


RESEARCH ARTICLE

Comparison of principal component analysis algorithms for imputation in agrometeorological data in high dimension and reduced sample size

Valter Cesar de Souza¹ ^{*}, Sergio Augusto Rodrigues¹ [Ⓞ], Luís Roberto Almeida Gabriel Filho² [Ⓞ]

1 São Paulo State University (Unesp), School of Agriculture, Botucatu, São Paulo, Brasil, **2** São Paulo State University (Unesp), School of Sciences and Engineering, Tupã, São Paulo, Brasil

 These authors contributed equally to this work.

* valter.souza@unesp.br



OPEN ACCESS

Citation: de Souza VC, Rodrigues SA, Filho LAG (2024) Comparison of principal component analysis algorithms for imputation in agrometeorological data in high dimension and reduced sample size. PLoS ONE 19(12): e0315574. <https://doi.org/10.1371/journal.pone.0315574>

Editor: Salim Heddami, University 20 Aout 1955 skikda, ALGERIA

Received: October 8, 2023

Accepted: November 23, 2024

Published: December 31, 2024

Copyright: © 2024 de Souza et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The datasets generated and analyzed during this study are available in the Mendeley Data repository, accessible via the link <https://data.mendeley.com/datasets/2ptckpw94f> or the DOI [10.17632/2ptckpw94f.1](https://doi.org/10.17632/2ptckpw94f.1).

Funding: This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Funding Code 001. The funders had no role in

Abstract

Meteorological data acquired with precision, quality, and reliability are crucial in various agronomy fields, especially in studies related to reference evapotranspiration (ET_o). ET_o plays a fundamental role in the hydrological cycle, irrigation system planning and management, water demand modeling, water stress monitoring, water balance estimation, as well as in hydrological and environmental studies. However, temporal records often encounter issues such as missing measurements. The aim of this study was to evaluate the performance of alternative multivariate procedures for principal component analysis (PCA), using the Nonlinear Iterative Partial Least Squares (NIPALS) and Expectation-Maximization (EM) algorithms, for imputing missing data in time series of meteorological variables. This was carried out on high-dimensional and reduced-sample databases, covering different percentages of missing data. The databases, collected between 2011 and 2021, originated from 45 automatic weather stations in the São Paulo region, Brazil. They were used to create a daily time series of ET_o. Five scenarios of missing data (10%, 20%, 30%, 40%, 50%) were simulated, in which datasets were randomly withdrawn from the ET_o base. Subsequently, imputation was performed using the NIPALS-PCA, EM-PCA, and simple mean imputation (IM) procedures. This cycle was repeated 100 times, and average performance indicators were calculated. Statistical performance evaluation utilized the following indicators: correlation coefficient (*r*), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Normalized Root Mean Square Error (nRMSE), Willmott Index (*d*), and performance index (*c*). In the scenario with 10% missing data, NIPALS-PCA achieved the lowest MAPE (15.4%), followed by EM-PCA (17.0%), while IM recorded a MAPE of 24.7%. In the scenario with 50% missing data, there was a performance reversal, with EM-PCA showing the lowest MAPE (19.1%), followed by NIPALS-PCA (19.9%). The NIPALS-PCA and EM-PCA approaches demonstrated good results in imputation (10% ≤ nRMSE < 20%), with NIPALS-PCA excelling in the 10%, 20%, and 30% scenarios, and EM-PCA in the 40% and 50% scenarios. Based on statistical evaluation, the NIPALS-PCA, EM-PCA, and IM

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

imputation models proved suitable for estimating missing ETo data, with PCA imputation models in the NIPALS and EM algorithms showing the most promise. Future research should explore the effectiveness of various imputation methods in diverse climatic and geographical contexts, as well as develop new techniques considering the temporal and spatial structure of meteorological data, to advance understanding and climate prediction.

