Pak. J. Agri. Sci., Vol. 60(2), 327-336; 2023
ISSN (Print) 0552-9034, ISSN (Online) 2076-0906
DOI:10.21162/PAKJAS/23.228
http://www.pakjas.com.pk

Comparison of missing data imputation methods using weather data

Hafiza Nida, Muhammad Kashif [*], Muhammad Imran Khan and Madiha Ghamkhar

Department of Mathematics and Statistics, Faculty of Sciences, University of Agriculture, Faisalabad
*Corresponding author's e-mail: mkashif@uaf.edu.pk

Researchers and data analysts commonly experience challenges while dealing with missing data for analyzing large data sets in their respective field of studies. It is necessary to handle missing data properly to obtain better and more reliable outcomes about any research. The objective of this research is to evaluate different imputation techniques for handling missing observations occurred in the weather data. For this purpose, weather data of the variables: daily rainfall, maximum temperature (T_{max}) and minimum temperature (T_{min}) of 23 stations of Pakistan have been taken from Pakistan Metrological department for the years 1981 to 2020. There are about 14610 total observations of each variable while each variable has different number of missing observations, called as size of missingness, at different stations. The techniques: mean imputation, k nearest neighbors (KNN) imputation, predictive mean matching (PMM) imputation and sample imputation have been considered for the estimation of missing observations found while analyzing data of each station. The minimal value of root mean square error (RMSE) is considered to decide about station-wise imputation technique because the size of missingness varied from station to station. The KNN technique is the most appropriate to estimate the missing observations of the rainfall variables for all the stations while mean imputation technique is recommended for T_{max} and T_{min} data; as compared to other imputation methods.
Keywords: Rainfall, temperature, missing data, imputation methods, root mean square error.

INTRODUCTION

Weather data have been gradually gain considerable attention from researchers of different field of studies to observe the different environmental changes in any country. Temperature and rainfall are the two most important weather variables and plays very important role in many sectors of the country. The agriculture sector is highly affected by the variability of these two variables (Marui *et al.*, 2002). The duration of the growing season is influenced by temperature and plant production is influenced by rainfall (Fiala *et al.*, 2009; Hatfield and Prueger, 2015). Almost all agricultural production forecasting models contain these two weather variables and yield of all major crops mainly depends on these two variables. Pakistan is an agricultural dependent country and is situated in arid regions of Asia; those have been usually characterized by low rainfall and high temperature. The economy of Pakistan depends on agriculture as it has fertile soil that can produce all types of agricultural goods (Awan and Aslam, 2015). The quantification of climate change depends on identifying variations in rainfall and temperature on a regional scale (Khattak and Ali, 2015; Rahman *et al.*,2017). The monitoring of weather data plays very important role for planning of different activities in any region especially for agriculture sector (Fahad and Wang, 2020).Weather is one of the most important sources that affects directly to crop yield, especially in the form of climate variation (Ceglar *et al*., 2017). The valid modelling and the use of up-to-date data analysis techniques will support the farmers to make decisions about timely plantation and managing their fields for drill and harvest (Ray *et al*., 2015). For planning purpose there is requirement of at least thirty year time series data (Azman and Zakaria, 2015), however the collection and storage of long term time series data show huge concerns in context to its quality and usage (Menne *et al*., 2018). One of the major issues with long-term weather data is the presence of missing data values (Afrifa-Yamoah *et al*., 2020). The missing data may occur due to inadequate sample, measurement mistakes, or data collection flaws (Saleem and Ahmed, 2016). Addressing missing data in a time series is a critical challenge for research. It not only causing difficulties in parameters estimation but also cause misconception about the temporal and spatial fluctuations of environmental variables (Houari *et al*., 2016). In different real-life situations, incomplete historical datasets required considered attention from the policy makers. The several researchers have recommended different imputation techniques while dealing with the weather data (Malarvizhi and Thanamani, 2012; Waljee *et al.*, 2013; Armitage *et al.*, 2015; Saleem and Ahmed, 2016; Mehrotra *et al.*, 2017; Yang and Kim, 2018;

Nida, H., M. Kashif, M.I. Khan and M. Ghamkhar. 2023. Comparison of missing data imputation methods using weather data. Pakistan Journal of Agricultural Sciences 60:327-336
[Received 15 Dec 2022; Accepted 25 Apr 2023; Published 27 Jun 2023]
Attribution 4.0 International (CC BY 4.0)S

Egigu, 2020). However; the accurate forecasting for climate data is a challenging task which requires non-missing data to compute precise estimates so that appropriate decisions can be purposed (Azman and Zakaria, 2015; Afrifa-Yamoah *et al.*, 2020). The data having missing values may provide imprecise analysis and misleading results. Therefore, it is always helpful to estimate missing value(s) before doing any statistical analysis, especially when efficient computing resources are available (Saleem and Ahmed, 2016).

***Statement of the Problems and Objective of the study*:** Missing data estimation is a cumbersome task and not commonly addressed through imputation methods while studying the weather data of Pakistan **(Saleem and Ahmed, 2016; Inam *et al*., 2017; Rahman *et al*., 2020);** therefore the objective of the present study is to select appropriate imputation method for estimating missing observations in weather related data of different districts of Pakistan.

MATERIALS AND METHODS

Data source, validity, categorizations and techniques: The daily weather data of three parameters: rainfall, maximum temperature (T_{max}) and minimum temperature (T_{min}) of 23 stations have been used in this research. The data have been collected from Pakistan metrological department for 40 years (1981 to 2020). The names of weather stations along with their geographical coordinates are listed in Table 1 and shown in Fig. 1.

***Missing Data Imputation Methods*:** The approach of estimating missing data from an observation based on relevant values of other variables is known as missing data imputation. Data imputation methods are classified into two types: single imputation and multiple imputation (Schmitt *et al.*, 2015).

Single Imputation: It refers to assigning a possible observation to every missing observation of a given variable in a dataset and then analyzing it as if all the data had been seen initially. The following are the most widely used single data imputation methods, including mean imputation and k nearest neighbors (KNN) imputation.

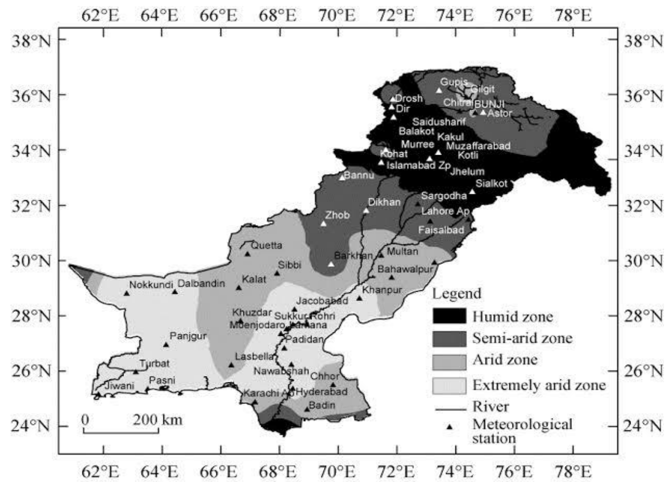

Figure 1. **Map of weather stations along with geographical coordinates.**

Mean Imputation: The mean imputation method is a frequently used approach of substituting missing data. It replaces missing observation with the sample arithmetic.

