Heliyon

ELSEVIER

Received:
22 October 2018
Revised:
16 December 2018
Accepted:
13 February 2019

Cite as: Amir Aieb, Khodir Madani, Marco Scarpa, Brunella Bonaccorso, Khalef Lefsih. A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria.
Heliyon 5 (2019) e01247.
doi: 10.1016/j.heliyon.2019.e01247

A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria

Amir Aieb [a,b,**], **Khodir Madani** [a], **Marco Scarpa** [c], **Brunella Bonaccorso** [c], **Khalef Lefsih** [a,*]

[a] *Laboratoire de Biomathématiques, Biophysique, Biochimie, et Scientométrie (L3BS), Université de Bejaia, 06000 Bejaia, Algérie*

[b] *Department of Computer Science, Faculty of Exact Science, Abderrahmane Mira University, Bejaïa 06000, Algeria*

[c] *Department of Engineering, University of Messina, Italy*

[*] Corresponding author.

[**] Corresponding author.

E-mail addresses: amir18informatique@gmail.com (A. Aieb), klefsih@yahoo.fr (K. Lefsih).

Abstract

Missing data is a very frequent problem in climatology, it influences on the quality of results that will afford in hydrological studies, as well as water resources management. This paper proposes a new imputation algorithm, based on the optimization of some regression methods, which are hot deck, k-nearest-neighbors imputation, weighted k-nearest-neighbors imputation, multiple imputation, linear regression and simple average method. The choice of these methods was justified by qualitative and quantitative statistical tests analysis. However, the reliability of obtained results depends mainly on percentage of missing data, choice of neighboring stations and data missingness mechanism which should be missing at random. During the study it was found that the most of stations in Soummam watershed don't have a good correlation because the large loss in rainfall data or the geology of watershed which gives a relationship between station position and rainfall variability. For this case, principal component analysis is applied on a set of stations; it showed a positive impact of altitude, latitude and longitude on correlation index between

https://doi.org/10.1016/j.heliyon.2019.e01247

selected stations. The graphical analysis of the normal law on RMSE values, which were obtained by applying the proposed technique in several random cases of missingness, that are 4%, 8%, 12% and 16% respectively, it confirmed the validity and the performance of this approach.

Keywords: Atmospheric science, Environmental science, Hydrology

1. Introduction

The precipitation is one of variables commonly used to study climate variability, flow estimation and even to understand floods and landslides (Hong et al., 2007; Medlin et al., 2007). These studies require complete and reliable records of rainfall data. For this case, data was also represented by satellite maps, deduced by Algorithms to overcome spatial coverage limitations of rainfall and to optimize rainfall measurement (Adler et al., 2000; Pettazzi and Salsón, 2012; Vila et al., 2009). The availability of climatic terrestrial networks is one of factors to obtain the best estimation of precipitation for remote sensing communities (Sharifi et al., 2016), unfortunately in practice the obtained databases contain gaps due to systematic errors or other source of malfunctioning such as the absence of the observer, the destruction of gauging devices, the power failure and the elimination of incorrect data (Sharifi et al., 2016).

The breaks in data acquisition of recording systems are more prevalent in Mediterranean countries (Gyau-Boakye and Schultz, 1994), it is an interesting topic for meteorologists, hydrologists and environmental managers to fill these gaps (Khosravi et al., 2015), however the problem is to find the real data while it has been judged that it is difficult to find a perfect method.

In literature, it can find two possibilities for filling gaps. Firstly, to exclude matrix rows containing even a single MD (missing data elimination techniques), this technique deletes the cells of matrix that contain MD, it is widely used because of its simplicity; however, it does not give the most efficient utilization of data and it can incur a bias, just only if the values are not missing completely at random. Consequently, it can be used only in case where the gaps are very low (Song et al., 2008). Secondly, to estimate gaps of data series by replacing MD with values (imputed missing data techniques) (Li et al., 2007), the choice of these filling techniques refers to case where the strategy is the best reliable. There are many methods which can be applied to complete dataset; these procedures are designed to reduce number of gaps by replacing MD with nearest values, this process called completion or substitution. The most significant advantages of these procedures are the conservation of data dimension, therefore, the strength of statistical data processing. In a more or less broad measurement, all procedures of replacement are not important to be used if

https://doi.org/10.1016/j.heliyon.2019.e01247

there is not random distribution of missing values. However, the replacement of MD is reliable when correlations between variables are very strong. On the other hand, it was shown that there is no ideal technique of estimation could exist (Presti et al., 2010). The effectiveness of each technique depends on a number of factors, such as the percentage of MD, the mechanism of data loss and characteristics of the variable under consideration.

In this regard, there are two types of imputations to be distinguished: simple imputation and multiple imputation. Simple imputation is the one in which the missing value is replaced by the average of all measured observations in the same series, knowing that the amount of MD must be less than 5%; we can also use simple regression methods which starts by studying correlation between observed data of neighboring stations and that of reference stations, this technique is applied when data are missing randomly and the amount of the lack must be between 5% and 10% (Johnson, 2003). In this context, there are two types of regression methods, logistic regression that is used to treat discrete variables, and the linear regression applied for treating continuous variables. Multiple imputation is a treatment that uses several similar databases. It will give many replacements for the same unobserved value. Finally, the subsequent inference obtained by combining all imputed values (Sovilj et al., 2016). The most of these techniques suffer from problem of obtained results reliability when the data missed randomly, the case that let statistical studies based on the other hand to evaluate models which ignored the existence of MD, for example: Principal Component Analysis (PCA) model building (Folch-Fortuny et al., 2015; Nelson et al., 1996) and Wavelet method which is used to study climate variability on discontinuous temporal series that contains missing observations (Turki et al., 2016).

This paper introduces new approach that uses hybrid methods to solve problem of reliability about obtained results of the filling, it was applied in missing climate data series. The proposed technique based on optimization of some imputation methods detailed in the content of this article.

2. Materials & methods

Usually, the choice of each filling method will be done according to missingness mechanisms (different ways in which data are missing). So to manage matrices characterized by incomplete data series properly, the first requirement is to identify the mechanism responsible for data losing (Little and Rubin, 1987). There are three different types of processes to be considered (Little and Rubin, 1987), the simplest case is when data loss occurs absolutely at random, which are indicated with the acronym MCAR (missing completely at random). It occurs when the probability

https://doi.org/10.1016/j.heliyon.2019.e01247

of missing value depends neither on the variable itself nor of another variable of the database; in mathematical term, it is written:

$$P(r|X_{obs}, X_{mis}) = P(r) \tag{1}$$

where X_{obs} is observed data, X_{mis} is MD and (r) is distribution condition of MD.

On the contrary, when the probability that a value is missing depends on the value of other variables, and only on them, the condition is called MAR (missing at random) (Schafer, 1997). That is:

$$P(r|X_{obs}, X_{mis}) = P(r|X_{obs}) \tag{2}$$

Finally, the third and the last case arises when the loss of data does not occur randomly at all (NMAR), in this case, the probability that a value is missing depends on the missing value itself. That is:

$$P(r|X_{obs}, X_{mis}) = P(r|X_{mis}) \tag{3}$$

In general, whether the missingness mechanism is related to study variables or not, it is very significant as to determine how it is difficult to handle MD at the same time (Little and Rubin, 1987).

2.1. Methods

There are many imputation methods in the literature based on different approaches, used in specific domains or even for specific datasets transportation (Tang et al., 2015), meteorology (Junger and De Leon, 2015), and others (Folguera et al., 2015). Although, the proposed methodology can be considered when the MAR hypothesis is assumed (Gómez-Carracedo et al., 2014), it should be borne in mind that it can be applied for MCAR case, which are more random than MAR. The first step is to pre-process the time series, it means to identify gaps distribution, then to select a set of neighboring data series (e.g. neighboring weather stations) to those affected by missing values (target station), respecting the temporal behavior of variables under consideration. In climatology, we can use data recoding from weather station in the same watershed as neighboring time series if there is similarity between data using Spearman coefficient and mean absolute error. The following methods used to fill MD in this paper presented in different descriptions.