Introduction

Measuring evapotranspiration represents a significant challenge in agricultural meteorology [1, 2], mainly due to the high costs associated with direct measurement techniques in terms of implementation, operation and maintenance of measuring equipment [3, 4]. As an alternative, indirect methods are used [5–10] by means of mathematical equations capable of adjusting to local climatic conditions, requiring historical series of meteorological data. However, data recorded over time is generally subject to flaws or errors [11], such as missing measurements, commonly referred to as missings [12].

Addressing missing data is crucial for accurate analysis and decision-making in meteorological studies. Among the various methods available for imputing missing data [13–22], Principal Component Analysis (PCA) has emerged as a versatile and effective tool for exploratory data analysis and characterization of spatial variability [23–25]. Traditional PCA, known for exploratory analysis and dimensionality reduction, can also serve as a viable imputation procedure for missing data [26]. Iterative methods such as the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm [27–33] and the Expectation-Maximization (EM) algorithm [34–38] are commonly employed in conjunction with PCA for imputation.

The application of the NIPALS-PCA and EM-PCA algorithms to the imputation of missing data has shown promising results in various domains. For example, Martí and Zarzo [24] demonstrated the superiority of NIPALS-PCA in imputing reference evapotranspiration data recorded along the Mediterranean coast of Spain compared to methods based on nearest neighbors. Similarly, Josse and Husson [39] introduced the EM-PCA method in the *missMDA* package of the R-Gui computing environment [40], making it easier to reconstruct the data matrix for high-dimensional data sets [23, 41–44].

In this context, even considering the advances, there is still a need for further exploration and evaluation of alternative imputation techniques that employ PCA, particularly with regard to their performance in different domains and datasets [44]. This study aims to fill part of this gap by investigating the performance of alternative multivariate principal component analysis procedures (NIPALS-PCA and EM-PCA) in the imputation of missing data in time series of meteorological variables. Specifically, considering databases in high-dimensional scenarios and reduced sample size, evaluating their performance under different percentages of missing data.

Material and methods

The research was structured in three phases, in which we sought to carry out an extensive literature review on the subject of Missing Value Imputation (MVI), detail an algorithm for how to carry out the simulation steps in MVI and compare alternative procedures for applying principal component analysis techniques in situations of high dimension, reduced sample size and lack of data, by evaluating performance against observed reference evapotranspiration data considering different contexts.

The phases were structured as follows: Phase 1: Understand the panorama of worldwide research on the subject of Missing Value Imputation through a bibliometric analysis; Phase 2: Detail a plan for simulation studies in MVI, focusing on the database and type of data, mechanism and missing rate, imputation technique and performance evaluation method, preparing concepts for the application, used in phase three; Phase 3: Compares the performance of alternative multivariate procedures of principal component analysis in the imputation of missing data in time series of meteorological variables, considering databases in the high-dimensional and reduced-sample scenario, with different rates of missing data.

Fig 1 illustrates the methodological development used by means of a structured framework. It provides an overview of the organization of the phases, emphasizing the deliverables of each phase.

This paper focuses on phase three: comparison of principal component analysis algorithms for imputation in high-dimensional agrometeorological data with a reduced sample size.

Simulation planning

There are some important definitions and steps when planning an MVI simulation study, which are: database and data type, mechanism and fault rate, imputation technique and performance evaluation method. Simulation studies, as is the case in this paper, usually aim to verify the performance of imputation methods, as shown in Fig 2, considering the interactions between application type, data type, mechanism, and fault rate.

Databases

Hourly databases were used, provided by *National Meteorological Institute* (INMET) [45] were used for each meteorological variable, from January 1, 2012, to December 31, 2021, evaluated at 45 automatic weather stations in the region of the State of São Paulo, Brazil.

For each station, the hourly databases covering the period in question were downloaded from the website of the *National Meteorological Institute* in.csv format files, totaling four hundred and fifty files (10 years x 45 stations).

