon," and the quantiles for the four Paragon stations (Mascara, Batna, Blida and Jijel) were estimated for return periods of 5, 10, 20, 50, 100, 500, and 1000 years, using the normal distribution law, for four PCA imputation variants.

Effect of filling on quantiles

The results and effect of filling on quantiles observed and predicted are shown. Table 5 shows numerical values of predicted quantiles according to return periods for the four Paragon stations (Mascara, Batna, Blida, and Jijel) based on simple imputation methods for various percentages of missing values.

For the Maskara station, Table 5(a) shows that the EM and regularized methods for 10, 30, and 40% of the missing data give a good estimate of the predicted quantiles compared to the observed one with an acceptable positive or negative margin. Also, for 20% of missing data, these methods give a good estimation and the same values of predicted quantiles compared to observed values.

For Batna station, Table 5(b) shows that EM and regularized methods for 10 and 30% of missing data give a good estimation and the same values of predicted quantiles compared to observed quantiles. Also, for 20 and 40% of missing data, these methods give a good estimation of predicted quantiles compared to observed ones with an acceptable positive or negative margin.

For Blida station, Table 5(c) shows that EM and regularized methods for 10 to 40% of missing data give a good estimation of predicted values of quantiles compared to observed values with an acceptable positive or negative margin.

For Jijel station, Table 5(d) shows that EM and regularized methods for 10 to 40% of missing data give a good estimation of predicted quantiles compared to observed quantiles with an acceptable positive or negative margin.

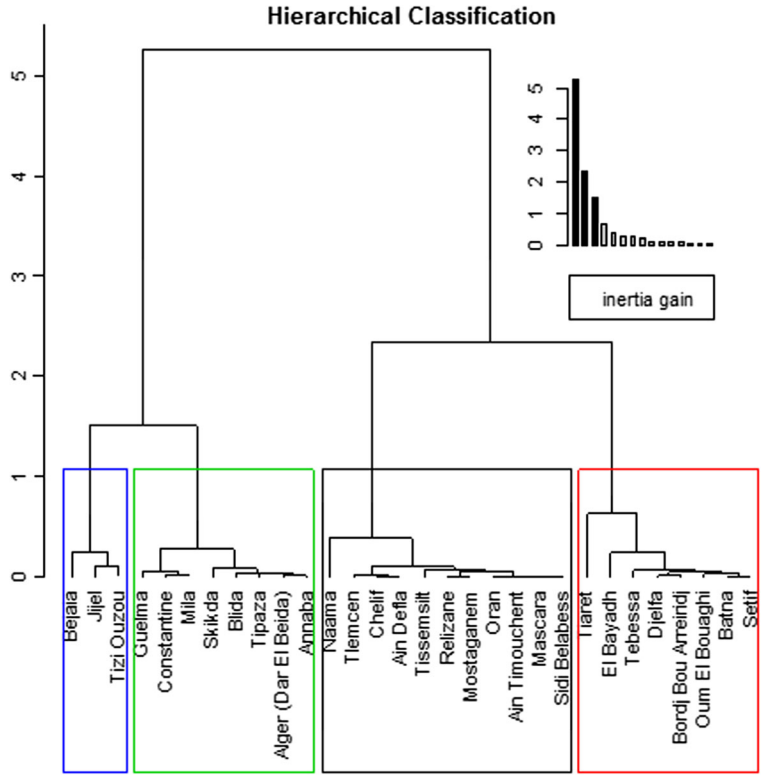

Fig. 4 Hierarchical cluster analysis

Finally, for each percentage of missing data (from 10 to 40%), the regularized method is found to be the best for four estimation methods used and the EM method is the second best; the lowest performances are given by the SVD and PPCA methods based on their values of the two performance criteria, CC and relative error (RE) indices.

CC and RE of observed quantiles with quantiles after filling

CC for the annual rainfall series filled with the variants of the PCA for 10 to 40% of quantiles observed with quantiles after filling for Paragon stations (Mascara, Batna, Blida and Jijel) are illustrated in Table 6. The values of the CC are acceptable and vary between 0.66 to 0.97 for EM and 0.74 to 0.97 for regularized.