Let y_1 , y_2 ,…, y_n be the observed values of a dataset containing n observations. The estimate of missing value, y_{mis} using the mean imputation is given by

$$\hat{y}_{mis} = \frac{1}{n_0}\sum_{i=1}^{n_0} y_i$$

$$\hat{y}_{mis} = \frac{\sum_{i=1}^{n_0} y_i}{n_0}, \quad (1)$$

where n_0 show the size of the observed values in the dataset (Mehrotra *et al.*, 2017).

KNN Imputation: In the KNN imputation method, the observations from identical records in the same dataset are copied to fill the missing observations. The similarity is

Table 1. List of 23 Weather stations of Pakistan along with geographical coordinates (Altitude, Latitude and Longitude) of each station.

Station	Altitude (m)	Latitude (°C)	Longitude (°C)	Station	Altitude (m)	Latitude (°C)	Longitude (°C)
Astore	2168	35.35	74.86	**Quetta**	1719	30.18	66.99
Chilas	1250	35.42	74.09	**Jacobabad**	55	28.28	68.44
Dara Ismail Khan	173	31.83	70.91	**Sibbi**	133	29.55	67.87
Dir	1375	35.19	71.86	**Hyderabad**	28	25.38	68.37
Faisalabad	186	31.42	73.08	**Islamabad**	507	33.74	73.08
Gilgit	1459	35.92	74.31	**Rohri**	66	27.68	68.90
Kakul	1308	34.18	73.25	**Karachi**	21	24.86	66.99
Lahore	213	31.58	74.33	**Multan**	122	30.18	71.49
Parachinar	1725	33.89	70.10	**Cherat**	1372	33.82	71.88
Saidu Sharif	961	34.75	72.36	**Jiwani**	56	10.39	-0.37
Sialkot	255	32.49	74.54	**Kalat**	2015	29.03	66.59
Skardu	2209	35.29	75.65				

usually determined using a distance function. The Euclidean function is considered the famous distance function. There is no need in the KNN method to create a prediction model for each variable, but the choice of k value is critical. The KNN defines a collection of k nearest neighbors and then replaces the missing observations for a given variable with the average of its neighbors. In order to estimate a missing observation, y_{mis} using the k nearest neighbors method, the methodology is as follows:

Select the appropriate number of k nearest neighbors.

Compute the distance between the missing values on variable i and the other observed values by Euclidean distance function given as

$$d(y_m, y_0) = \sqrt{(y_{mi} - y_{0i})^2} \quad (2)$$

where y_{mi} is the value of variable i on target population, y_m, y_{0i} is the value of variable i on the other observed value, y_0, and $d(y_m, y_0)$ is the distance between the target observation y_m and the observed value, y_0. Select k nearest observations, based on values with the shortest distance.

Calculate the weights of the k nearest values. The weights can be calculated as

$$W_j = \frac{1}{d(y, y^n{}_j)^2} \quad (3)$$

The estimate of the missing value in the weighted average of k nearest neighbors can be computed as

$$\hat{y}_{mis} = \frac{\sum_{j=1}^{k} W y^n{}_j}{W} \quad (4)$$

Where $y^n{}_j$, j=1,2,…,k, are the k nearest neighbors, W_j are weights of the k neighbors and $W = \sum_{j=1}^{k} W_j$. The KNN has the benefit of being applicable whether the variables are continuous or discrete. Nevertheless, the procedure takes a long time when the size of the data is enormous, and selecting the value of k is also challenging (Malarvizhi and Thanamani, 2012). In this paper the five nearest neighbor values are selected such as $k = 1$, $k = 2$, k =3, k =4 and k =5 (Gou, 2011).

***Multiple Imputation*:** It replaces multiple possible observations (m different complete data sets) for each missing observation rather than a single observation to represent ambiguity about which observation to impute The commonly used multiple imputation methods for weather data are predictive mean matching (PMM) method and sample method.

***PMM Imputation*:** The PMM imputation method is used to estimate missing value by using distance function approach and imputed values are realistic and do not introduce implausible values into the data. The PMM method used different regression models like liner regression, decision tree and random forest and choice of appropriate regression model affect the accuracy of the imputations. Moreover, this method preserved the distributional properties of the variable being imputed and capturing non-linear relationship between variables. Suppose in any dataset , the variable κ has some missing values. There are another variable I with no missing data and is used to impute the missing values of variable κ then the algorithm of PMM imputation methods works in the following way (White *et al*., 2010, Eekhout *et al.*, 2014, Khan. and Hoque. 2020, Abdullah et al.,2022).

i. In order to estimate the missing value of variable κ , a linear regression model is fitted between variable κ and I by considering κ as response variable. The resulting regression coefficients are denoted by β.
ii. The predicted values of the missing observation come from the posterior distribution of complete observation and for this purpose draw a random samples from posterior distribution of β which can be further use to generate multiple imputed datasets. Let β^* represents the new set of coefficients.
iii. From each imputed dataset, the predicted values for κ are generated for all cases.
iv. In order to combine estimates from these multiple imputed data sets, Rubin’s rules which are based on asymptotic theory in a Bayesian framework, are used which provide more accurate and reliable estimates of the missing values. For each missing values of κ this method identifies a set of cases that are closest to the predicted values with missing data.
v. Finally from close cases where multiple values can be used to impute the missing values , one value is randomly chosen to replace the missing value.
vi. Repeat steps ii to v several times to perform multiple imputation for every completed data set.

Sample Imputation: A sample imputation method replaces a missing value by randomly choosing one of the observed data values from a simple random drawn sample of the available observed ones.

Let y_1 , y_2 ,…, y_{n0} be the observed values and $y^1{}_{mis}$, $y^2{}_{mis}$,…, $y^{n_p}{}_{mis}$ be the missing observations of a dataset containing n=n_0+n_p observations; the random sample imputation replaces the missing observations with a sample of n_p from n_0 observed values. The random sample estimation has the benefit of working with both continuous and discrete variables (Mohammed *et al*., 2021).

Measure of Accuracy: The performance of imputation methods could be evaluated using a frequently used measure of accuracy named root mean square error (RMSE). It is a grading method that is quadratic in nature, which also calculates the error's mean magnitude. It is the square root of the mean of the squared differences between imputed values and original values. The RMSE is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \quad (5)$$

where x_i is the observed value and $\hat{x}_i$ is the imputed value concerning imputation methods, while n denotes the size of data (Kamble and Deshmukh, 2017).

RESULTS

***Summary Statistics of Selected Variables*:** The summary of missing observations of rainfall, maximum temperature (T_{max}) and minimum temperature (T_{min}) of each station from 1981 to 2020 are given in Table 2 and presented accordingly in Fig. 2.

From Table 2 it can be observed that the largest value of the percentage of missing observations in rainfall is 11.60% which occurred at Kalat station, followed by Quetta and Jacobabad stations respectively where missing percentage is 8.71% and 8.61% respectively. On the other than no missing observations were observed at Dara Ismail Khan station while the lowest value of percentage (0.01%) of missing observation was observed at Kakul station. Similarly, the Kalat station also showed the highest missing value percentage in both maximum and minimum temperature (T_{max} and T_{min}) respectively with 10.70% and 16.70%. The lowest percentage of missing observations in T_{max} has been observed at Sialkot and Dara Ismail Khan stations respectively. In all other stations, different cases of missing observations have been noticed for all three variables in the study years. Both single and multiple imputation methods can be used for estimating missing values when the missing data percentage lies between 5 to 10% where as any imputation method can be used if the missing percentage is less than 5% ((Jadhav *et al.*, 2019; Presti *et al.*, 2010).