Extracting weather data from INMET

To download weather data from INMET's historical series, several steps are required:

1. Log on to the INMET website: <https://bdmep.inmet.gov.br/>;
2. Choose the annual data package option for all automatic stations separated by year, and you will be taken to the page for annual historical data;
3. Choose the years of interest, among which data is available from the year 2000 onwards. For each year selected, a file in ".csv", comma separated values, format will be available for each station;
4. Choose the stations of interest for the particular survey. To choose the stations of interest, view the geographical distribution of the stations on the map of stations on the link: <https://mapas.inmet.gov.br/>;
5. Rename all the files (.csv), this can be done manually or automatically. The automatic method is preferable due to the number of files to be handled by a data processing routine, for example, for a choice of 45 stations for a period of 10 years, there are 450 files. In order to merge and automatically process the data contained in these files, it is necessary to

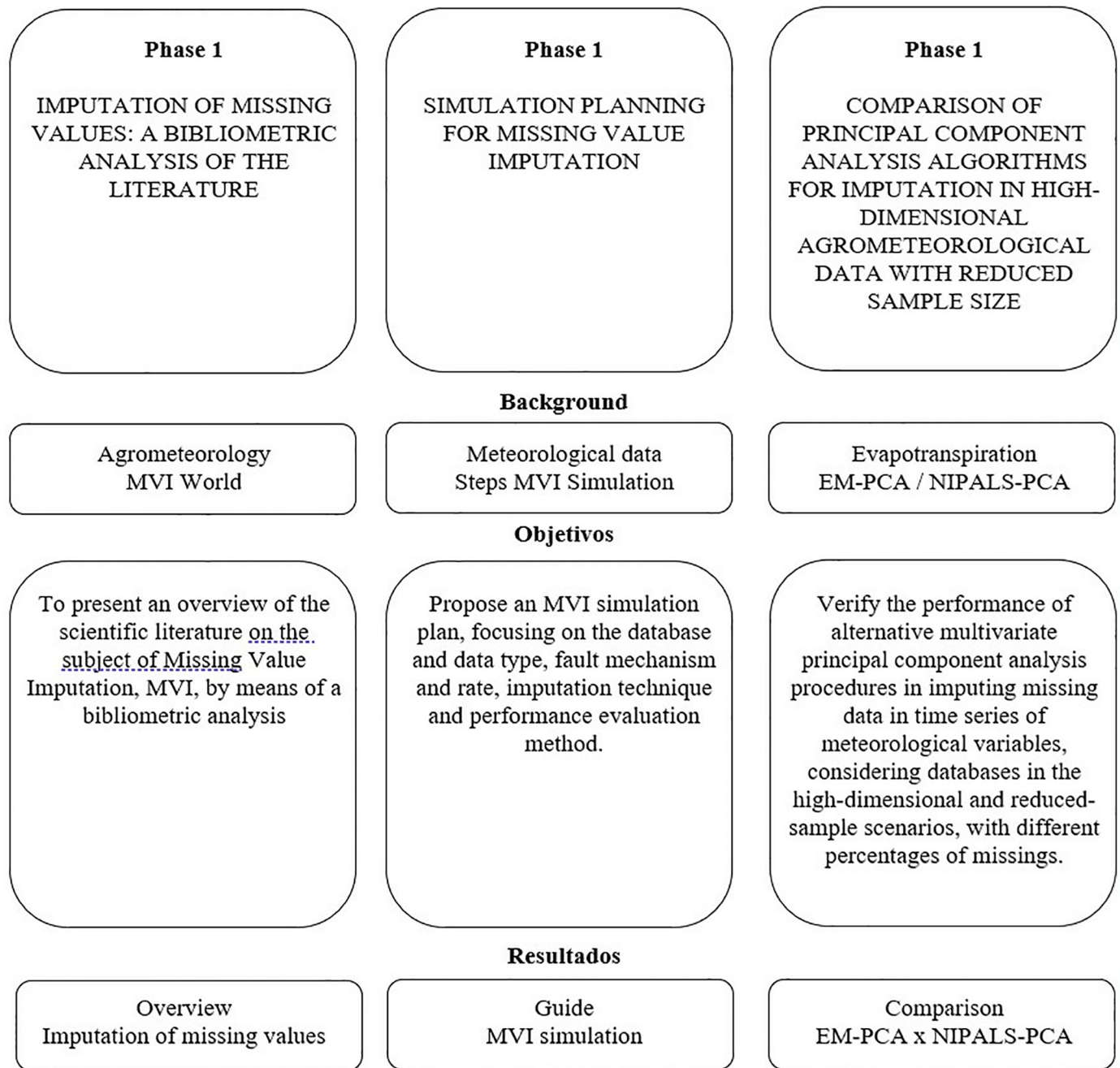


Fig 1. Research methodological framework. Source: Own authorship.

<https://doi.org/10.1371/journal.pone.0315574.g001>

standardize the names. Going from a full name, for example, INMET_SE_SP_A725_A-VARE_01-01-2011_A_31-12-2011 to a shortened name A725_2011;

6. To facilitate the routine reading of these files, create a folder for each station with the spreadsheet files for the years of interest;
7. Create a routine using a script in the **R** environment that automatically reads the data files obtained, considering the following steps:

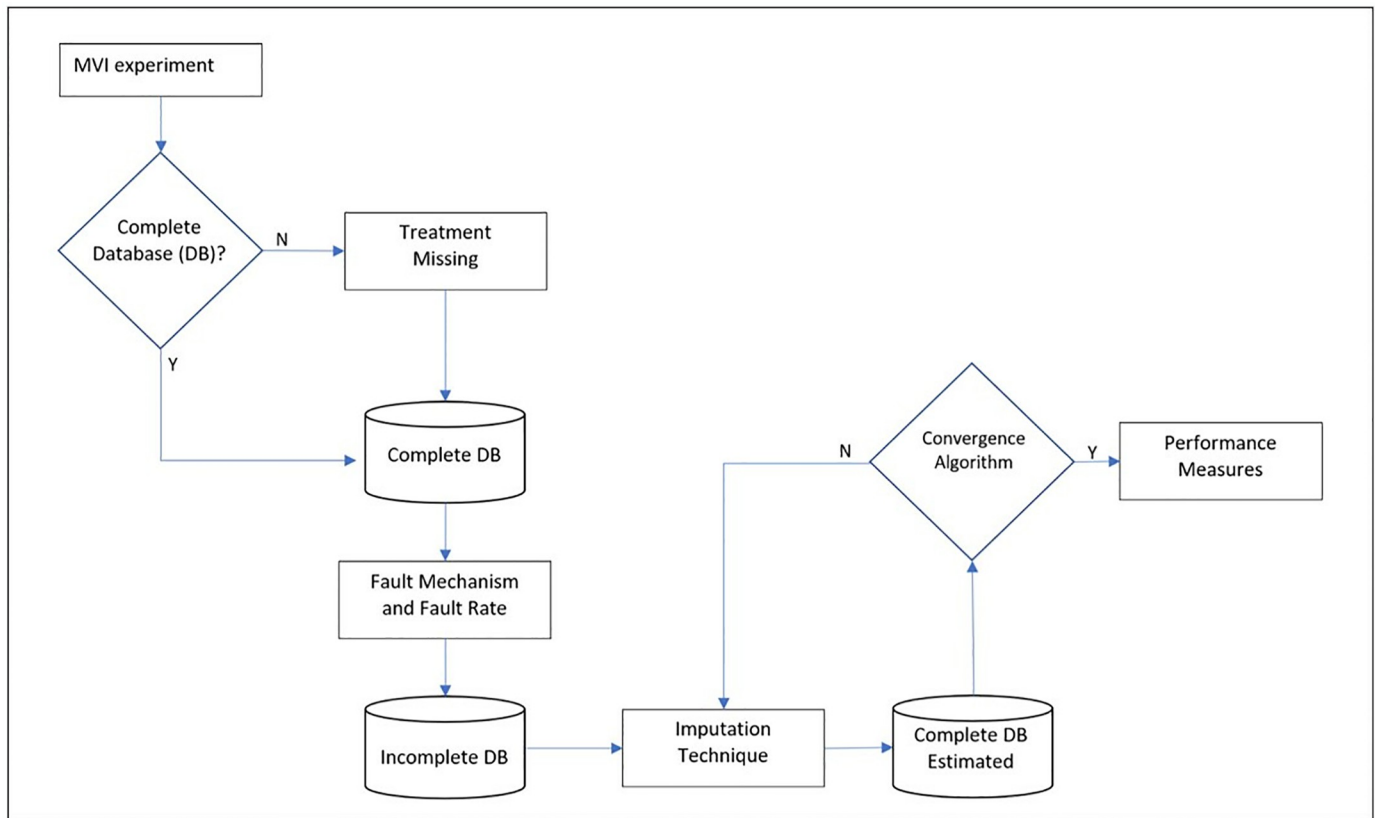


Fig 2. Steps in an MVI experiment. Source: Own authorship.

<https://doi.org/10.1371/journal.pone.0315574.g002>

- Read the files with the data from each station for all the years of the research;
- Exclude the first 9 lines, as they contain information on the weather stations from the research data source (INMET);
- Use common column names for all the databases read into the *R* environment;
- Replace all "-9999" values with "NA";
- Convert the date-time (Greenwich time zone) to local time, with a specific adjustment for São Paulo, subtracting three hours.
- Create a database aggregating all the files;
- Recalculate the variables of interest on a daily basis.

Automatic weather stations

[Table 1](#) provides information on the automatic weather stations used in this research.

Variables of interest

A routine was created in *R* [40] was created to aggregate and generate daily data for the variables of interest: global solar irradiation (R_s , MJ m⁻² hour⁻¹), maximum and minimum air

Table 1. Information on automatic weather stations.