2.1.1. k-nearest-neighbors imputation (KNNI)

This method is based on the k observed values of the most similar time series, then all values are used into a single estimate using approaches such as the average

https://doi.org/10.1016/j.heliyon.2019.e01247

methods or kernel function (Amiri and Jensen, 2016). If we can find single nearest neighbor data series (k = 1), in this case it called hot deck (Batista and Monard, 2003). These methods must be appropriate when data are MAR, meaning that the missing value is correlated with other observed variables.

2.1.1.1. Hot-deck

This method performs the estimation of MD of incomplete records P_x using values of similar complete records P_i, belonging to the same data set when there is a large similarity between data, that means (k = 1) (Rahman et al., 2015). In our case study, the nearest station that has the greatest correlation between its values and that of the reference station which allows us to take the same value of this similar station in the same time.

$$P_x = P_i \tag{4}$$

2.1.1.2. Arithmetic average method

When the number of similar data series is equal two or more, a simple average method can estimate MD. In our case study, when rainfall values of each similar station have difference less than 10%, comparing to other measurement of record stations, we can apply KNNI shown in Eq. (5). But, when this difference is very large, a normal-ratio method was recommended. We can point out that the selection of neighboring precipitation stations for estimating MD must be based on meteorological judgment (Teegavarapu and Chandramouli, 2005).

$$P_x = \frac{\sum_{i=1}^{k} P_i}{k} \tag{5}$$

where P_i represents daily rainfall data of similar stations, k is the number of similar stations.

2.1.2. Weighted-nearest-neighbors imputation (WKNNI)

It is another estimation method based on the weighting coefficient of similar time series. In climate data series we used Euclidean distance between similarity stations and reference station to calculate this coefficient, in order to obtain the best estimation of MD then, the final result is determined by a weighted average of all neighboring data (Teegavarapu and Chandramouli, 2005; Troyanskaya et al., 2001). It is given by the following equation:

https://doi.org/10.1016/j.heliyon.2019.e01247

$$P_x = \frac{\sum_{i=1}^{k}(P_i * w_i)}{\sum_{i=1}^{k} w_i} \text{ Such as; } \mathrm{Wi} = d_{xi}^{-K} \tag{6}$$

where P_x is the MD observed in reference station x, k is the number of similar stations; P_i is imputed data, that means the observed rainfall data on neighboring station,w_i is the weighting coefficient that equald_{xi}^{-K},d_{xi} is the Euclidean distance between location of neighboring station i and the reference station x; and K is referred to as friction distance (Vieux, 2001), which ranges from 1.0 to 6.0.

2.1.3. Multiple imputation (MI)

Multiple Imputation is a filling method that provides valid statistical inferences under MAR condition. This method processes MD sets by using the standard procedures of regression, then to make combination between imputing results from these analyses for obtaining final result, as shown in Eq. (7) (Rubin, 1987; Schafer, 1997; Troyanskaya et al., 2001). The multiple imputation has been implemented in software such as SAS (Jerez et al., 2010) and Amelia II package (Honaker et al., 2011). To apply this method, it must follow the following steps: (i) Find k similar databases for each missing value; then the observed values will be used to impute MD. (ii) For each MD (P_x), we use data imputations P_i by applying regression methods to obtain k different estimate results I_i (P_x,P_i). (iii) The final result P_x will be obtained by combining all imputations results I_i using average of all the k complete data values, that is:

$$P_x = \frac{\sum_{i=1}^{k} I_i(P_x, P_i)}{k} \tag{7}$$

2.1.4. Linear regression (LR)

The linear regression is a very fundamental computational procedure which forms the basis of many elaborate algorithms, like the alternating least squares algorithms (ALS) (Wang et al., 2003), it also assumes that values are missing at random. This method requires two steps, initially to estimate the relationship between predictors and missing values, then to use a trend equation for filling the gaps (Bárdossy and Pegram, 2014), and we can express it by Eq. (8):

$$P_x(t) = a + b.P_i(t) \tag{8}$$

The value of a and b can be estimated by using respectively Eqs. (9) and (10):

$$a = \bar{y} - b\bar{x} \tag{9}$$

https://doi.org/10.1016/j.heliyon.2019.e01247

$$b = \frac{\sum_{i=1}^{n} xy - \frac{\sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n}}{\sum_{i=1}^{n} x^2 - \frac{\left(\sum_{i=1}^{n} x\right) x)^2}{n}} \tag{10}$$

Where $\overline{y}$ and $\overline{x}$ are mean values of the data series, respectively, the reference and similarity stations (Bárdossy and Pegram, 2014; Khosravi et al., 2015).

2.1.5. Simple average method

According to (Johnson, 2003), for an amount of MD less than 5%, we can use any filling method; in our case study, this percentage represent just single missing value in one month. He reports that replacing MD with total average of all observations of the same month gives a good result, only if there is low correlation between variables (Presti et al., 2010), we can write that by:

$$P_x = \frac{\sum_{i=1}^{n} P_i}{n} \tag{11}$$

Where P_i is the observed rainfall data, n is the number of days for each month which can be (28, 29, 30 or 31) days.

2.2. Study area

This study focuses on the filling MD of daily precipitation measurements at Bejaia airport weather station of Algeria (Fig. 1), over 32 years. This station is located in the Soummam watershed with an area surface of 9200 km^2, which is a part of eastern Algeria. It is bounded on the north by the mountains of Djurdjura (Lala Khadija with an altitude equal 2308 m), on the east by the plateau of Setif and Babors mountain chains (with an altitude equal 2004 m), as well as on the West by Bouira. Soummam consists of 10 sub-watersheds, equipped with 34 meteorological stations. This area offers a diverse climate and different morphological zones.

In this paper, the proposed filling methodology was applied to daily precipitation time series from January 1st, 1982 to December 31th, 2014, which were obtained from National Agency of Hydraulic Resources A.N.R.H and National Centers for Environmental Information (NOAA), https://www.ncdc.noaa.gov/cdo-web/. The dataset was affected by 3.47% of MD during the entire time of the study. The amount missing was detailed for each month in Fig. 2.

https://doi.org/10.1016/j.heliyon.2019.e01247

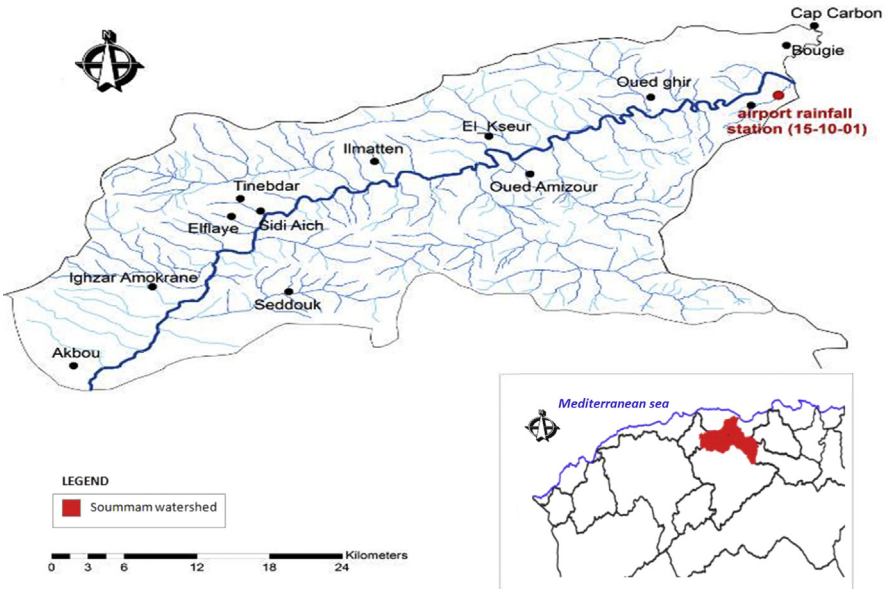

Fig. 1. Geographical map of Soummam watershed borders, followed by its position on north of algerian (medallion map).