RE for the annual rainfall series filled with the variants of the PCA for 10 to 40% of observed quantiles with quantiles after filling for Paragon stations (Mascara, Batna, Blida, and Jijel) are illustrated respectively in Table 7.

The values of the RE for Mascara station vary between 1.7 and 3.4% for EM and 0.20 and 3.5% for regularized (Table 7(a)).

The values of the RE for Batna station vary between 0.17 and 2.71% for EM and 0.46 and 3.64% for regularized (Table 7(b)).

Table 4 Classification of rainfall stations and their paragons

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Ain Defla, Chelif, Tlemcen, Sidi Belabess, Mostaganem, Mascara, Oran, Tissemsilt, Naama, Timouchent, Relizane	Djelfa, O.Bouaghi Batna, Tebessa, Setif B.B.Arreirdj, Tiaret, ElBayadh	Blida, Alger, Tipaza Skikda, Annaba, Guelma, Constantine, Mila	Tizi Ouzou, Bejaia, Jijel
Paragon: Mascara	Paragon: Batna	Paragon: Blida	Paragon: Jijel

Table 5 Quantiles observed and calculated with PCA methods according to return periods for the fourth station for 10 to 40% of filling, (a) Mascara, (b) Batna, (c) Blida, and (d) Jijel

(a)									
Maskara	T (year)	5	10	20	50	100	200	500	1000
Quantiles	Observed	434.0	446.6	453.0	456.8	458.1	458.7	459.1	459.3
10%	EM	441.5	454.7	461.3	465.3	466.6	467.3	467.7	467.8
	Regularized	439.5	452.4	458.9	462.8	464.1	464.8	465.2	465.3
	PPCA	438.5	451.2	457.7	461.5	462.8	463.5	463.9	464.0
	SVD	438.7	451.5	457.9	461.8	463.1	463.7	464.1	464.3
20%	EM	420.3	431.9	437.8	441.4	442.6	443.2	443.6	443.7
	Regularized	420.1	431.7	437.5	441.1	442.2	442.8	443.2	443.3
	PPCA	412.3	423.7	429.5	433.0	434.1	434.7	435.1	435.2
	SVD	412.1	423.5	429.3	432.8	433.9	434.5	434.9	435.0
30%	EM	441.2	454.5	461.2	465.2	466.6	467.3	467.7	467.8
	Regularized	430.0	442.0	448.0	451.7	452.9	453.5	453.9	454.0
	PPCA	471.1	428.0	433.6	436.9	438.0	438.6	438.9	439.0
	SVD	417.4	428.4	433.9	437.3	438.4	439.0	439.3	439.4
40%	EM	440.7	453.3	459.7	463.5	464.8	465.4	465.8	465.9
	Regularized	432.9	444.8	450.8	454.4	455.7	456.3	456.6	456.7
	PPCA	417.0	427.3	432.5	435.7	436.7	437.3	437.6	437.7
	SVD	419.2	429.6	434.9	438.1	439.2	439.7	440.0	440.1
(b)									
Batna	T (year)	5	10	20	50	100	200	500	1000
Quantiles	Observed	447.0	459.0	465.0	468.6	469.8	470.4	470.8	470.9
10%	EM	445.1	456.6	462.4	465.9	467.1	467.6	468.0	468.1
	Regularized	445.0	456.5	462.3	465.7	466.9	467.5	467.8	468.0
	PPCA	444.0	455.4	461.2	464.7	465.8	466.4	466.8	466.9
	SVD	443.1	455.4	461.2	464.7	465.8	466.4	466.8	466.9
20%	EM	446.3	457.9	463.7	467.2	468.4	469.0	469.3	469.5
	Regularized	444.8	456.2	462.0	465.5	466.7	467.3	467.6	467.7
	PPCA	443.1	454.5	460.3	463.8	465.0	465.6	465.9	466.0
	SVD	443.2	454.7	460.5	464.0	465.1	465.7	466.1	466.2
30%	EM	435.3	446.7	452.5	456.0	457.1	457.7	458.1	458.2
	Regularized	435.8	446.8	452.3	455.6	456.7	457.3	457.6	457.7
	PPCA	434.8	445.5	450.9	454.1	455.2	455.8	456.1	456.2
	SVD	434.5	445.2	450.6	453.9	454.9	455.5	455.8	455.9
40%	EM	440.9	451.4	456.7	459.9	461.0	461.5	461.8	461.9
	Regularized	433.8	443.8	448.8	451.9	452.9	453.4	453.7	453.8
	PPCA	432.1	441.7	446.5	449.4	450.4	450.9	451.2	451.3
	SVD	427.7	436.9	441.5	444.3	445.2	445.7	445.9	446.1
(c)									
Blida	T (year)	5	10	20	50	100	200	500	1000
Quantiles	Observed	796.4	824.4	838.6	847.2	850.0	851.5	852.3	852.6
10%	EM	762.9	787.9	800.6	808.3	810.9	812.1	812.9	813.2
	Regularized	762.2	787.1	799.7	807.4	809.9	811.2	811.9	812.2
	PPCA	761.6	786.4	798.9	806.5	809.0	810.3	811.1	811.3
	SVD	761.6	786.4	798.9	806.5	809.0	810.3	811.1	811.3