ID	Location	State	Latitude [°]	Longitude [°]	Altitude [m]
1	ARIRANHA	SP	21°7'59"S	48°50'26"W	525,4
2	AVARE	SP	23°6'6"S	48°56'28"W	776,4
3	BARRA BONITA	SP	22°28'16"S	48°33'27"W	533,7
4	BARRA DO TURVO	SP	24°57'46"S	48°24'59"W	659,9
5	BARRETOS	SP	20°33'33"S	48°32'42"W	534,4
6	BARUERI	SP	23°31'26"S	46°52'10"W	776,5
7	BAURU	SP	22°21'29"S	49°1'44"W	636,2
8	JORDAN FIELDS	SP	22°45'1"S	45°36'14"W	1.663,0
9	WHITE HOUSE	SP	21°46'50"S	47°4'31"W	734,2
10	FRANCE	SP	20°35'4"S	47°22'57"W	1.002,8
11	IBITINGA	SP	21°51'20"S	48°47'59"W	496,8
12	IGUAPE	SP	24°40'18"S	47°32'45"W	2,7
13	ITAPEVA	SP	23°58'55"S	48°53'9"W	743,3
14	ITAPIRA	SP	22°24'54"S	46°48'19"W	634,9
15	ITUVERAVA	SP	20°21'35"S	47°46'31"W	610,6
16	JALES	SP	20°9'54"S	50°35'42"W	460,4
17	LINS	SP	21°39'58"S	49°44'5"W	460,7
18	PIRACICABA	SP	22°42'11"S	47°37'24"W	566,5
19	PRADOPOLIS	SP	21°20'18"S	48°6'50"W	540,4
20	PRESIDENT PRUDENTE	SP	22°7'12"S	51°24'31"W	431,9
21	RANCHARIA	SP	22°22'22"S	50°58'29"W	398,8
22	SAO CARLOS	SP	21°58'49"S	47°53'2"W	859,3
23	SAO LUIS PARAITINGA	SP	23°13'42"S	45°25'1"W	862,3
24	SAO MIGUEL ARCANJO	SP	23°51'7"S	48°9'53"W	675,7
25	SAO PAULO—MIRANTE	SP	23°29'47"S	46°37'12"W	785,6
26	SOROCABA	SP	23°25'34"S	47°35'8"W	609,3
27	TAUBATE	SP	23°2'30"S	45°31'15"W	582,3
28	VALPARAISO	SP	21°19'9"S	50°55'49"W	381,9
29	VOTUPORANGA	SP	20°24'12"S	49°57'58"W	510,4
30	PARATY	RJ	23°13'25"S	44°43'37"W	3,0
31	RESENDE	RJ	22°27'5"S	44°26'42"W	438,8
31	RESENDE	RJ	22°27'5"S	44°26'42"W	438,8
32	JAPIRA	PR	23°46'24"S	50°10'50"W	692,9
33	MARINGA	PR	23°24'19"S	51°55'58"W	548,5
34	NEW FATIMA	PR	23°24'55"S	50°34'40"W	664,3
35	PARANAPOEMA	PR	22°39'30"S	52°8'4"W	308,7
36	VENTANIA	PR	24°16'49"S	50°12'37"W	1.093,4
37	AGUA CLARA	MS	20°26'40"S	52°52'33"W	323,6
38	PARANAIBA	MS	19°41'44"S	51°10'54"W	408,1
39	CALDAS	MG	21°55'5"S	46°22'59"W	1.077,3
40	CAMPINA VERDE	MG	19°32'21"S	49°31'5"W	559,1
41	CONCEICAO DAS ALAGOAS	MG	19°59'9"S	48°9'6"W	572,5
42	MONTE VERDE	MG	22°51'42"S	46°2'36"W	1.544,9
43	PASS FOUR	MG	22°23'45"S	44°57'43"W	1.017,1
44	STEPS	MG	20°44'43"S	46°38'2"W	781,7
45	SACRAMENTO	MG	19°52'31"S	47°26'3"W	913,1