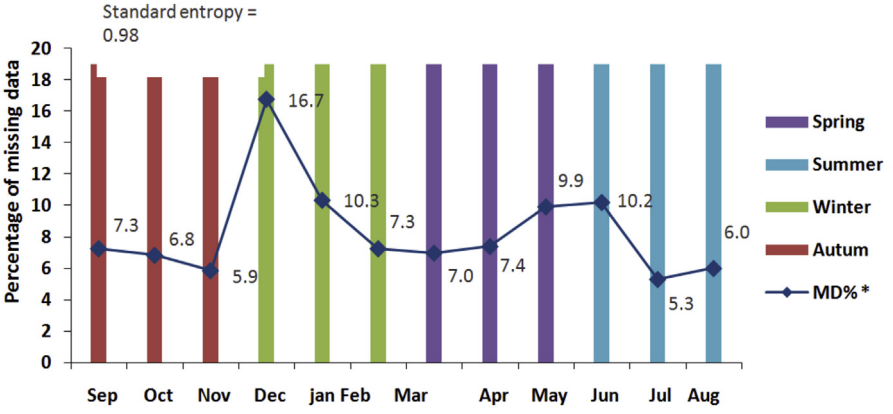

Fig. 2. Seasonal Percentage of missing data distribution per month over 32 year observation at the Bejaia Airport station and the related Standard entropy. *: The percentage of missing data for each month.

2.3. Testing of data distribution

As already indicated above the ultimate objective of this paper is to apply different imputation methods for filling gaps in rainfall time series, in this case it can only be achieved if the mechanism producing data loss is not NMAR (Presti et al., 2010). To exclude this hypothesis, it is possible to test from empirical knowledge that the rainfall has a random distribution (MAR). To this aim, it was assumed that the loss of data was due to failure assessment, which is caused in particular by intense rainy

https://doi.org/10.1016/j.heliyon.2019.e01247

events. If this hypothesis is correct, the following statements should be verifiable: The amount of MD should be affected by obvious seasonal behavior, which usually results in autumn and winter. A positive correlation between the amount of MD and the elevation of any station should exist; this means that rainfall tends to increase as function of altitude.

In this context, it easy to verify the randomness of MD distribution by using standardized entropy, showed in Eq. (12) (Shannon, 1948).

$$H = \frac{\sum_{m=1}^{k} [\ln p(m)]^* p(m)}{\ln(K)} \tag{12}$$

where m represents month, k is the number of months in one year which equal 12 and $p(m)$ is the percentage of MD in each month.

In our case, the standardized entropy coefficient was applied on inter-monthly rainfall data in all 32 years of study, to verify MD distribution during different seasons.

Fig. 2 shows that no relationship exists between the amount of MD and the seasonality data distribution, showing that the greatest percentage of MD measured in summer is 10.2% but in winter it is only 7.3 %. On the other hand, the entropy value is evidently equal 0.98, it is close to 1, allowing us to reject the first hypothesis, therefore rejecting NMAR mechanism. To exclude any dependence of MD on rainfall measurements, MAR and MCAR mechanisms should be tested.

An important consideration should be made to choose between these last assumptions, having to accept the influence of other variables on rainfall measurement, it can be reasonably affirmed that there are some influence also on instrumental failures that caused data losses. Accepting this last statement is equivalent to rejecting the MCAR mechanism. For this purpose (Rubin, 1976; Scheffer, 2002), point out that "MD is very rarely MCAR and they proved that daily rainfall distribution law is not Gaussian but gamma.

2.4. Station similarity

In order to determine the similarities between meteorological stations used to solve problem of MD observed in Bejaia-Airport weather station, the Spearman coefficients and residual average were compared, by coupling the time series values of the target station with the nearby stations values. To determine the similarity in index values, the whole statistical series was considered by including the "dry" days; characterized by no rainfall. This choice was proposed by respecting the following considerations: MD can also include no rainy days, as it was hypothesized that they are missed at random. To use only rainy days could bias

https://doi.org/10.1016/j.heliyon.2019.e01247

the similarity index values as it would not allow for consideration of those events where a zero value was registered in the target station, whilst a rainfall event occurred in a neighboring station (Presti et al., 2010).

Data represented relating to set of stations by PCA graph (Fig. 3) for three years that were selected randomly, this figure gives information about criteria to choose similar climatological stations to that of Bejaia-Airport, in order to fill MD observed in this station.

Fig. 3A shows the PCA graph, as a function of two main components F1 and F2, which are respectively coefficient of correlation (r) and mean absolute error (MAE). They are represented from low to strong, reading respectively; from left to right and from bottom to top. This figure shows three clusters, the choice is mainly based on the degree of correlation, knowing that F1 gives 96% of information rate.

The PCA graph (Fig. 3A) shows that Tichy and Taza are the two nearest stations to Bejaia-Airport, representing the first cluster, they are very similar to each other according to the information obtained by Dendrogram (Fig. 4). However, it is preferable to choose Tichi owing to the availability of data. It provides a strong correlation index (0.98) (Table 1).

Ziama, Tigzirt and Azeffoun stations belong to the same cluster, however Ziama and Tigzirt are more similar than Azeffoun (Fig. 4), and the latter also shows a strong correlation (Table 1).

Annaba and Boumerdes have the same degree of similarity, the both of them belong to the second cluster, shown in Fig. 3A; on the other hand Dar Sghir is a station of the third cluster which gives a lower similarity compared to all stations (Fig. 4).

Fig. 3. **B** represents the PCA graph as a function of two principal components (F1 and F2), which are respectively Euclidean distance between similar stations and

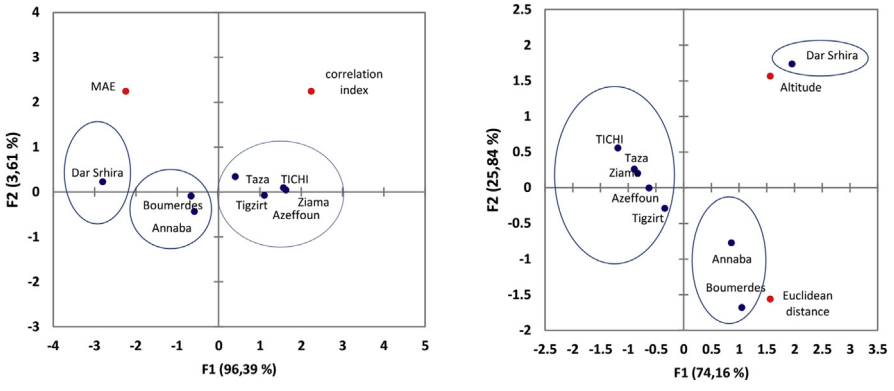

Fig. 3. Principal Component Analysis (PCA) score plot show the similarity between nearby stations setof Bejaia, airport station.

https://doi.org/10.1016/j.heliyon.2019.e01247

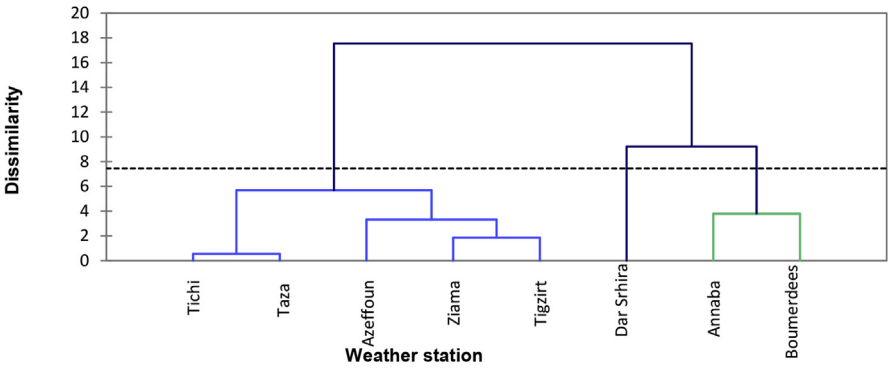

Fig. 4. Dendrogram plot show different clusters of nearest stations of Bejaia, airport weather station.

Table 1. Statistical parameters relating to weathers stations selected for filling missing data of Bejaia-Airport station.