Table 5 (continued)

20%	EM	770.1	795.2	807.9	815.5	818.1	819.4	820.2	820.4
	Regularized	766.9	791.6	804.0	811.5	814.1	815.3	816.1	816.3
	PPCA	770.9	795.5	808.0	815.5	818.0	819.3	820.0	820.3
	SVD	770.9	795.5	808.1	815.7	818.2	819.5	820.2	820.5
30%	EM	776.9	805.2	819.5	828.2	831.1	832.6	833.4	833.7
	Regularized	771.3	798.1	811.7	820.0	822.7	824.1	824.9	852.2
	PPCA	770.4	796.5	809.7	817.7	820.4	821.7	822.5	822.8
	SVD	770.7	796.8	810.1	818.0	820.7	822.0	822.8	823.1
40%	EM	783.8	812.5	827.0	835.8	838.7	840.2	841.1	841.4
	Regularized	766.8	793.3	806.8	814.9	817.6	819.0	819.8	820.1
	PPCA	786.9	815.8	830.5	839.3	842.3	844.8	844.7	845.0
	SVD	826.1	858.5	884.9	884.8	888.1	889.8	890.8	891.1
(d)									
Jijel	T (year)	5	10	20	50	100	200	500	1000
Quantiles	Observed	1112.6	1242.2	1257.1	1266.1	1269.1	1270.6	1271.5	1271.8
10%	EM	1227.3	1256.8	1271.7	1280.6	1283.6	1285.1	1286.0	1286.3
	Regularized	1225.3	1254.3	1269.0	1277.8	1280.7	1282.2	1283.1	1283.4
	PPCA	1237.2	1267.5	1282.7	1291.9	1295.0	1296.5	1297.5	1297.8
	SVD	1241.2	1271.9	1287.5	4296.8	1300.0	1301.5	1302.5	1302.8
20%	EM	1187.8	1218.0	1233.2	1242.4	1245.4	1247.0	1274.9	1248.2
	Regularized	1183.7	1212.8	1227.4	1236.2	1239.2	1240.7	1241.6	1241.8
	PPCA	1189.1	1217.6	1232.0	1240.7	1243.6	1245.0	1245.9	1246.2
	SVD	1192.1	1221.0	1235.6	1244.4	1247.3	1248.8	1249.7	1250.0
30%	EM	1178.1	1200.3	1211.4	1218.1	1220.4	1221.5	1222.2	1222.4
	Regularized	1175.2	1196.9	1207.8	1214.4	1216.6	1217.7	1218.3	1218.6
	PPCA	1170.2	1191.7	1202.5	1209.0	1211.1	1212.2	1212.9	1213.1
	SVD	1174.3	1196.4	1207.5	1214.5	1216.5	1217.6	1218.3	1218.5
40%	EM	1205.5	1230.9	1243.7	1251.5	1254.1	1255.4	1256.1	1256.4
	Regularized	1196.9	1221.5	1233.9	1241.4	1243.9	1245.1	1245.9	1246.1
	PPCA	1177.8	1201.9	1214.0	1221.3	1223.7	1224.9	1225.7	1225.9
	SVD	1168.8	1193.3	1205.6	1213.1	1215.5	1216.8	1217.5	1217.8