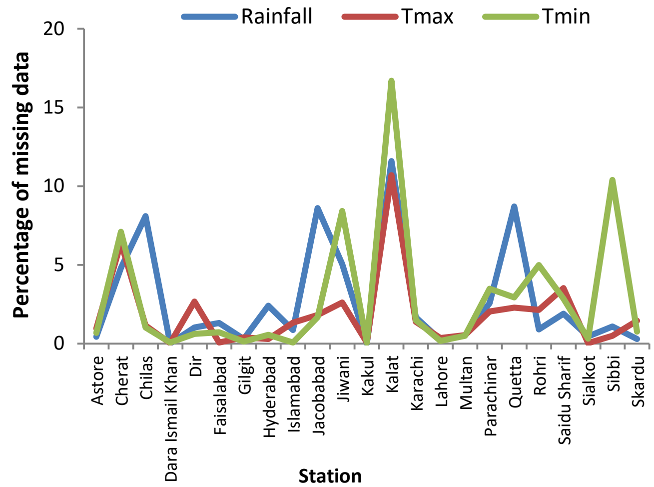

Figure 2. Percentage of missing data with respect to weather stations of three variables (Rainfall, Maximum temperature and Minimum temperature) from 1981 to 2022.

Table 2. Summary of missing data with respect to weather stations of three variables (Rainfall, Maximum temperature and Minimum temperature) from 1981 to 2022.

	Missing observations			Percentage of missing observations		
Station name	**Rainfall**	**Tmax**	**Tmin**	**Rainfall**	**Tmax**	**Tmin**
Astore	62	139	98	0.42	0.95	0.67
Cherat	701	937	1036	4.80	6.41	7.09
Chilas	1182	167	128	8.09	1.14	1.01
Dara Ismail Khan	0	4	10	0.00	0.03	0.07
Dir	148	90	388	1.01	2.66	0.62
Faisalabad	190	9	10	1.30	0.06	0.72
Gilgit	41	55	19	0.28	0.38	0.13
Hyderabad	352	43	74	2.41	0.29	0.56
Islamabad	123	5	10	0.85	1.33	0.07
Jacobabad	1257	264	243	8.61	1.81	1.66
Jiwani	736	382	1232	5.04	2.61	8.43
Kakul	2	9	6	0.01	0.06	0.04
Kalat	1693	1567	2368	11.60	10.70	16.70
Karachi	243	202	204	1.67	1.39	1.49
Lahore	32	48	23	0.22	0.34	0.16
Multan	74	77	72	0.51	0.53	0.49
Parachinar	376	298	507	2.57	2.04	3.47
Quetta	1272	321	424	8.71	2.27	2.93
Rohri	131	310	727	0.90	2.14	4.98
Saidu Sharif	277	513	417	1.90	3.51	2.85
Sialkot	65	4	39	0.45	0.03	0.31
Sibbi	157	64	1519	1.07	0.48	10.40
Skardu	42	211	103	0.29	1.44	0.75

In Fig. 2 the highest percentage of missing data has been observed in the T_{min} variable and the lowest percentage of missing data has been observed in the rainfall variable. The T_{max} has observed a low percentage of missing data as compared to rainfall and T_{min} variables during the whole study period (1981 to 2020). In all 23 stations mostly, the largest value of percentage of missing observations in rainfall has been observed at Kalat station, followed by Quetta and Jacobabad stations.

The summary statistics such as minimum (Min) value, mean, standard deviation (SD), maximum (Max) value, skewness and kurtosis of rainfall, T_{max} and T_{min} data from 1981 to 2020 of all selected stations under study are reported in Table 3.

In case of rainfall data, the minimum value is zero at each station under study, which indicates that there is no rain at those stations during different time periods. The maximum 481 mm per day rain has been observed in the Jacobabad station during the whole study period followed by Lahore station where 300 mm rain was observed. The Dir station showed the highest average rainfall of 3.89 mm with standard deviation of 9.9 mm whereas the lowest average rainfall of 0.23 mm was observed in Jiwani station with standard deviation 2.54 mm during the whole study period. Overall, there observed high value of standard deviation as compared to mean rainfall which indicates that there observed a large variation in the rainfall data at each station. The skewness coefficient in all stations showed that the rainfall data is positively skewed as its value is greater than 0 at all stations. On the other hand, the kurtosis coefficient indicates that the shape of the data at all stations is leptokurtic because kurtosis is greater than 3. The maximum value 27.48 of coefficient of skewness was observed in the Skardu district, while the minimum skewness value of 4.68 was observed in the Parachinar district. On the other hand, the maximum kurtosis 2386.11 was observed in Jacobabad station and the minimum kurtosis was observed in Parachinar district, which is 34.14.

The descriptive analysis of maximum temperature data indicates that the highest average temperature value 35.06 ℃ was observed in the Sibbi station while the minimum average value, 15.97℃ was observed in Astore station. The minimum value of SD was observed in Karachi, which is 3.84, while the maximum value of SD, 11.25, was observed in Skardu. The distribution of T_{max} data is negatively skewed as the skewness value is less than zero for all stations under study except Skardu, which indicates positive skewness. In most stations, the distribution indicates platykurtic except Karachi and Jiwani stations those showed leptokurtic distribution. Similarly in case of minimum temperature data, Skardu stations showed minimum temperature 24 ℃ whereas Jacobabad, Karachi and Sibbi stations indicate highest

Table 3. Summary statistics with respect to weather stations of three meteorological variables (Rainfall, Maximum temperature and Minimum temperature) from 1981 to 2022.