Station	Altitude	Latitude	Longitude	Min	Max	$\overline{X}$	σ	r	ER*	R^2
Tichy	1	36.67	5.16	0	33.0	4.37	7.71	0.98	$0.93X_1$	0.97
Ziama	1	36.67	5.48	0	33.2	4.41	7.74	0.96	$0.97X_2+0.44$	0.94
Azeffoun	1	36.79	4.42	0	33.0	4.49	7.74	0.88	$0.74X_3+0.75$	0.77
Annaba	1	36.83	7.76	0	26.9	4.03	6.87	0.74	$0.40X_4+1.78$	0.55

*: Equation of Regression.**r** : Correlation, $\overline{X}$: Average, σ: Standard deviation, R^2 coefficient of determination.

Bejaia-Airport station and Altitude. The first component gives an information rate equal to 74%, which represents more information, giving that the Euclidean distance for each station is obtained by using latitude and longitude. This graph shows the same clusters that were obtained in Fig. 3A.

A relationship between the geographical positions of stations and the similarity of stations can be noticed. Such as, all stations of the first cluster have small Euclidean distances compared to others stations and altitude measurements between (0.9 and 1.1) which are near to Bejaia-Airport station geographical position that have an altitude equals 1m (Table 2). On the other hand, compared to the information obtained from Fig. 3A, we can deduce that when Euclidean distance decreases, the correlation index increases, as well as, when the altitude of the station increases or decreases the MAE error also increased.

We can confirm this information after interpreting the results of the second and third clusters. Boumerdes and Annaba are two stations belonging to different watersheds, that have the same Euclidean distance (Fig. 3B), and the correlation also gives the same information (Fig. 3A). However it is preferable to take Annaba station for filling gaps in data, because the MAE value of Boumerdes is larger than that of Annaba, we can also find the same meaning to this information using the altitude

https://doi.org/10.1016/j.heliyon.2019.e01247

Table 2. Daily rainfall datasets obtained in February 2005, used for Example 1.

Day	Original dataset	Dataset with MD*	Station similarity			
	Px^1	Px^1	Pn^1	Pn^2	Pn^3	Pn^4
1	12.1	-	11.8	11.1	10.3	3
2	2.1	2.1	2.4	2.8	3.7	5.6
3	17.2	-	-	17	17.4	0.5
4	3.1	3.1	3.4	3	3	4.7
5	4.1	-	-	4	1	-
6	0	0	0	0	0	3
7	0	0	1	0.5	0	1
8	0	0	0	1	3	0
9	0.5	0.5	0.5	0.5	0.4	13.9
10	0	0	0.2	2.8	0	5.9
11	4	-	-	0.6	-	7
12	0	0	0.1	0.5	0	0.5
13	0	0	0.3	3.7	1.3	4.2
14	7.2	7.2	7.6	6.8	9.1	1
15	22.3	22.3	20	25.2	20.7	26.9
16	24.1	-	-	23.4	-	16
17	24.1	24.1	24.1	23.8	24	13.9
18	19.2	19.2	16.7	19	4	5.5
19	6.2	-	-	-	5.8	-
20	3.1	3.1	2.8	3.7	4	4.2
21	1	1	2.1	0.5	0	3
22	9	9	9.3	3.5	6.5	7.1
23	0.5	0.5	1.3	0.5	0.5	0
24	0	0	0	0.5	2.3	0
25	0	0	0	1	1	2.5
26	1	1	1	1.5	1	3
27	1	1	0.5	0	1	3
28	3.1	-	-	-	-	-

* MD (Missing Data). x^1: Bejaia airport station. n^1: Tichy statio,n^2: Ziama station,n^3: Azeffoun station,n^4: Annaba station.

variable (Fig. 3B), knowing that altitude of Annaba is closer to that of Bejaia-Airport compared to Boumerdes.

With these results we can also confirm the previous hypothesis that rainfall is missing at random; this it means that it has a deterministic role on precipitation series, which explain the relationship between geographic coordinates and correlation index.

https://doi.org/10.1016/j.heliyon.2019.e01247

The dendrogram plot shown in Fig. 4 gives information regarding the best choice of the nearest station, according to the amount of MD recorded at each station. The graph shows that the most similar stations are respectively (Tichy, Ziama), (Azeffoun, Tigzirt) and (Annaba, Boumerdes). But in data processing, one station for each cluster can be taken, depending on the filling method used, in this case the following can be selected (Tichy, Ziama, Azeffoun and Annaba).

2.5. Algorithm description

The various methods used to fill the missing climate data series showed before were summarized in the flowchart of the new algorithm (Fig. 5). The proposed methodology was applied on space−time continuous data (i.e. rainfall time series referred to different monitoring stations belonging to the same network). The data was arranged in matrices with the following structure: each matrix represents one year of each climatological station, which has a row that it uniquely associated to the date of the observation and each column is uniquely associated to month.

The first step of this process began with pretreatment of the dataset and tested to find similarities between stations by using the geographical coordinates of each station (Fig. 2) then to calculate the percentage of MD. The chosen method shown in the second step for filling MD is determined as a function of similarity indices between stations, proximity of values representing toleration between imputation values of similarity stations, in addition to the availability of neighboring stations. In the third step, the percentage of MD after every step of the filling must be calculated, in order to check the rate of the gaps remaining after the second step, so that it can be filled either using a simple regression method (LR) or an arithmetic average (SAM). The processing is done iteratively until amount of MD is null.

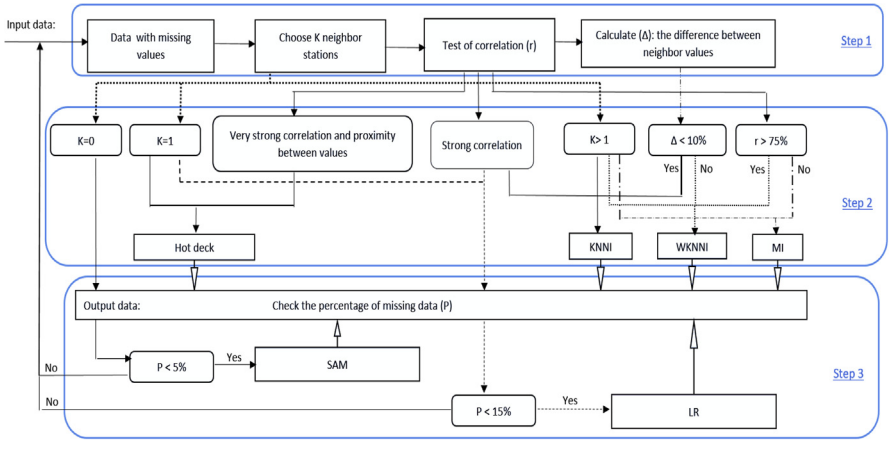

Fig. 5. Flowchart summarizing the filling daily rain fall dataset. Pretreatment of climate data bases (Step1), filling missingdata with the appropriate method (Step2), checking the percentage of missing data (Step3).

https://doi.org/10.1016/j.heliyon.2019.e01247

2.6. Example

Table 2 shows daily precipitation measurements at Bejaia-airport station during February 2005. The effectiveness of this new approach is shown by using 25% of MD that is chosen randomly throughout these series. In this table, stations noted Pn^1, Pn^2, Pn^3 and Pn^4 are respectively, Tichy, Ziama, Azeffoun and Annaba which are used as neighboring stations to process the MD in this example. Table 1 shows statistic parameters of the neighboring stations and large correlation indexes can also be seen, between (0.74 and 0.98), meaning that the similarity between selected stations and the reference station is already existent.

In this example MD are noted by '-' and 'K' means the number of neighboring stations; 'Δ' is the toleration between the rainfall data imputed values of all similar stations in the same time. The calculation steps are the following:

Step1 (Find the K nearest stations):

According to Table 1 and the result obtained in (Fig. 4), Tichy station can be used to fill gaps by hot deck method, the neighboring stations which are respectively, Ziama, Azeffoun and Annaba have been used for filling data by KNNI, WKNNI, LR and MI methods.