Table 6 Correlation coefficient of quantiles observed with quantiles after filling for the fourth station

Paragon	% of filling	CC EM	CC REG	CC SVD	CC PPCA
Mascara	10	0.97	0.98	0.97	0.97
	20	0.88	0.89	0.87	0.86
	30	0.89	0.93	0.91	0.91
	40	0.78	0.77	0.83	0.82
Batna	10	0.98	0.98	0.97	0.97
	20	0.97	0.97	0.98	0.98
	30	0.86	0.88	0.89	0.89
	40	0.73	0.74	0.79	0.78
Blida	10	0.93	0.93	0.90	0.90
	20	0.85	0.85	0.79	0.79
	30	0.85	0.87	0.93	0.93
	40	0.66	0.74	0.72	0.78
Jijel	10	0.97	0.97	0.95	0.96
	20	0.92	0.93	0.86	0.87
	30	0.75	0.76	0.68	0.69
	40	0.76	0.79	0.66	0.73

Table 7 Percentage (%) of relative error of quantiles observed with quantiles after filling for the fourth station for 10 to 40% of filling, (a) Mascara, (b) Batna, (c) Blida, and (d) Jijel

(a)									
Maskara	T (year)	5	10	20	50	100	200	500	1000
% of Filling	PCA	Relative error (RE (%))							
10%	EM	1.70	1.80	1.80	1.90	1.90	1.90	1.90	1.90
	Regularized	1.30	1.30	1.30	1.30	1.30	1.30	1.30	1.30
	PPCA	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10
	SVD	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10
20%	EM	3.20	3.30	3.30	3.40	3.40	3.40	3.40	3.40
	Regularized	3.20	3.30	3.40	3.40	3.50	3.50	3.50	3.50
	PPCA	5.00	5.10	5.20	5.20	5.20	5.20	5.20	5.20
	SVD	5.00	5.20	5.20	5.30	5.30	5.30	5.30	5.30
30%	EM	1.60	1.80	1.80	1.80	1.90	1.90	1.90	1.90
	Regularized	0.90	1.00	1.10	1.10	1.10	1.10	1.10	1.20
	PPCA	3.90	4.20	4.30	4.40	4.40	4.40	4.40	4.40
	SVD	3.80	4.10	4.20	4.30	4.30	4.30	4.30	4.30
40%	EM	1.60	1.50	1.50	1.50	1.50	1.50	1.50	1.50
	Regularized	0.20	0.40	0.50	0.50	0.50	0.50	0.50	0.50
	PPCA	3.90	4.30	4.50	4.60	4.70	4.70	4.70	4.70
	SVD	3.40	3.80	4.00	4.10	4.10	4.10	4.20	4.20
(b)									
Batna	T (year)	5	10	20	50	100	200	500	1000
% of Filling	PCA	RE (%)							
10%	EM	0.44	0.52	0.56	0.58	0.59	0.59	0.60	0.60
	Regularized	0.46	0.54	0.59	0.61	0.62	0.63	0.63	0.63
	PPCA	0.69	0.77	0.82	0.84	0.85	0.85	0.86	0.86
	SVD	0.69	0.78	0.82	0.84	0.85	0.865	0.86	0.86
20%	EM	0.17	0.24	0.28	0.30	0.31	0.31	0.31	0.31
	Regularized	0.51	.059	.064	0.66	0.67	0.67	0.67	0.67
	PPCA	0.89	0.96	1.00	1.02	1.03	1.03	1.04	1.04
	SVD	0.86	0.93	0.97	0.99	1.00	1.00	1.00	1.00
30%	EM	2.62	2.67	2.69	2.70	2.70	2.71	2.71	2.71
	Regularized	2.52	2.66	2.73	2.77	2.79	2.79	2.80	2.80
	PPCA	2.74	2.94	3.03	3.09	3.11	3.12	3.13	3.13
	SVD	2.80	2.99	3.09	3.15	3.16	3.17	3.18	3.18
40%	EM	1.38	1.65	1.78	1.86	1.89	1.90	1.91	1.91
	Regularized	2.96	3.31	3.48	3.58	3.61	3.63	3.64	3.64
	PPCA	3.35	3.77	3.97	4.10	4.14	4.16	4.17	4.17
	SVD	4.32	4.81	5.05	5.19	5.24	5.26	5.27	5.28
(c)									
Blida	T (year)	5	10	20	50	100	200	500	1000
% of Filling	PCA	RE (%)							
10%	EM	4.21	4.42	4.53	4.59	4.61	4.62	4.63	4.63
	Regularized	4.29	4.52	4.63	4.70	4.72	4.73	4.74	4.74
	PPCA	4.36	4.61	4.73	4.80	4.82	4.83	4.84	4.84
	SVD	4.36	4.61	4.73	4.80	4.82	4.83	4.84	4.84