	Rainfall						Tmax						Tmin					
Station	**Min**	**Mean**	**SD**	**Max**	**Skewness**	**Kurtosis**	**Min**	**Mean**	**SD**	**Max**	**Skewness**	**Kurtosis**	**Min**	**Mean**	**SD**	**Max**	**Skewness**	**Kurtosis**
Astore	0.00	1.27	4.64	118.20	7.96	105.95	-6.10	15.97	9.20	41.50	-0.02	1.86	-17.70	4.20	7.89	34.00	-0.09	2.43
Cherat	0.00	1.75	7.25	257.00	9.45	160.99	-0.30	21.33	8.58	45.20	-0.07	2.02	-3.90	12.68	6.82	32.20	-0.06	1.95
Chilas	0.00	0.54	2.82	70.00	10.90	170.94	0.80	26.39	10.26	48.10	-0.03	1.77	-3.50	14.38	9.32	35.80	0.08	1.77
D.I.Khan	0.00	0.89	4.54	202.00	11.02	173.76	2.50	31.67	8.02	51.00	-0.33	2.19	-4.50	17.04	8.26	38.50	-0.30	1.78
Dir	0.00	3.89	9.90	166.00	4.69	33.77	0.00	23.28	8.52	46.00	-0.40	2.14	-13.90	7.74	7.73	32.00	0.12	1.99
Faisalabad	0.00	1.08	5.19	180.30	9.61	130.23	0.00	31.01	7.95	48.20	-0.36	2.27	-1.00	17.48	8.51	37.10	-0.27	1.76
Gilgit	0.00	0.41	2.06	64.50	12.21	219.52	0.70	24.18	9.66	46.30	-0.01	1.86	-9.60	7.67	7.46	32.00	-0.11	2.01
Hyderabad	0.00	0.46	5.07	203.00	20.30	561.05	3.40	34.19	6.01	48.50	-0.49	2.45	1.50	21.11	6.03	39.20	-0.60	2.17
Islamabad	0.00	3.49	12.03	297.00	7.60	96.88	5.00	28.73	7.55	46.60	-0.26	2.17	-2.70	14.46	8.18	38.00	-0.14	1.76
Jacobabad	0.00	0.54	7.59	481.00	42.81	2386.11	7.50	33.97	7.86	52.50	-0.23	2.24	0.00	20.16	8.18	48.00	-0.05	1.83
Jiwani	0.00	0.28	2.54	165.00	27.45	1061.87	11.00	30.43	3.90	48.00	-0.22	3.09	0.00	21.28	5.04	41.50	-0.63	2.62
Kakul	0.00	3.61	10.01	243.00	5.86	69.29	-0.50	23.23	7.27	40.00	-0.38	2.26	-7.30	10.42	7.07	31.10	-0.06	1.81
Kalat	0.00	0.58	3.15	88.00	11.48	186.85	-8.10	22.22	8.47	54.00	-0.32	2.05	-13.00	5.80	7.93	47.50	0.02	1.95
Karachi	0.00	0.52	4.89	223.00	20.14	582.30	2.50	32.38	3.84	48.50	-0.39	3.38	0.50	21.44	6.18	48.00	-0.57	2.06
Lahore	0.00	1.87	8.89	300.00	12.88	274.23	7.40	30.59	7.52	47.40	-0.33	2.33	1.10	18.88	7.68	35.60	-0.28	1.80
Multan	0.00	0.65	4.21	127.00	11.83	180.98	6.20	32.38	7.95	50.00	-0.35	2.19	-2.00	18.71	8.74	38.30	-0.28	1.75
Parachinar	0.00	2.83	7.95	115.00	4.68	34.14	-0.60	21.15	8.43	39.50	-0.24	1.96	-16.60	7.38	8.56	31.00	-0.13	2.11
Quetta	0.00	0.72	3.67	113.00	12.08	221.93	-1.00	25.18	9.40	42.00	-0.32	1.96	-15.00	8.83	8.47	33.50	-0.06	1.99
Rohri	0.00	0.32	3.90	173.70	26.45	937.08	10.00	34.23	7.74	51.00	-0.32	2.03	-5.00	19.71	7.62	46.00	-0.34	1.93
Saidu sharif	0.00	2.87	8.41	187.00	5.27	41.79	1.20	25.99	8.38	45.50	-0.30	1.99	-7.00	12.01	7.55	32.00	-0.04	1.85
Sialkot	0.00	2.75	8.89	251.00	8.53	113.45	6.50	29.35	7.63	47.30	-0.23	2.34	-2.00	16.55	7.95	32.90	-0.28	1.79
Sibbi	0.00	0.49	3.45	81.00	10.86	145.39	1.70	35.06	8.23	53.00	-0.30	2.04	-6.00	19.40	9.11	48.00	-0.30	1.91
Skardu	0.00	0.66	5.24	232.00	27.48	1070.87	-6.60	20.42	11.25	50.50	0.02	2.15	-24.00	6.08	9.77	39.50	0.07	2.64

minimum temperature value of 48℃. The maximum average value of 21.44 ℃ was observed in Karachi and the minimum average value of 5.80 ℃ was observed in Kalat in case of minimum temperature data. The minimum SD value of 5.04 was observed in Jiwani, while the maximum value of SD was observed in Skardu, which is 9.77. The distribution of minimum temperature data is positively skewed for Skardu, Dir, Chilas stations, while it is negatively skewed and platykurtic for other stations.

Imputation Methods**:** To estimate the missing values in rainfall, Tmax and Tmin data sets, the performance of four different imputation techniques is compared. These techniques are mean imputation, k nearest neighbors (KNN), predictive mean matching (PMM) and sample method. In mean imputation, the mean of the non-missing values for each variable was computed and then missing values were imputed by this mean. In k nearest neighbors (KNN) method, five data sets based on five nearest neighbor values were generated such as ($k = 1, k = 2, \dots, k = 5$). In predictive mean matching (PMM) method and sample method the five data sets were also generated including ($p = 1, p = 2, \dots p = 5$) and ($s = 1, s = 2, \dots, s = 5$) where p shows five data sets for prediction mean matching and s shows five data sets based on random sampling method. The root mean square error (RMSE) was used as an accuracy measure for selecting appropriate techniques for estimation of missing observations in the three selected data sets. The results of all four methods are reported in Table 4 to Table 6. The bold value in these tables indicate the minimum RMSE in each imputation method.

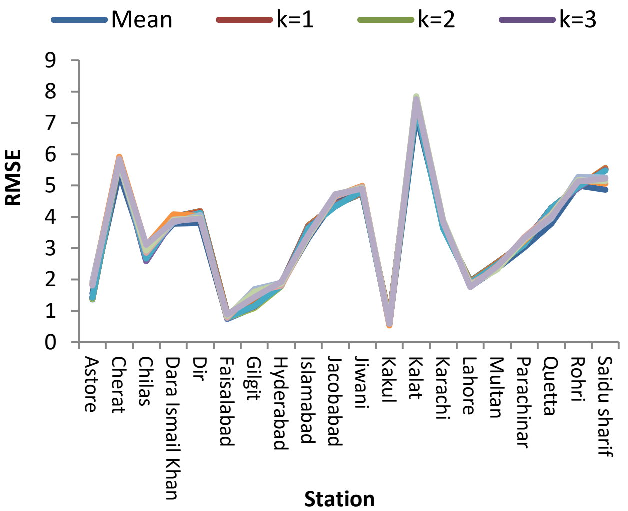

Figure 3. Comparison of different Imputation methods of rainfall data at various missing percentage values based on RMSE.