Step2 (Filling gaps):

In the 1st day we have P_x is MD values and K is the number of similar stations that equal 4, but in this case it is better to choose Tichy station (Table 1) for filling gaps P_x by using Hot-deck method:

$$P_x = \text{Hot} - \text{deck}\,(Pn^1) = 11.8$$

For the 3rd day: P_x is missing and the number of similar stations K is 3, which are Ziama, Azeffoun and Annaba. So, according to Fig. 5 and Table 1, we can choose between KNNI or WKNNI methods, this implies calculating the difference between imputed values (Δ):

$$\Delta\left(Pn^2, Pn^3\right) = (17.4 - 17)*100/17.4 = 2.29 < 10\%$$

$$\Delta\left(Pn^2, Pn^4\right) = (17 - 0.5)*100/17 = 97.05 > 10\%$$

According to these results, it is better to use only Ziama (Pn^2) and Azeffoun (Pn^3) stations for filling data by applying KNNI method:

$$P_x = KNNI\left(Pn^2, Pn^3\right) = (17 + 17.1)/2 = 17.05$$

For 5th day: in this case, there are just two similar stations (Ziama and Azeffoun), so to calculate MD (P_x), we need to choose between WKNNI and MI methods.

https://doi.org/10.1016/j.heliyon.2019.e01247

$$\Delta\left(Pn^3, Pn^4\right) = (4-1)*100/4 = 75 > 10\%$$

Δ Value is more than 10% and Table 1 shows that the correlation index of rainfall values between "Ziama and Bejaia-Airport" and also between "Annaba and Bejaia-Airport" are 96% and 88%, respectively. So it is preferable to apply WKNNI method:

$$P_x = WKNNI\left(Pn^3, Pn^4\right) = \left(Pn^3 * w_3 + Pn^4 * w_4\right)/(w_3 + w_4)$$

$$w_3 = 1/(d_{n3}) = 1/(58.4) = 0.017$$

$$w_4 = 1/(d_{n4}) = 1/(122.2) = 0.008$$

$$p_x = (4*0.017 + 1*0.008)/(0.017 + 0.008) = 3.04$$

For 11th day: we have two similar stations that are Ziama and Annaba.

$$\Delta\left(Pn^3, Pn^4\right) = (7-0.6)*100/7 = 91 > 10\%$$

One can see in Table 1 that correlation index, between Annaba and Bejaia-Airport station is 74%. In this case, according to Fig. 5, we can fill the datum using MI method:

$$P_x = MI\left(Pn^2, Pn^4\right) = Average(\text{Imput1}, \text{Imput2})$$

To use this method, we must calculate the regression equations, shown on (Table 1) that will be applied in each imputation, which are:

$$Y_2 = 0.97X_2 + 0.44$$

$$Y_4 = 0.40X_4 + 1.78$$

$$\text{Imput1} = 0.97*0.6 + 0.44 = 1.02$$

$$\text{Imput2} = 0.40*7 + 1.78 = 4.58$$

$$P_x = MI(1.02,\ 4.58) = 2.8$$

For 16th day: the number of similar stations K is two (Ziama and Annaba), they have correlation index of 96% and 74%, respectively (Table 1). In this case Δ equal 31%, so we have to use only WKNNI (Fig. 5).

$$P_x = WKNNI\left(Pn^2, Pn^4\right) = 21.24$$

For 19th day: the percentage of MD obtained on Bejaia-Airport station series is 7.14%, so it can fill these gaps using LR method (Fig. 5):

https://doi.org/10.1016/j.heliyon.2019.e01247

$$Y_3 = 0.74X_3 + 0.75$$

$$P_x = LR\left(Pn^3\right) = 5.04$$

For 28th day: the percentage of MD in Bejaia-airport station series is 3.57%, it is less than 5% and it doesn't have at least one similar station, so in this case we can apply only SAM to fill the gaps (Fig. 5):

$$P_x = SAM(P_{x_{obs}}) = 5.08$$

2.7. Experimental

In this section, experimental work was done to compare all methods used in the proposed technique of filling missing climate data. The validation criteria are applied on daily rainfall data for three years which were randomly selected (1997, 2003 and 2013). The calculations applied to fill lack of data in this part were obtained by the implementation of the proposed algorithm on the Delphi language platform (version XE2).

We have chosen four cases where data was missing, which are respectively (4%, 8%, 12% and16%) for showing the reliability of these methods respecting to the same conditions proposed by (Presti et al., 2010). To validate filling data results, it is necessary to compare the obtained rainfall data with actual measurements of the same station using a coefficient of determination R^2 showed in Eq. (13) and an adjusted coefficient of determination R^2_{Adj} showed in Eq. (14). The Error measurements must also be accepted to ensure their reliability, for this case we have used others parameters, such as root mean squared error *RMSE*, mean absolute error *MAE*, which are given by Eqs. (15) and (16), Knowing that the lower RMSE value gives the best imputation (Chai and Draxler, 2014).

$$R^2 = \frac{\sum_{i=1}^{N}\left(X_{model} - \overline{X_{obs}}\right)}{\sum_{i=1}^{N}\left(X_{obs} - \overline{X_{obs}}\right)} \tag{13}$$

$$R^2{}_{Adj}\frac{(1-R^2)(N-1)}{(N-K'-1)} \tag{14}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(X_{obs} - X_{\mathrm{model}}\right)^2}{n}} \tag{15}$$

https://doi.org/10.1016/j.heliyon.2019.e01247

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_{obs} - X_{model}| \quad (16)$$

Where 'n' is the total number of observations, X_{obs} are the observed values and X_{model} are the modeled values. N is the number of points in our data sample. K' is the number of independent repressors, i.e. the number of variables in our model excluding the constant.

3. Results & discussion

In this comparison, different cases of missing values amount were applied to justify the best use of the filling techniques. The number of similar stations considered in this estimation is three and the nearest station, which are Ziama, Azeffoun, Annaba and Tichy respectively.

Fig. 6 show Residual Analysis plots followed by linear regression graphs between imputation results and real rainfall measurements for each filling method when data missed 4%, 8%, 12% and 16%. These graphs show that in the case of 4%, all methods are acceptable and this result confirmed the proposition of Presti about the choice of methods (Presti et al., 2010), we can take into consideration the best choice according to Table.3 which shows that Hot-deck, KNNI, WKNNI and MI methods perform better than LR and SAM.

When the global rate of MD is 8%, the residual values corresponding to SAM graph show that most of the values are negative, flowing with a high fit of data which is remarked on the regression plot (Fig. 6), this method is not usable in this case of study, as it noted a bad correlation (Table 3). The graph shown in Fig. 7 proves that Hot deck and KNNI methods perform the best. WKNNI and MI are similar, having a correlation index that equals 0.82 and 0.84 respectively, however the histogram graph represented by Fig. 8 shows that RMSE values obtained when using MI are higher compared to WINNI.

For a case where the missing data is 12%, the graph shown in Fig. 6 pointed out that the results obtained when LR method was used is less reliable in comparison with all previous cases.

Table.3 shows that hot deck, KNNNI, WKNNI are the methods that perform the most having a strong correlation which are more than 95%, on the other hand the filling by using MI method is still feasible and giving regression index of 81%.

In the last case, the results are given in Table 3. When the percentage of MD was 16%, suggesting that the hot-deck method always provides the best performance, based on both of RMSE and MAR parameters, it can also be found that KNNI and WKNNI provided good values of R^2, which are respectively; 0.997 and 0.923.