Source: Own authorship based on INMET data.

<https://doi.org/10.1371/journal.pone.0315574.t001>

temperatures (T_{max} and T_{min} , °C), maximum and minimum relative humidity (RH_{max} and RH_{min} , %) and wind speed (u_2 , m s⁻¹) measured at a height of 2 meters from the surface.

Reference evapotranspiration

Next, with the daily values of the variables (R_s , T_{max} , T_{min} , UR_{max} , UR_{min} , u_2) and using the Penman-Monteith model [46], recommended by the Food and Agriculture Organization, in bulletin FAO-56, a daily database of reference evapotranspiration (ET_o) was obtained for this region. Naturally, this initial database, a matrix of 45 stations (in rows) by 3653 days (in columns), totaling 164,385 elements, had missing data, about 9.45%, which corresponds to 15,531 missing data in total, which were fully filled in by the *average value of the column* corresponding to the position of the missing data, resulting in a complete database. This complete database was used to verify, using a *script* implemented in the **R** environment, the performance of the NIPALS-PCA algorithms [33, 47] EM-PCA [39] and imputation by the mean of the columns (IM) for filling in missing data, carried out by means of a simulation to evaluate the methods for imputing *missings* in the complete daily ET_o data matrix.

Equação de penman-monteith

The Penman-Monteith model [46] for calculating reference evapotranspiration, ET_o , given by the equation:

$$ET_o = \frac{0,408 \Delta (R_n - G) + \gamma \frac{900}{T_{med} + 273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0,34 u_2)} \quad (1)$$

Where: ET_o —reference evapotranspiration (mm dia⁻¹), R_n —net radiation at the crop surface (MJ m⁻² dia⁻¹), G —soil heat flux density (MJ m⁻² dia⁻¹), T_{med} —mean daily air temperature at 2 m height (°C), u_2 —wind speed at 2 m height (m s⁻¹), e_s —saturation vapour pressure (kPa), e_a —actual vapour pressure (kPa), $e_s - e_a$ —saturation vapour pressure deficit (kPa), Δ —slope vapour pressure curve (kPa °C⁻¹), γ —psychrometric constant (kPa °C⁻¹).