Table 4. Comparison of different Imputation methods of rainfall data at various missing percentage values. Smaller values are better. Best values are displayed in boldface based on RMSE

Stations	**Single imputation methods**						**Multiple imputation methods**									
	Mean	**KNN**					**PMM**					**Sample**				
		***k*=1**	***k*=2**	***k*=3**	***k*=4**	***k*=5**	***p*=1**	***p*=2**	***p*=3**	***p*=4**	***p*=5**	***s*=1**	***s*=2**	***s*=3**	***s*=4**	***s*=5**
Astore	0.08	0.04	0.05	0.04	0.03	**0.02**	0.10	0.55	0.16	0.05	0.24	0.53	0.21	0.32	0.42	0.19
Cherat	**0.38**	0.88	0.63	0.50	0.42	0.41	1.12	1.00	1.14	0.99	0.72	1.51	2.65	1.85	1.92	1.71
Chilas	**0.15**	0.69	0.58	0.19	0.16	0.16	0.72	0.90	0.98	0.82	1.13	0.65	1.03	0.95	1.15	0.57
Dir	0.39	1.01	0.63	0.41	0.37	**0.25**	1.03	1.26	1.01	1.11	0.90	0.83	1.58	1.38	0.67	0.93
Faisalabad	0.12	0.31	0.18	0.09	0.05	**0.02**	0.19	0.81	0.56	0.45	0.43	0.36	0.70	0.53	1.69	0.83
Gilgit	0.02	0.01	**0.00**	**0.00**	**0.00**	**0.00**	0.04	0.20	0.01	0.03	0.14	0.13	0.03	0.21	0.29	0.14
Hyderabad	0.07	0.13	0.09	0.06	0.03	**0.00**	1.02	1.24	1.45	1.49	1.26	0.31	0.79	0.52	0.14	0.65
Islamabad	0.32	1.76	1.99	0.56	0.43	0.34	0.62	0.76	0.91	**0.16**	1.15	1.52	1.70	0.88	0.71	0.83
Jacobabad	0.16	0.65	0.51	0.12	0.08	**0.04**	0.16	4.13	0.89	1.02	0.97	1.64	0.81	1.40	4.09	1.11
Jiwani	**0.06**	0.56	0.79	0.15	0.35	0.07	0.41	0.50	0.57	0.51	1.44	1.02	1.29	0.82	0.62	0.53
Kakul	0.04	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.11	**0.00**	**0.00**	**0.00**	0.16
Kalat	**0.20**	0.76	0.72	0.56	0.51	0.49	1.40	1.78	1.36	1.83	1.36	1.31	1.12	0.94	1.17	0.57
Karachi	**0.07**	0.62	0.37	0.44	0.30	0.29	1.06	0.66	1.21	0.32	0.35	1.11	0.18	0.96	0.47	0.17
Lahore	0.09	0.06	0.84	0.83	0.47	0.10	0.14	0.15	0.16	0.36	0.25	0.45	**0.04**	0.27	0.16	0.90
Multan	0.05	0.10	0.01	0.01	0.01	**0.00**	0.34	0.14	0.02	0.12	0.63	0.27	0.04	0.17	0.19	0.02
Parachinar	**0.45**	1.12	1.05	0.68	0.56	0.46	1.55	1.54	1.48	1.23	1.32	1.41	1.60	0.89	1.40	1.13
Quetta	**0.21**	1.12	0.90	0.64	0.62	0.55	1.57	1.48	1.56	1.43	1.57	1.40	1.54	1.14	1.36	1.36
Rohri	0.03	0.14	0.17	**0.01**	0.08	0.07	1.02	0.62	1.49	0.86	0.18	0.18	0.14	0.07	0.51	0.76
Saidu sharif	0.40	0.82	0.64	0.67	0.58	0.48	0.96	1.39	1.25	1.14	1.25	1.17	1.13	1.24	0.29	**0.14**
Sialkot	0.18	0.29	0.26	0.07	0.05	**0.01**	0.91	0.20	0.23	0.66	0.57	1.26	0.73	1.04	0.27	0.41
Sibbi	**0.05**	0.69	0.37	0.07	0.07	0.07	0.52	0.60	0.35	0.36	0.43	0.46	0.13	0.38	0.19	0.25
Skardu	0.04	0.05	0.09	0.02	0.03	**0.01**	0.12	0.28	0.08	0.05	0.10	0.06	0.06	0.03	0.53	0.13

Table 5. Comparison of different Imputation methods of maximum temperature data at various missing percentage values. Smaller values are better. Best values are displayed in boldface based on RMSE.

Stations	Single imputation methods						Multiple imputation methods									
	Mean	KNN					PMM					Sample				
		k=1	k=2	k=3	k=4	k=5	p=1	p=2	p=3	p=4	p=5	s=1	s=2	s=3	s=4	s=5
Astore	1.56	1.43	1.41	1.39	1.41	1.41	**1.36**	1.38	1.37	1.43	1.43	1.87	1.95	1.84	1.84	1.81
Cherat	**5.40**	5.72	5.73	5.74	5.74	5.74	5.80	5.81	5.78	5.81	5.82	5.92	5.69	5.87	5.77	5.82
Chilas	2.82	2.72	2.70	2.69	2.69	2.70	**2.59**	2.63	2.62	2.59	2.67	3.08	2.88	2.86	2.93	3.11
D.I. Khan	**3.78**	3.98	3.99	4.01	4.00	3.99	3.92	3.94	3.91	3.91	3.93	4.08	3.83	3.92	3.91	3.87
Dir	**3.80**	4.17	4.17	4.18	4.18	4.18	4.09	4.16	4.13	4.12	4.12	4.03	4.06	3.97	3.97	3.95
Faisalabad	0.77	0.89	0.92	0.93	0.94	0.91	**0.74**	0.77	0.78	0.80	0.78	0.78	0.78	0.80	0.84	0.88
Gilgit	1.48	1.20	1.17	1.16	1.17	1.22	1.11	1.14	**1.09**	1.14	1.16	1.65	1.69	1.62	1.61	1.44
Hyderabad	1.85	1.83	1.82	1.86	1.86	1.84	1.80	1.81	**1.78**	1.84	1.80	1.88	1.89	1.80	1.87	1.93
Islamabad	**3.32**	3.70	3.72	3.72	3.72	3.72	3.66	3.69	3.67	3.66	3.64	3.42	3.38	3.39	3.39	3.44
Jacobabad	4.57	4.41	4.39	4.39	4.39	4.39	4.37	**4.34**	4.36	4.36	**4.34**	4.66	4.66	4.65	4.71	4.72
Jiwani	4.92	4.77	4.76	**4.75**	**4.75**	4.76	4.82	4.84	4.81	4.81	4.83	4.99	4.94	4.97	4.93	4.90
Kakul	0.58	0.77	0.76	0.77	0.76	0.77	0.59	0.65	0.63	0.66	0.60	**0.54**	0.59	0.66	0.65	0.59
Kalat	**7.28**	7.60	7.59	7.58	7.58	7.58	7.54	7.60	7.53	7.57	7.52	7.77	7.82	7.81	7.85	7.76
Karachi	3.82	3.66	3.66	3.68	3.68	3.69	3.68	3.67	3.68	3.67	**3.63**	3.87	3.85	3.86	3.90	3.85
Lahore	1.77	1.95	1.94	1.95	1.96	1.94	1.87	1.87	1.89	1.88	1.89	1.77	**1.76**	1.85	1.82	**1.76**
Multan	2.35	2.54	2.54	2.55	2.54	2.54	2.49	2.48	2.47	2.49	2.49	2.44	2.40	2.35	**2.32**	2.43
Parachinar	**3.02**	3.17	3.14	3.16	3.16	3.17	3.13	3.12	3.11	3.15	3.11	3.23	3.28	3.35	3.27	3.31
Quetta	**3.78**	4.28	4.27	4.28	4.27	4.27	4.28	4.27	4.27	4.28	4.28	4.04	3.95	4.08	4.05	4.00
Rohri	5.00	4.95	4.96	4.96	4.97	4.97	4.93	4.95	4.93	**4.92**	4.96	5.18	5.28	5.17	5.17	5.13
Saidu sharif	**4.87**	5.54	5.52	5.53	5.54	5.56	5.51	5.51	5.49	5.49	5.48	5.06	5.25	5.13	5.15	5.22
Sialkot	0.73	0.70	0.64	0.69	0.72	0.70	0.63	0.66	0.64	0.70	0.63	0.79	0.77	0.80	**0.62**	0.68
Sibbi	2.43	2.87	2.85	2.85	2.85	2.84	2.78	2.82	2.79	2.82	2.82	2.51	2.56	**2.37**	2.51	2.49
Skardu	**2.45**	3.75	3.66	3.72	3.69	3.70	3.36	3.34	3.27	3.36	3.34	2.85	2.94	2.91	2.92	2.79