https://doi.org/10.1016/j.heliyon.2019.e01247

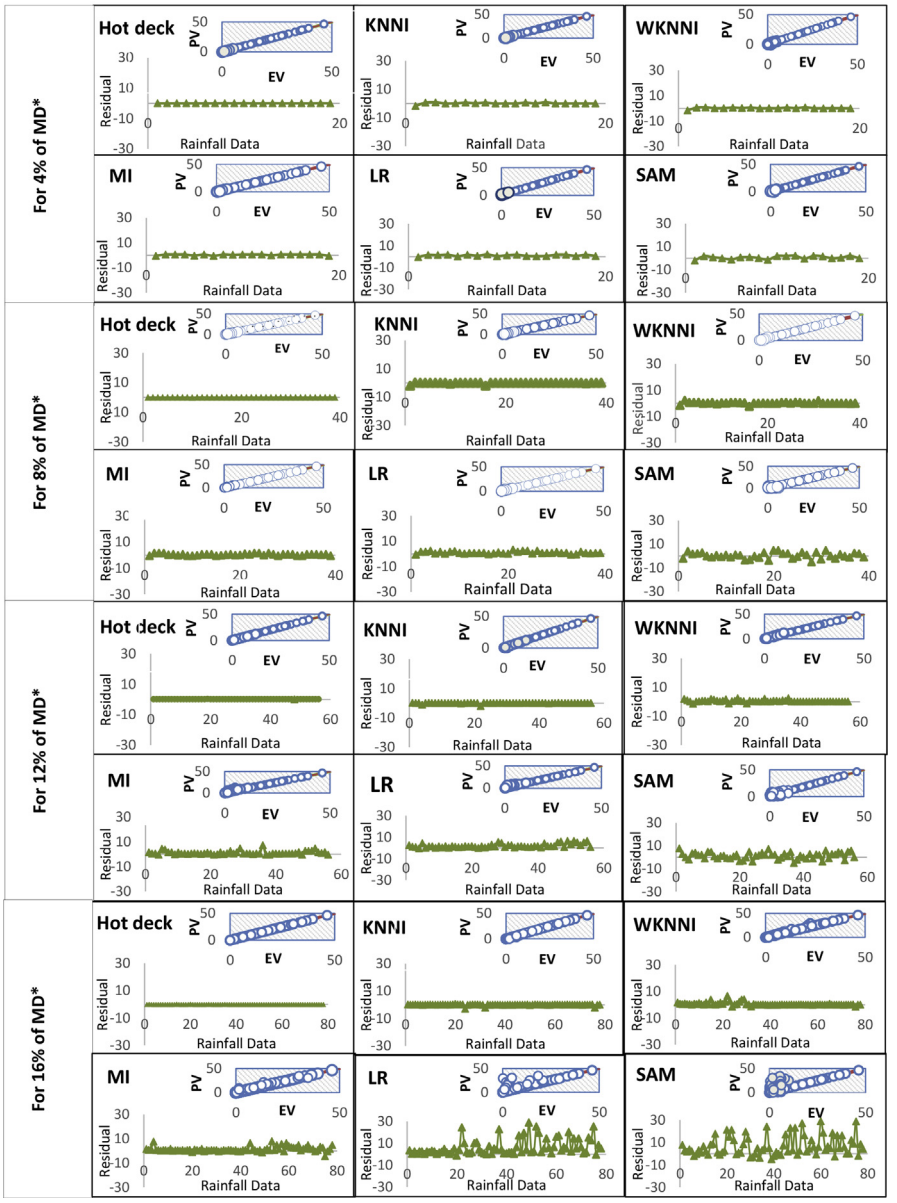

Fig. 6. Residual curves obtained from linear regression graphs (medallion graph) between predicted and experimental values of rainfall data when the amount of missing data equal 4%, 8%, 12% and 16%. (EV) Experimental values, (PV) predicted values, (*): Missing Data.

However, the MI method shows a very high performance, compared to WKNNI (Table.3), as most of the imputation cases to fill the data gap in this study use the data observed at the Annaba station. The latter has a degree of correlation of 74% (Table.1). The regression equations used under MI that were obtained by the correlation study between the observed data of each station are respectively:

$$Y_b = 1.001x_{Zi} + 0.017 \tag{17}$$

https://doi.org/10.1016/j.heliyon.2019.e01247

Table 3. Performance indicators for rainfall missing data when amount are equal to 4, 8, 12 and 16%. Determination coefficient (R^2), adjusted determination coefficient (R^2_{adj}), root mean square error (*RMSE*), mean absolute error *MAE*.

	Hot-Deck	KNNI	WKNNI	MI	LR	SAM
1st case: 4 % of MD*						
R^2	0.999	0.930	0.853	0.924	0.725	0.562
R^2_{Adj}	0.999	0.930	0.852	0.924	0.723	0.56
RMSE	0.009	0.069	0.011	0.113	0.254	0.269
MAE	0.001	0.001	0.007	0.015	0.042	0.024
2nd case: 8 % of MD*						
R^2	0.995	0.918	0.826	0.847	0.712	0.341
R^2_{Adj}	0.995	0.918	0.826	0.847	0.711	0.338
RMSE	0.013	0.137	0.045	0.241	0.396	0.709
MAE	0.002	0.004	0.018	0.04	0.086	0.039
3rdcase: 12 % of MD*						
R^2	0.999	0.994	0.956	0.817	0.679	0.261
R^2_{Adj}	0.999	0.994	0.956	0.816	0.678	0.258
RMSE	0.035	0.094	0.08	0.637	0.959	1.077
MAE	0.003	0.003	0.042	0.136	0.246	0.156
4th case: 16 % of MD*						
R^2	0.999	0.997	0.983	0.952	0.397	0.265
R^2_{Adj}	0.999	0.997	0.983	0.952	0.394	0.246
RMSE	0.128	0.218	0.548	0.978	3.891	4.327
MAE	0.004	0.002	0.086	0.197	1.029	1.083

* MD (Missing Data).

$$Y_b = 1.006x_{Az} + 0.113 \tag{18}$$

$$Y_b = 0.621x_{An} + 0.813 \tag{19}$$

Where Y_b is daily rainfall data imputation in Bejaia-Airport station, x_{zi} is daily rainfall data observed in Ziama station, x_{az} is daily rainfall data observed in Azeffoun station, x_{An} is daily rainfall data observed in Annaba station.

A bad linear relationship and a great deviation of residual values which was observed between experimental and imputing values when LR and SAM was used (showing in Fig. 6), demonstrated that it should be rejected.

After testing all cases of MD, the filling final results explained that the hot-deck method outperforms in all slices filling (Fig. 6). The validation tests relative to KNNI and WKNNI shown in Fig. 7 and Fig. 8, highlight that there is no significant difference in relationship with the respecting amount of missing values.

https://doi.org/10.1016/j.heliyon.2019.e01247

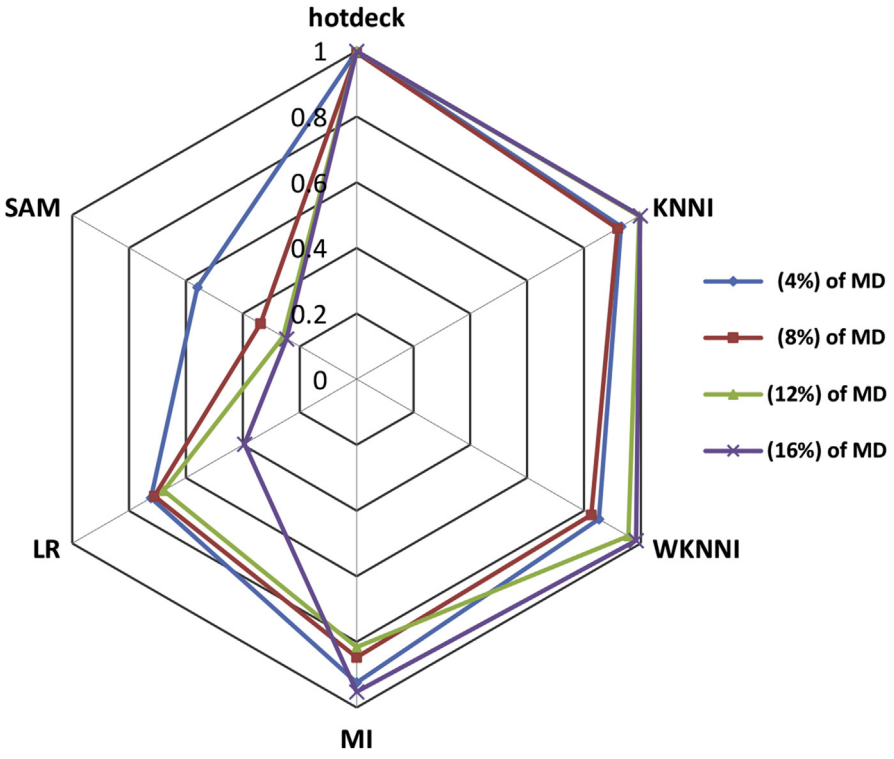

Fig. 7. Radar graph of the adjusted determination coefficientR^2_{Adj}obtained from differentmissing rainfall data imputation methods with different amount of missing data at Bejaia airport station.