Table 7 (continued)

20%	EM	3.30	3.54	3.66	3.73	3.76	3.77	3.78	3.78
	Regularized	3.70	3.98	4.12	4.23	4.25	4.25	4.25	4.26
	PPCA	3.20	3.50	3.65	3.74	3.77	3.78	3.79	3.79
	SVD	3.19	3.49	3.63	3.72	3.74	3.76	3.77	3.77
30%	EM	2.45	2.33	2.27	2.24	2.23	2.22	2.22	2.22
	Regularized	3.15	3.18	3.20	3.21	3.22	3.22	3.22	3.22
	PPCA	3.26	3.38	3.44	3.48	3.49	3.50	3.50	3.50
	SVD	3.22	3.34	3.41	3.41	3.45	3.46	3.46	3.47
40%	EM	1.58	1.45	1.38	1.34	1.33	1.32	1.32	1.32
	Regularized	3.72	3.77	3.79	3.81	3.82	3.82	3.82	3.82
	PPCA	1.19	1.04	0.97	0.93	0.91	0.91	0.90	0.90
	SVD	3.73	4.13	4.32	4.44	4.48	4.50	4.51	4.51
(d)									
Jijel	T (year)	5	10	20	50	100	200	500	1000
% of Filling	PCA	RE (%)							
10%	EM	1.21	1.18	1.16	1.15	1.14	1.14	1.14	1.14
	Regularized	1.05	0.98	0.94	0.92	0.92	0.91	0.91	0.91
	PPCA	2.03	2.04	2.04	2.04	2.04	2.04	2.04	2.04
	SVD	2.36	2.40	2.42	2.43	2.43	2.44	2.44	2.44
20%	EM	2.04	1.95	1.90	1.88	1.87	1.86	1.86	1.86
	Regularized	2.38	2.37	2.36	2.36	2.36	2.36	2.36	2.36
	PPCA	1.93	1.98	2.00	2.01	2.01	2.01	2.02	2.02
	SVD	1.69	1.71	1.71	1.72	1.72	1.72	1.72	1.72
30%	EM	2.84	3.37	3.63	3.79	3.84	3.87	3.88	3.89
	Regularized	3.08	3.64	3.92	4.08	4.14	4.17	4.18	4.19
	PPCA	3.49	4.07	4.35	4.51	4.57	4.60	4.61	4.62
	SVD	3.16	3.69	3.94	4.10	4.15	4.17	4.19	4.19
40%	EM	0.59	0.91	1.06	1.15	1.19	1.20	1.21	1.21
	Regularized	1.30	1.66	1.84	1.95	1.99	2.01	2.02	2.02
	PPCA	2.87	3.25	3.43	3.54	3.58	3.60	3.61	3.61
	SVD	3.61	3.94	4.10	4.19	4.22	4.24	4.25	4.25

The values of the RE for Blida station vary between 1.32 and 4.63% for EM and 3.15 and 4.74% for regularized (Table 7(c)).

The values of the RE for Jijel station vary between 0.59 and 3.89% for EM and 0.91 and 4.19% for regularized (Table 7(d)).