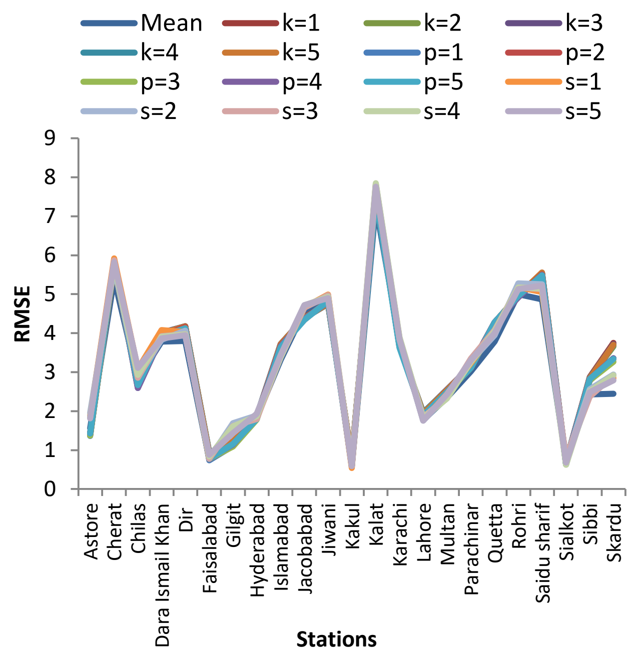

Figure 4. Comparison of different Imputation methods of maximum temperature data at various missing percentage values based on RMSE.

The results from **Table** 4 and **Fig.** 3 indicate that the mean imputation approach is suitable for Parachinar, Cherat, Sibbi, Karachi, Chilas, Kalat, Quetta, and Jiwani stations in estimating rainfall data. On the other hand, the KNN imputation method at $\boldsymbol{k} = \mathbf{3}$ is appropriate for Rohri, Gilgit, Skardu, Astore, Dir, Faisalabad, Sialkot, Multan, Hyderabad and Jacobabad stations. The KNN at $\boldsymbol{k} = \mathbf{5}$, is appropriate for Rohri, Gilgit, Skardu, Astore, Dir, Faisalabad, Sialkot, Multan, Hyderabad, and Jacobabad. The PMM imputation method at $\boldsymbol{p} = \mathbf{4}$ is appropriate for Islamabad, while the sample method at $\boldsymbol{s} = \mathbf{4}$ is appropriate for Saidu Sharif except for $\boldsymbol{s} = \mathbf{1}$ and $\boldsymbol{s} = \mathbf{5}$. There is no missing rainfall data observed in Dara Ismail khan, so there is no need of imputation for this station.

Like rainfall data the missing value is also computed in temperature data and results of all selected methods are reported in Table 5 for T_{max} and in Table 6 for T_{mini} data. In case of maximum data, mean approach is appropriate for Saidu Sharif, Skardu, Dir, Parachinar, Dara Ismail Khan, Islamabad, Quetta, Cherat, and Kalat stations based on smallest value of RMSE. The KNN method at $k = 3$ and $k = 4$, is suitable for Jiwani station only. Whereas the PMM method is appropriate for Astore, Chilas and Faisalabad stations at $p = 1$, Gilgit and Hyderabad at $p = 3$, Rohri at $p = 4$ and Karachi at $p = 5$, while at $p = 2$ and $p = 5$ this method is appropriate for Jacobabad station. The sample imputation method is appropriate for Kakul and Lahore stations at $s = 1$, and $s = 5$ respectively. The sample imputation method is also suitable for Sibbi, Sialkot and

Table 6. Comparison of different Imputation methods of minimum temperature data at various missing percentage values. Smaller values are better. Best values are displayed in boldface based on RMSE.

	Single imputation methods						Multiple imputation methods									
Stations	**Mean**	**KNN**					**PMM**					**Sample**				
		$\boldsymbol{k}$=1	$\boldsymbol{k}$=2	$\boldsymbol{k}$=3	$\boldsymbol{k}$=4	$\boldsymbol{k}$=5	$\boldsymbol{p}$=1	$\boldsymbol{p}$=2	$\boldsymbol{p}$=3	$\boldsymbol{p}$=4	$\boldsymbol{p}$=5	$\boldsymbol{s}$=1	$\boldsymbol{s}$=2	$\boldsymbol{s}$=3	$\boldsymbol{s}$=4	$\boldsymbol{s}$=5
Astore	**0.34**	0.64	0.62	0.66	0.65	0.65	0.66	0.68	0.66	0.67	0.68	0.71	0.71	0.69	0.73	0.70
Cherat	**3.38**	3.74	3.71	3.72	3.71	3.73	3.63	3.58	3.66	3.64	3.66	3.79	3.87	3.80	3.77	3.84
Chilas	1.44	1.35	1.33	1.33	1.33	1.33	1.33	1.36	1.33	**1.32**	1.34	1.65	1.72	1.64	1.79	1.76
D.I. Khan	0.45	0.44	0.43	0.42	0.42	0.42	0.33	**0.30**	0.32	0.31	**0.30**	3.22	0.48	0.41	0.49	0.55
Dir	**0.61**	0.97	0.98	0.99	0.98	0.99	1.01	1.02	1.02	0.97	1.00	0.87	0.89	0.90	0.86	0.89
Faisalabad	1.48	1.30	1.30	1.28	1.29	1.29	1.30	**1.27**	1.29	1.28	1.29	1.69	1.61	1.77	1.64	1.71
Gilgit	0.28	0.34	0.33	0.33	0.33	0.32	0.25	0.24	**0.19**	0.20	0.25	0.31	0.41	0.37	0.40	0.35
Hyderabad	1.58	1.51	1.51	1.51	1.50	1.50	1.49	1.49	1.48	**1.47**	**1.47**	1.62	1.68	1.64	1.66	1.55
Islamabad	0.38	0.34	0.34	0.34	0.34	0.34	0.24	0.26	0.24	0.29	**0.21**	3.73	0.40	0.39	0.41	0.47
Jacobabad	2.60	2.61	2.58	**2.56**	**2.56**	**2.56**	2.59	2.60	2.59	2.61	2.57	2.91	2.87	2.81	2.84	2.84
Jiwani	**6.18**	6.46	6.46	6.46	6.45	6.46	6.40	6.42	6.38	6.42	6.44	6.39	6.31	6.38	6.36	6.45
Kakul	**0.21**	0.35	0.36	0.36	0.37	0.37	0.27	0.25	0.28	0.26	0.28	0.28	0.34	0.25	0.24	0.28
Kalat	**2.37**	4.07	3.97	3.96	3.92	3.90	4.04	4.04	4.04	4.08	4.03	3.92	4.07	4.04	3.87	4.05
Karachi	2.62	2.31	**2.29**	2.32	2.31	2.30	2.32	2.32	2.34	2.32	2.32	2.78	2.75	2.79	2.80	2.71
Lahore	0.75	0.72	0.72	0.72	0.71	0.71	0.67	0.68	0.67	0.68	0.67	0.81	**0.66**	0.87	0.81	0.86
Multan	**1.31**	1.49	1.49	1.50	1.49	1.50	1.51	1.50	1.49	1.50	1.50	1.46	1.50	1.34	1.54	1.44
Parachinar	**1.38**	2.39	2.35	2.35	2.33	2.34	2.21	2.19	2.17	2.20	2.23	2.22	2.17	2.17	2.15	2.22
Quetta	**1.51**	2.47	2.44	2.45	2.42	2.41	2.39	2.44	2.34	2.38	2.34	2.21	2.22	2.21	2.04	2.15
Rohri	**4.40**	5.14	5.13	5.13	5.13	5.13	4.98	4.97	5.01	4.98	4.98	4.79	4.68	4.76	4.68	4.80
Saidu Sharif	**2.03**	2.90	2.89	2.90	2.89	2.88	2.83	2.84	2.85	2.85	2.84	2.55	2.45	2.49	2.36	2.44
Sialkot	0.92	0.92	0.92	0.92	0.92	0.92	0.94	0.95	0.95	0.95	**0.91**	1.02	1.06	1.02	1.11	1.03
Sibbi	**6.26**	7.04	7.03	7.04	7.03	7.05	6.99	7.02	7.00	7.00	7.05	6.98	7.06	7.00	7.00	7.04
Skardu	**0.53**	1.90	1.87	1.88	1.81	1.81	1.31	1.36	1.47	1.39	1.37	1.11	1.04	0.96	1.02	0.94