Imputation methods

The treatment of missing data can begin with the decision to eliminate or estimate the missing values [48]. To eliminate missing values, techniques such as complete deletion (listwise deletion) and pairwise deletion are used. In listwise deletion, all cases with at least one missing value are eliminated, which can result in the loss of a lot of data.

In pairwise deletion, only observations with missing values for the variable of interest are excluded, which allows different subsets of data to be used in different analyses, depending on the availability of data for each variable. Although paired exclusion can be more efficient than complete exclusion of cases, it can result in different sample sizes for each analysis and affect the validity of comparisons between variables. These approaches are simple and easy to implement but can result in significant loss of information [49].

Imputation refers to replacing missing data with estimated values. There are several ways in which missing values can be imputed, depending on the nature of the problem and the data. Depending on the nature of the problem, imputation techniques can be broadly classified as basic imputation techniques that do not take time into account are replaced by a constant value, which can be some descriptive measure of position (mean, median or mode) of each column in which the missing values are located [50, 51]. Now for the basic techniques that take time into account, such as time series, there are the techniques of forward fill, back fill and

linear interpolation [52]. Linear interpolation is an imputation technique that assumes a linear relationship between the observed and missing values.

Advanced methods can be classified as multivariate statistical techniques or machine learning. Advanced machine learning imputation techniques use machine learning algorithms to impute missing values in a data set. One such technique is K-nearest neighbor [53], which uses the observed values of the nearest neighbors to replace the missing value. Among the statistical techniques, we highlight the application of principal component analysis (PCA), in conjunction with the NIPALS algorithm [33] and the EM algorithm [38].

Principal component analysis

Principal component analysis (PCA), introduced by Karl Pearson [54] and based on Hotelling [55], aims to reduce the dimension of a data set by explaining the variance and covariance structure of a random vector made up of p random variables, by constructing new variables obtained by linearly combining the original variables. These linear combinations are called principal components and are not correlated with each other [56]. PCA traditionally seeks to find the directions of maximum variance in the data and represents each observation in terms of these directions, principal components. However, in incomplete data sets, conventional PCA requires association with other algorithms, such as NIPALS [33] and EM [38], i.e. NIPALS-PCA and EM-PCA, respectively.

NIPALS-PCA

NIPALS-PCA is an extension of PCA that uses the NIPALS algorithm to find the principal components. The NIPALS algorithm is iterative and calculates the principal components one at a time using the partial least squares technique. This allows it to deal with non-linearities in the data and is especially useful in data sets with high dimensionality or complex correlations between variables. Its ability to deal with these complexities makes it a valuable tool for exploratory analysis and data modeling. This work makes use of the NIPALS-PCA algorithm implemented in the *R-Gui* computing environment, NIPALS package [33].

EM-PCA

The EM-PCA iterative method [57] seeks to minimize the least squares criterion in the observed inputs. Minimization is achieved through an iterative procedure, missing values are replaced by random values. PCA is then applied to the completed data set and the missing values are updated by the fitted values using a predefined number of dimensions. This procedure is repeated until convergence [44]. This method provides estimates for individuals and variables, and an imputation for missing values. An important question concerns the number of dimensions that must be defined at the start of the iterative algorithm. The researchers Josse and Husson [58] suggested methods based on cross-validation to estimate this parameter from an incomplete data set. The method is implemented in the *R-Gui* computing environment, using the `imputePCA` function from the `missMDA` package [23, 39]. Further details can be found in the works [38, 41, 42, 57, 59].

Number of components

The `missMDA` package [39] of the *R-Gui* computing environment provides functions for calculating the number of components (`estim_ncpPCA`) and some imputation methods, including EM-PCA (`imputePCA` function). For NIPALS-PCA, the `nipals` function from the `nipals` package was used. [33] implemented in *R*.