Conclusion

In the present study, a comparison of four simple imputation methods (probabilistic PCA, expectation maximization PCA, regularized PCA, and singular value decomposition PCA) is performed, based on a real dataset of different rainfall stations in Algeria according to the MCAR hypothesis. The validation of the results and the choice of the best method of imputation is an important step, so the prediction performances of the four methods are assessed by different statistical criteria like root mean square error, mean absolute error, quadratic error, and correlation coefficient. The study examined the effect of the simple imputations on the quantiles of the rainfall series of 30 stations for the period ranging from 1935 to 2004 (69 years), located in northeast Algeria. The results of the imputations for four different percentages of missing values (PMVs), namely, 10, 20, 30, and 40%, suggests that the regularized PCA and expectation maximization PCA are the best methods which could be used with success to filling gaps. The singular value decomposition PCA and the probabilistic PCA methods (Table 3, Fig. 2) give the lowest performances. Moreover, the regularized PCA and expectation maximization PCA methods are the best in estimating of quantiles compared to the reference observed one, for the four Paragons determined by the cluster analysis, and result in very good to acceptable predicted quantiles regarding the values of CC and RE, such as (CC = 0.97 with 10% of PMV and CC = 0.66 with 40% of PMV; RE = 4.74% with 10% of PMV and RE = 3.82% with 40% of PMV).

References

Audigier, V., Husson, F., & Joss, J. (2015). Multiple imputation for continuous variables using a Bayesian principal component analysis. *Journal of Statistical Computation and Simulation., 86*, 2140–2156. https://doi.org/10.1080/00949655.2015.1104683.

Audigier, V., Husson, F., & Josse, J. (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification, 10*, 5–26. https://doi.org/10.1007/s11634-014-0195-1.

Bigot, S. (2002). Détection des discontinuités temporelles au sein des séries climatiques: point méthodologique et exemple d'application. Actes des Journées de Climatologie de la Commission «Climat et Société» du Comité National Français de Géographie, 27–46.

Boke, A. S. (2017). Comparative evaluation of spatial interpolation methods for estimation of missing meteorological variables over Ethiopia. *Journal of Water Resource and Protection, 9*, 945–959. https://doi.org/10.4236/jwarp.2017.98063.

Brito, T. T., Oliveira-Júnior, J. F., Lyra, G. B., Gois, G., & Zeri, M. (2016). Multivariate analysis applied to monthly rainfall over Rio de Janeiro state, Brazil. *Meteorology and Atmospheric Physics, 129*(5), 469–478. https://doi.org/10.1007/s00703-016-0481-x.

Brock, G., Shaffer, J., Blakesley, R., Lotz, M., & Tseng, G. (2008). Which missing value imputation method to use in expression profiles: A comparative study and two selection schemes. *BMC Bioinformatics, 9*(1), 12. https://doi.org/10.1186/1471-2105-9-12.

Cantat, O. (2004). Critical analysis of rainfalls trends during the 20th century in low-Normandy. Considerations about reliability of data and climate change. https://doi.org/10.4267/climatologie.963.

Dastorani, M. T., Moghadamnia, A., Piri, J., & Rico-Ramirez, M. (2010). Application of ANN and ANFIS models for reconstructing missing flow data. *Environmental Monitoring and Assessment, 166*(1–4), 421–434. https://doi.org/10.1007/s10661-009-1012-8.

El Methni, J. (2013). Contributions to the estimation of extreme quantiles. Applications to environmental data. Dissertation, University of Grenoble.

Faizah, C. R., Hiroyuki, T., Lariyah, M. S., & Hidayah, B. (2016). Homogeneity and trends in long-term rainfall data, Kelantan River Basin, Malaysia. *International Journal of River Basin Management, 14*, 151–163. https://doi.org/10.1080/15715124.2015.1105233.

Josse, J., & Husson, F. (2016). MissMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software, 70*(1), 1–31. https://doi.org/10.18637/jss.v070.i01.

Josse, J., Chavent, M., Liquet, B., & Husson, F. (2012). Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification, 29*, 91–116. https://doi.org/10.1007/s00357-012-9097-0.