Multan at $s = 3$ and $s = 4$. The Fig. 4 has displayed the variants of maximum temperature (Tmax) for each weather station.

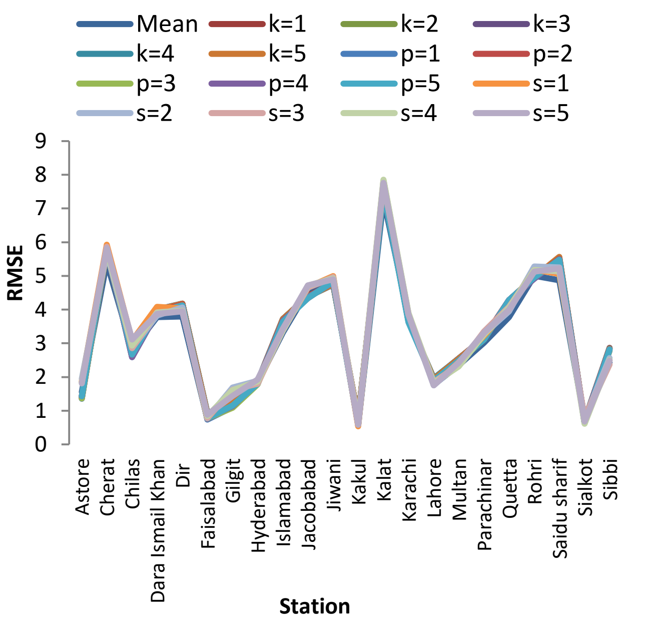

Figure 5. Comparison of different Imputation methods for minimum temperature data at various percentages of missing values based on RMSE.

In case of minimum temperature data, the comparison of four imputation methods is given in Table 6 and display in Fig. 5. The selection of imputation method in each station is based on minimum value of root mean square error (RMSE). The table shows that the mean imputation method is appropriate for Saidu Sharif, Skardu, Astore, Dir, Parachinar, Kakul, Quetta, Multan, Rohri, Cherat, Kalat, Sibbi and Jiwani station The KNN imputation method is suitable for Karachi at $\boldsymbol{k} = \mathbf{2}$ and Jacobabad at $\boldsymbol{k} = \mathbf{3}, \mathbf{4}$ and $\mathbf{5}$. The PMM imputation method is suitable for Faisalabad at $\boldsymbol{p} = \mathbf{1}$, Gilgit at $\boldsymbol{p} = \mathbf{3}$, Chilas at $\boldsymbol{p} = \mathbf{4}$, Sialkot and Islamabad at $\boldsymbol{p} = \mathbf{5}$ while Dara Ismail khan at $\boldsymbol{p} = \mathbf{2}$ and $\boldsymbol{p} = \mathbf{5}$ and Hyderabad at $\boldsymbol{p} = \mathbf{4}$ and $\boldsymbol{p} = \mathbf{5}$. The sample method is only appropriate for Lahore at $\boldsymbol{s} = \mathbf{2}$. From the whole data it can be concluded that for rainfall data, the KNN imputation method is most appropriate while for T_{max} and T_{min} data the mean imputation method is most appropriate as compared to other methods based on RMSE values.

DISCUSSION

The objective of this paper is to estimate missing observations in the chosen weather data of Pakistan. For this purpose, the data of rainfall and temperature (T_{max} and T_{min}) of 23 stations of Pakistan from 1981 to 2020 have been collected. The weather data having missing values can cause bias in estimation due to systematic errors because of missingness.

Therefore, it would be wise approach to estimate the missing values precisely, so that rational conclusions can be made (Alsaber *et al*., 2021). Anderson and Gough (2018) argued that missing data is a common occurrence in climatological time series and its estimation is big challenge for researchers. The main contribution of this study is to find the most appropriate method for estimating the missing observations in the weather data set of Pakistan. Single and multiple imputation methods have been used while comparing the performance of each using the RMSE. The single imputation methods including Mean method and KNN method while the multiple imputation including PMM method and Sample method have been used for estimation purpose. The results showed that KNN method precisely estimate the missing values of rainfall (Aieb *et al.*, 2019). In case of maximum temperature (T_{max}) and minimum temperature (T_{min}) data the mean imputation method is recommend (Twumasi-Ankrah, *et al*., 2019) performed better at different missing rates. The major finding of the research is that selection of appropriate imputation method should be based on the size of missing observation in the weather data but may not be the case all the times (Jadhav *et al*., 2019).

Conclusion**:** Missing values in the time series data is common problem in many fields of study and imputation of the missing values is of great concern in these days. The objective of this paper is to compare four imputation techniques known as mean imputation, KNN imputation, PMM imputation and sample imputation for estimating missing observations in rainfall and temperature data of 23 stations of Pakistan for the years 1981 to 2020. RMSE is used as an accuracy measure to test and evaluate the performance of different imputation techniques. The results have shown that the selection of imputation method is different in each station of rainfall and temperature data due to different missing rates. For rainfall data, the KNN is the most appropriate imputation method while for temperature data mean imputation method is most appropriate as compared to other methods.

Conflicts of interest**:** The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' contribution statements**:** HN and MK collect data and analyze it whereas MIK, MG and MK conceived the idea and supervised the work.

Acknowledgments**:** The authors are thankful to Pakistan Metrological department for providing time series data of rainfall and temperature.