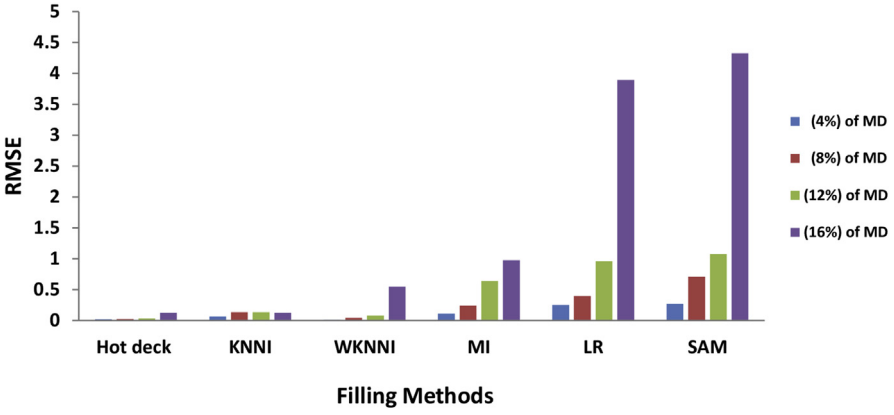

Fig. 8. Histogram of root mean square error*RMSE* obtained from different missing rainfall data imputation methods with 4, 8, 12 and 16% of missing data at Bejaia airport station.

In Table 4 we can see that in all cases where the distance is varying, the choice of K equal 1 gives the best filling results when WKNNI method is used. But when this distance is very large, the similarity indexes of neighboring stations are less than 75%, it can be seen that MI method performs better because the small variation on rainfall intensity can introduce significant changes in the runoff values generated from distributed rain-runoff models (Vieux, 2001). Considering the amounts of

https://doi.org/10.1016/j.heliyon.2019.e01247

Table 4. Weighting coefficients and adjustment quality parameters obtained from WKINNI and MI methods taking into account the neighboring stations and their separating distances (the amount of missing data is equal to 16%).

Stations similarity (X_i)	ED (Km)	Method	K	R^2	R^2_{Adj}	RMSE	MAE
Ziama	41	WKNNI	1	0.9989	0.9989	0.0276	0.0427
Azeffoun	58.4		2	0.9986	0.9986	0.0312	0.0441
			3	0.9983	0.9983	0.0343	0.0454
			4	0.998	0.9980	0.037	0.0465
			5	0.9978	0.9977	0.0392	0.0474
			6	0.9975	0.9974	0.0409	0.0481
		MI	-	0.997	0.9969	0.0447	0.0774
Taza	42.5	WKNNI	1	0.8001	0.7955	0.3837	1.1604
Tigzirt	86.4		2	0.7772	0.7721	0.4305	1.2729
			3	0.7599	0.7544	0.468	1.3750
			4	0.7495	0.7437	0.4915	1.4378
			5	0.7438	0.7379	0.5045	1.4784
			6	0.7227	0.7163	0.435	1.3303
		MI	-	0.8117	0.8074	0.3658	0.0862
Annaba	122.2	WKNNI	1	0.6976	0.6906	0.4641	1.1433
Dar sghir	98.5		2	0.6859	0.6787	0.4746	1.1664
			3	0.6723	0.6648	0.4858	1.2037
			4	0.6576	0.6497	0.4972	1.2387
			5	0.6272	0.6184	0.5083	1.2707
			6	0.6272	0.6184	0.5189	1.2995
		MI	-	0.7435	0.7377	0.4139	0.8448

Euclidean distance between the neighboring station and Bejaia Airport station(ED), coefficient using to calculate weighting coefficient (K), determination coefficient(R^2), adjusted determination coefficient(R^2_{Adj}), root mean square error($RMSE$) and mean absolute error(MAE).

missing values, it is obvious that the results are quite similar to what is obtained for 5% of MD. When this percentage rises, the RMSE raises slightly, showing in Fig. 7. All the errors of obtained measurements and the good fit criterion point conclude that hot-deck, KNNI and WKNNI are better methods compared to regression methods such as MI and LR.

Fig. 9 represents a simulation of RMSE results obtained over the same period of study after to fill (4%, 8%, 12% and 16%) of MD by the proposed algorithm. In each case, RMSE was calculated 50 times randomly for all climatic seasons, in order to study density of error distribution relating to this technical applied on daily rainfall dataset. All graphs show that the distribution of this random variable follows normal law, knowing that in this case the representation of RMSE values were done in class.

https://doi.org/10.1016/j.heliyon.2019.e01247

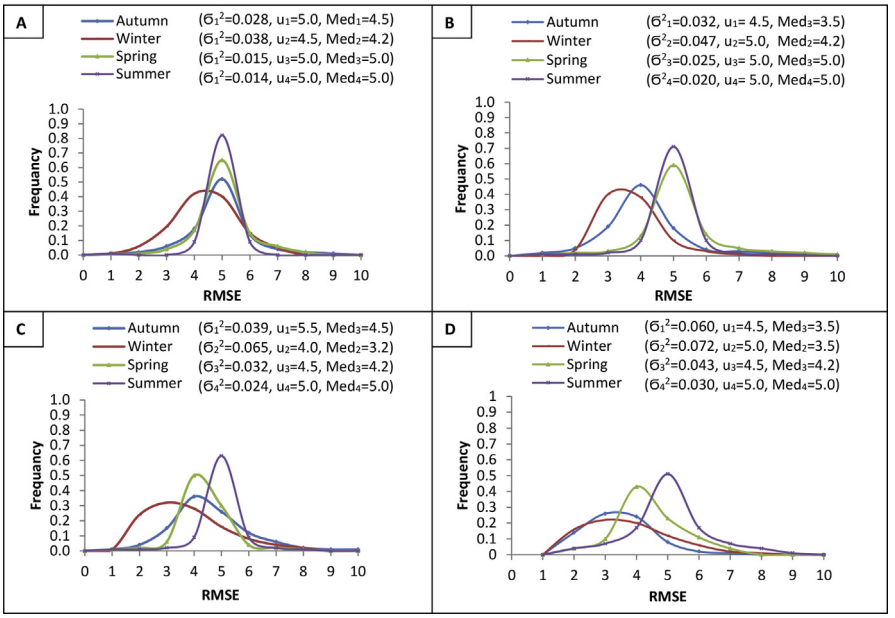

Fig. 9. Simulation of RMSE density obtained by generating randomly 50 times of missing data of each season, using 4% (A), 8% (B), 12% (C) and 16% (D) of missing data.

When the missing data is 4%, the graph of Fig. 9A shows a symmetric distribution recorded in autumn, spring and summer, which for each case gives an average (U = 5) and median (med = 5), the variability changed from weak to strong, respectively, from the driest to the wettest season. On the other hand, the filling data in winter graphically shows an asymmetric distribution (right skewed distribution) that gives positive errors justified by a variability index, which equal 0.038.

When this amount of MD increases to 8%, 12% or 16%, it is found that RMSE distribution is always symmetrical in summer (Fig. 9. B, C, D), only the variability index (бi) increases relatively to the amount of MD; this means that the filling data in this season gives a large proportion of RMSE which always turns around the average. Fig. 9. B, C, and D show that the filling data in autumn and winter when the percentage of MD increases is marked by positive distribution, the variability increases as function of this amount. Whereas the density of RMSE recorded in spring is right skewed distribution only when the lack of data equals 12% and 16%.

After all of these cases, it can be deduced that the filling of data using the proposed algorithm respects the seasonality of the same rainfall series. The density of errors is always positive, which increases relatively with the amount of MD and converges to small values.

https://doi.org/10.1016/j.heliyon.2019.e01247

4. Conclusion

Climate databases suffer from the problem of MD, however the best choice of each method is still on studying missingness mechanism and percentage of MD. The main contribution of this article is to propose a new hybrid technique of filling gaps for estimating missing climate databases more reliable, using climatological network data.