Karlsson, I. B., Sonnenborg, T. O., Refsgaard, J. C., Trolle, D., Børgesen, C. D., Olesen, J. E., Jeppesen, E., & Jensen, K. H. (2016). Combined effects of climate models, hydrological model structures and land use scenarios on hydrological impacts of climate change. *Journal of Hydrology*. https://doi.org/10.1016/j.jhydrol.2016.01.069.

Kebede, A., Diekkrüger, B., & Moges, S. A. (2014). Comparative study of a physically based distributed hydrological model versus a conceptual hydrological model for assessment of climate change response in the Upper Nile, Baro-Akobo basin: A case study of the Sore watershed, Ethiopia. *International Journal of River Basin Management, 12*(4), 299–318. https://doi.org/10.1080/15715124.2014.917315.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software, 25*(1), 1–18.

Lee, H., & Kang, K. (2015). Interpolation of missing precipitation data using kernel estimations for hydrologic modeling, Hindawi Publishing Corporation *Advances in Meteorology*. https://doi.org/10.1155/2015/935868,.

Lin, T. H. (2010). A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. *Quality and Quantity, 44*, 277–287. https://doi.org/10.1007/s11135-008-9196-5.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken: Wiley-Interscience. https://doi.org/10.1002/9781119013563.

Luengo, J., García, S., & Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems, 32*(1), 77–108. https://doi.org/10.1007/s10115-011-0424-2.

Melanie, M., & Maria, P. L. (2018). Hydrostatistical study ofthe Paraná and Uruguay rivers. *International Journal of River Basin Management.*, 1–12. https://doi.org/10.1080/15715124.2018.1446962.

Nkiaka, E., Nawaz, N. R., & Lovett, J. C. (2016). Using self-organizing maps to infill missing data in hydro-meteorological time series from the Logone catchment, Lake Chad basin. *Environmental Monitoring and Assessment, 188*(7), 400. https://doi.org/10.1007/s10661-016-5385-1.

Ouarda, T. B. M. J., Ba, K. M., Diaz-Delgado, C., Carstenu, A., Chokmani, K., Gingras, H., Quentin, E., Trujillo, E., & Bobée, B. (2008). Regional flood frequency estimation at ungauged sites in the Balsas River basin, Mexico. *Journal of Hydrology, 348*, 40–58. https://doi.org/10.1016/j.hydrol.2007.09.031.

Presti, R. L., Barca, E., & Passarella, G. (2010). A methodology for treating missing data applied to daily rainfall data in the Candelaro River basin (Italy). *Environmental Monitoring and Assessment, 160*(1–4), 1–22. https://doi.org/10.1007/s10661-008-0653-3.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in survey*. Hoboken: Wiley.

Sattari, M.-T., Rezazadeh-Joudi, A., & Kusiak, A. (2017). Assessment of different methods for estimation of missing data in precipitation studies. *Hydrology Research, 48*(4), 1032–1044. https://doi.org/10.2166/nh.2016.364.

Stacklies, W., & Redestig, H. (2017). The pcaMethods package. CAS-MPG Partner Institute for Computational Biology (PICB) Shanghai. P.R. China And Max Planck Institute for Molecular Plant Physiology Potsdam, Germany. http://bioinformatics.mpimp-golm.mpg.de/.

Stekhoven, D. J., & Bühlmann, P. (2011). Miss Forest-nonparametric missing value imputation formixed-type data.

Bioinformatics, 28(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597.

Suzuki, R., & Shimodaira, H. (2006). Piculs: An R package for assessing the uncertainty in hierarchical clustering. *Bioinfomatics, 22*(12), 15401542. https://doi.org/10.1093/bioinformatics/btl117.

Tabari, H., & Talaee, P. H. (2011). Temporal variability of precipitation over Iran: 1966–2005. *Journal of Hydrology, 396*(3), 313–320. https://doi.org/10.1016/j.jhydrol.2010.11.034.

Tabari, H., Kisi, O., Ezani, A., & Talaee, P. H. (2012). SVM, ANFIS, regression and climate-based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment. *Journal of Hydrology, 444*, 78–89. https://doi.org/10.1016/j.jhydrol.2012.04.007.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520.

Van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton: Chapman & Hall/CRC Press 342 pages. ISBN 9781439868249.