REFERENCES

Abdullah, A. Y. M., M.H. Bhuian, G. Kiselev, A. Dewan, Q.K. Hassan and M. Rafiuddin. 2022. Extreme temperature and rainfall events in Bangladesh: a comparison between coastal and inland areas. International Journal of Climatology 42: 3253-3273.

Afrifa-Yamoah, E., U.A. Mueller, S.M. Taylor and A.J. Fisher. 2020. Missing data imputation of high-resolution temporal climate time series data. Meteorological Applications 27:1-18.

Aieb, A., K. Madani, M. Scarpa. 2019. Bonacorso and K. Lefsih. A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria. Heliyon 55:1-27.

Alsaber, A. R., J. Pan, A. Al-Hurban. 2021. A. Handling complex missing data using random forest approach for an air quality monitoring dataset: a case study of Kuwait environmental data (2012 to 2018). International Journal of Environmental Research and Public Health 18: 1333.

Anderson, C. I. and W. A. Gough. Accounting for missing data in monthly temperature series: Testing rule-of-thumb omission of months with missing values. 2018. International Journal of Climatology 38:4990-5002.

Armitage, E. G., J. Godzien, V. Alonso-Herranz, A. Lopez-Gonzalvez and C. Barbas. 2015. Missing value imputation strategies for metabolomics data. Electrophoresis 36: 3050-3060.

Awan, A.G. and A. Aslam. 2015. Impact of agriculture productivity on economic growth: A case study of Pakistan. Global Journal of Management and Social Sciences 1:57-71.

Azman, M.A.Z. and R. Zakaria 2015. Estimation of missing rainfall data in Pahang using modified spatial interpolation weighting methods. In: AIP Conference Proceedings, Ekaterinburg.pp. 65-72.

Ceglar, A., M. Turco, A. Toreti and F.J. Doblas-Reyes. 2017. Linking crop yield anomalies to large-scale atmospheric circulation in Europe. Agricultural and forest meteorology 240: 35-45.

Eekhout, I., H.C. de Vet, J.W. Twisk, J.P. Brand, M.R. de Boer and M.W. Heymans. 2014. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. Journal of clinical epidemiology 67:335-342.

Egigu, M. L. 2020. Techniques of Filling Missing Values of Daily and Monthly Rain Fall Data: A Review. Journal of Environmental and Earth Science 3:1036.

Fahad, S., and J. Wang. 2020. Climate change, vulnerability, and its impacts in rural Pakistan: a review. Environmental Science and Pollution Research. 27:1334-1338.

Fiala, K., I. Tuma and P. Holub. 2009. Effect of manipulated rainfall on root production and plant belowground dry mass of different grassland ecosystems. Ecosystems 12:906-914.

Gou, J.,T. Xiong and Y. Kuang. 2011. A novel weighted voting for K-nearest neighbor rule. Journal of computers 6:833-840.

Hatfield, J. L. and J.H. Prueger. 2015. Temperature extremes: Effect on plant growth and development. Weather and climate extremes 10:4-10.

Houari, R., A. Bounceur, M.T. Kechadi, A.K. Tari and R. Euler. 2016. Dimensionality reduction in data mining: A Copula approach. Expert Systems with Applications 64:247-260.

Inam, A., J. Adamowski, S. Prasher and R. Albano. 2017. Parameter estimation and uncertainty analysis of the Spatial Agro Hydro Salinity Model (SAHYSMOD) in the semi-arid climate of Rechna Doab, Pakistan. Environmental Modelling and Software 94:186-211.

Jadhav, A., D. Pramod and K. Ramanathan. 2019. Comparison of performance of data imputation methods for numeric dataset. Applied Artificial Intelligence 33:913-933.

Kamble, V.B. and S.N. Deshmukh. 2017. Comparision Between Accuracy and MSE, RMSE by Using Proposed Method with Imputation Technique. Oriental Journal of Computer Science and Technology 10:773-779.

Khan, S. I. and A. S. M. L Hoque. 2020. SICE: an improved missing data imputation technique. Journal of big Data 7:11-21.

Khattak, M.S. and S. Ali. 2015. Assessment of temperature and rainfall trends in Punjab province of Pakistan for the period 1961-2014. Journal of Himalayan Earth Sciences 48:42-61.

Malarvizhi, R. and A.S. Thanamani. 2012. K-nearest neighbor in missing data imputation. International Journal of Engineering Research and Development 5:5-7.

Marui, A., T. Haraguchi, Y. Nakano, T. Funakoshi and K. Yuge. 2002. Irrigation scheduling to control the capillary saltish water movement at crop fields newly reclaimed from the sea. In: 2002 ASAE Annual Meeting, Chicago.pp.1-11.

Mehrotra, D.V., F. Liu and T. Permutt. 2017. Missing data in clinical trials: control-based mean imputation and sensitivity analysis. Pharmaceutical statistics 16:378-392.

Menne, M. J., C.N. Williams, B.E. Gleason, J.J. Rennie and J.H. Lawrimore. 2018. The global historical climatology network monthly temperature dataset, version 4. Journal of Climate 31: 9835-9854.

Mohammed, M.B., H.S. Zulkafil., M.B. Adam., N. Ali and I. A. Baba. 2021. Comparison of five imputation methods in handling missing data in a continuous frequency table. AIP Conference Proceedings.

Presti, L.R., E. Barca, and G. Passarella. 2010. A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). Environmental monitoring and assessment 160:1-22.

Rahman, K.U., S. Shang, M. Shahid and Y. Wen. 2020. An appraisal of dynamic Bayesian model averaging-based merged multi-satellite precipitation datasets over complex topography and the diverse climate of Pakistan. Remote Sensing 12:1-30.

Rahman, M. A., A. Saboor, I.A. Baig, U. Shakoor and H. Kanwal. 2017. An investigation of the impact of climate change on rice crop in Pakistan: a multivariate analysis. Pakistan Journal of Agricultural Sciences 54: 561-566.

Ray, D. K., J.S. Gerber, G.K. MacDonald and P.C. West. 2015. Climate variation explains a third of global crop yield variability. Nature communications 6:1-9.

Schmitt, P., J. Mandel and M. Guedj. 2015. A comparison of six methods for missing data imputation. Journal of Biometrics and Biostatistics 6:1-9.

Saleem, M.U., and S.R Ahmed. 2016. Missing data imputations for upper air temperature at 24 standard pressure levels over Pakistan collected from Aqua satellite. Journal of Data Analysis and Information Processing 4:132-146.

Twumasi-Ankrah, S., B. Odoi, W. Adoma and E.H. Gyamfi. (2019). Efficiency of imputation techniques in univariate time series 8:430-453.

White, I.R.,P. Royston and A.M. Wood.2010. Multiple imputation using chained equations: Issues and guidance for practice. Biostatistics 30:377–399.

Waljee, A.K., A. Mukherjee, A.G. Singal, Y. Zhang, J. Warren, U. Balis, J. Marrero, J. Zhu and P.D.R. Higgins. 2013. Comparison of imputation methods for missing laboratory data in medicine. Gastroenterology and hepatology 3:1-7.

Yang, S. and J.K. Kim. 2018. Predictive mean matching imputation in survey sampling 1:1-2.