The work shows two main results: (I) explaining the best choice between several of methods, according to the results of comparison using statistical tests on different cases of MD amount; the results shows that the Hot-Deck, KNNI and WKNNI methods give the best filling. It does not depend on the percentage of data. It can be seen that when Tichy, Jijel and Azeffoun stations were chosen for filling 4%, 8%, 12% and 16% of MD, the RMSEs values are always between 0.001 and 0.548. On the other hand, the choice of Annaba and Dar Sghir stations show that the MI method performs better than WKNNI, which have RMSEs of 0.4139 and 0.4641 respectively; these results depend on a degree of similarity between stations, which is less than 75% (II) It shows the influence of altitude, latitude and longitude on the correlation index and the residual average between neighboring stations and reference station. This variation affects the imputation results obtained by MI, LR and SAM. The proposed filling methodology can be also applied to estimate data sets of other case studies when the missing distribution is random.

5. Related work

We want to propose as a future work of this article, the implementation of our filling technique in form of new software that will serve not only climate data bases but we will try to include other methods to generalize the model of filling data. The algorithm will verify the reliability of the results and even the asymptotic complexity of the algorithm comparing to other software and scripts like (Amelia II script for R) and (MDI toolbox for Matlab) and also SPSS software.

Declarations

Author contribution statement

Amir Aieb: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Khodir Mandani: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

https://doi.org/10.1016/j.heliyon.2019.e01247

Marco Scarpa; Brunella Bonaccorso: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Khalef Lefsih: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

We wish to thank the staff of biomathematics, biophysics biochemistry and scientometry Laboratory of Algeria (BBBS) for their precious help about climatology and for giving us opportunity to use their subset data. We also thank the staff of Mobile and Distributed Systems Laboratory in Italy (MDSLab) for their comments and advices on earlier drafts of the paper.

References

Adler, R.F., Huffman, G.J., Bolvin, D.T., Curtis, S., Nelkin, E.J., 2000. Tropical rainfall distributions determined using TRMM combined with other satellite and rain gauge information. J. Appl. Meteorol. 39 (12), 2007−2023.

Amiri, M., Jensen, R., 2016. Missing data imputation using fuzzy-rough methods. Neurocomputing 205, 152−164.

Bárdossy, A., Pegram, G., 2014. Infilling missing precipitation records−A comparison of a new copula-based method with other techniques. J. Hydrol. 519, 1162−1170.

Batista, G.E., Monard, M.C., 2003. An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. 17 (5-6), 519−533.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?−Arguments against avoiding RMSE in the literature. Geosci. Model Dev. (GMD) 7 (3), 1247−1250.

 https://doi.org/10.1016/j.heliyon.2019.e01247

Folch-Fortuny, A., Arteaga, F., Ferrer, A., 2015. PCA model building with missing data: new proposals and a comparative study. Chemometr. Intell. Lab. Syst. 146, 77−88.

Folguera, L., Zupan, J., Cicerone, D., Magallanes, J.F., 2015. Self-organizing maps for imputation of missing data in incomplete data matrices. Chemometr. Intell. Lab. Syst. 143, 146−151.

Gómez-Carracedo, M., Andrade, J., López-Mahía, P., Muniategui, S., Prada, D., 2014. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemometr. Intell. Lab. Syst. 134, 23−33.

Gyau-Boakye, P., Schultz, G., 1994. Filling gaps in runoff time series in West Africa. Hydrol. Sci. J. 39 (6), 621−636.

Honaker, J., King, G., Blackwell, M., 2011. Amelia II: a program for missing data. J. Stat. Software 45 (7), 1−47.

Hong, Y., Adler, R.F., Negri, A., Huffman, G.J., 2007. Flood and landslide applications of near real-time satellite rainfall products. Nat. Hazards 43 (2), 285−294.

Jerez, J.M., et al., 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif. Intell. Med. 50 (2), 105−115.

Johnson, M., 2003. Lose Something? Ways to Find Your Missing Data. Houston Center for Quality of Care and Utilization. Studies Professional Development Series: 17-09. http://www.hsrd.houston.med.va.gov/Documents/MJ%20Missing%20Data%20PDS%20091703.ppt.

Junger, W., De Leon, A.P., 2015. Imputation of missing data in time series for air pollutants. Atmos. Environ. 102, 96−104.

Khosravi, G., Nafarzadegan, A.R., Nohegar, A., Fathizadeh, H., Malekian, A., 2015. A modified distance-weighted approach for filling annual precipitation gaps: application to different climates of Iran. Theor. Appl. Climatol. 119 (1-2), 33−42.

Li, J., Ruhe, G., Al-Emran, A., Richter, M.M., 2007. A flexible method for software effort estimation by analogy. Empir. Softw Eng. 12 (1), 65−106.

Little, R.J., Rubin, D.B., 1987. Statistical Analysis with Missing Data. J. Aufl., New York et al. https://www.wiley.com/en-us.

https://doi.org/10.1016/j.heliyon.2019.e01247

Medlin, J.M., Kimball, S.K., Blackwell, K.G., 2007. Radar and rain gauge analysis of the extreme rainfall during Hurricane Danny's (1997) landfall. Mon. Weather Rev. 135 (5), 1869−1888.

Nelson, P.R., Taylor, P.A., MacGregor, J.F., 1996. Missing data methods in PCA and PLS: score calculations with incomplete observations. Chemometr. Intell. Lab. Syst. 35 (1), 45−65.

Pettazzi, A., Salsón, S., 2012. Combining radar and rain gauges rainfall estimates using conditional merging: a case study. In: The Seventh European Conference on Radar in Meteorology and Hydrology, MeteoGalicia. Galician Weather Service Santiago de Compostela, Spain, p. 5. http://www.meteo.fr/cic/meetings/2012/ERAD/extended_abs/QPE_302_ext_abs.pdf.

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). Environ. Monit. Assess. 160 (1-4), 1.

Rahman, S.A., Huang, Y., Claassen, J., Heintzman, N., Kleinberg, S., 2015. Combining Fourier and lagged k-nearest neighbor imputation for biomedical time series data. J. Biomed. Inf. 58, 198−207.

Rubin, D., 1987. Multiple Imputations for Non Responses in Surveys, vol. 2. Wiley, New York.

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

Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman and Hall/CRC.

Scheffer, J., 2002. Dealing with missing data. Res. Lett. Inf. Math. Sci. 3, 153−160. http://hdl.handle.net/10179/4355.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Sys. Tech. J. 27 (3), 379−423.

Sharifi, E., Steinacker, R., Saghafian, B., 2016. Assessment of GPM-IMERG and other precipitation products against gauge data under different topographic and climatic conditions in Iran: preliminary results. Rem. Sens. 8 (2), 135.

Song, Q., Shepperd, M., Chen, X., Liu, J., 2008. Can k-NN imputation improve the performance of C4. 5 with small software project data sets? A comparative evaluation. J. Syst. Software 81 (12), 2361−2370.

Sovilj, D., et al., 2016. Extreme learning machine for missing data using multiple imputations. Neurocomputing 174, 220−231.

https://doi.org/10.1016/j.heliyon.2019.e01247

Tang, J., Zhang, G., Wang, Y., Wang, H., Liu, F., 2015. A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. Transport. Res. C Emerg. Technol. 51, 29−40.

Teegavarapu, R.S., Chandramouli, V., 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J. Hydrol. 312 (1-4), 191−206.

Troyanskaya, O., et al., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17 (6), 520−525.

Turki, I., et al., 2016. Hydrological variability of the Soummam watershed (Northeastern Algeria) and the possible links to climate fluctuations. Arabian J. Geosci. 9 (6), 477.

Vieux, B.E., 2001. Distributed hydrologic modeling using GIS. In: Distributed Hydrologic Modeling Using GIS. Springer, pp. 1−17.

Vila, D.A., De Goncalves, L.G.G., Toll, D.L., Rozante, J.R., 2009. Statistical evaluation of combined daily gauge observations and rainfall satellite estimates over continental South America. J. Hydrometeorol. 10 (2), 533−543.

Wang, J.-H., Hopke, P.K., Hancewicz, T.M., Zhang, S.L., 2003. Application of modified alternating least squares regression to spectroscopic image analysis. Anal. Chim. Acta 476 (1), 93−109.

 https://doi.org/10.1016/j.heliyon.2019.e01247