Cross-validation was used to define the number of components to be used in imputation via PCA [23, 60] using the kfold method [61–63]. The percentage of missing values (*pNA*) is removed and estimated with an EM-PCA model using the range of dimensions [*n_{cp.min}*, *n_{cp.max}*]. This process was repeated *n_{bsim}* times. Each cell is estimated using the *imputePCA* function, i.e. using the iterative PCA algorithm (*EM cross-validation*). The number of components resulting in the lowest mean square error was set as the number of components for imputation.

Missing scenarios

Five *miss* scenarios were simulated (10%, 20%, 30%, 40% and 50%), using the mechanism *Missing Completely at Random*, MCAR [64]. To create each *missings* scenario, a data set was randomly generated with some positions taken from the ET database. This procedure begins with the random generation of seeds with the *sample* function of the basic R package, for each specific seed (*set.seed*) the random positions of the *missings* were generated, again with the *sample* command, according to a specific rate of missing data. For all positions, the observed values were replaced by the value "NA", i.e. missing data. For each *missings* scenario specified, imputations were made using the NIPALS-PCA, EM-PCA and IM procedures. This cycle was run 100 times and, at the end, the average performance indicators were calculated.

Statistical performance evaluation

The following indicators were used to assess the statistical performance of the NIPALS-PCA, EM-PCA and IM imputation procedures: correlation coefficient (*r*) [65–67], Mean Absolute Error (MAE) [68–70], Mean Absolute Percentage Error (MAPE) [68], Mean Square Error (MSE) [71–73], Root Mean Square Error (RMSE) [67, 68, 70, 72–75], Normalized Root Mean Square Error (nRMSE) [66, 68], Willmott Index (*d*) [67, 73, 76] e o performance index (*c*) [77]. The indicators can be calculated by Eq (2) through Eq (9):

$$r = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^m (x_i - \bar{x})^2 \sum_{i=1}^m (y_i - \bar{y})^2}} \tag{2}$$

$$MAE = \frac{1}{m} \sum_{i=1}^m |x_i - y_i| \tag{3}$$

$$MAPE = 100 * \frac{1}{m} \sum_{i=1}^m \left| \frac{x_i - y_i}{x_i} \right| \tag{4}$$

$$MSE = \frac{1}{m} \sum_{i=1}^m (x_i - y_i)^2 \tag{5}$$

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (x_i - y_i)^2} \tag{6}$$

$$nRMSE = 100 * \frac{RMSE}{\frac{1}{m} \sum_{i=1}^m x_i} \tag{7}$$

$$d = 1 - \frac{\sum_{i=1}^m (x_i - y_i)^2}{\sum_{i=1}^m (|x_i - \bar{x}| + |y_i - \bar{x}|)^2} \tag{8}$$

$$c = r * d \quad (9)$$

Where “ x_i ” is i -th observed value ($i = 1, \dots, m$), “ \bar{x} ” is average of observed values, “ y_i ” is i -th imputed value ($i = 1, \dots, m$), “ \bar{y} ” is average of imputed values, “ m ” is number of missings.

Results and discussion

In the missing data scenarios (10%, 20%, 30%, 40%, 50%) simulated in the ET database, Principal Component Analysis was used using the NIPALS (with 45 components) and EM (with 5 components) algorithms and simple mean imputation (MI) to reconstruct the database and, consequently, obtain the estimated values of the simulated missings. Seven performance measures (r , MAE , $MAPE$, MSE , $nRMSE$, d and c) were implemented to evaluate the performance of the methods: NIPALS-PCA, EM-PCA and IM.

Dispersion and the correlation coefficient

Figs 3–5 show the dispersion of the ordered pairs corresponding to the imputed values (by NIPALS-PCA, EM-PCA and IM) and the observed ETo values in a typical simulation. When a point coincides with the ideal line, represented by the black curve, this indicates that the

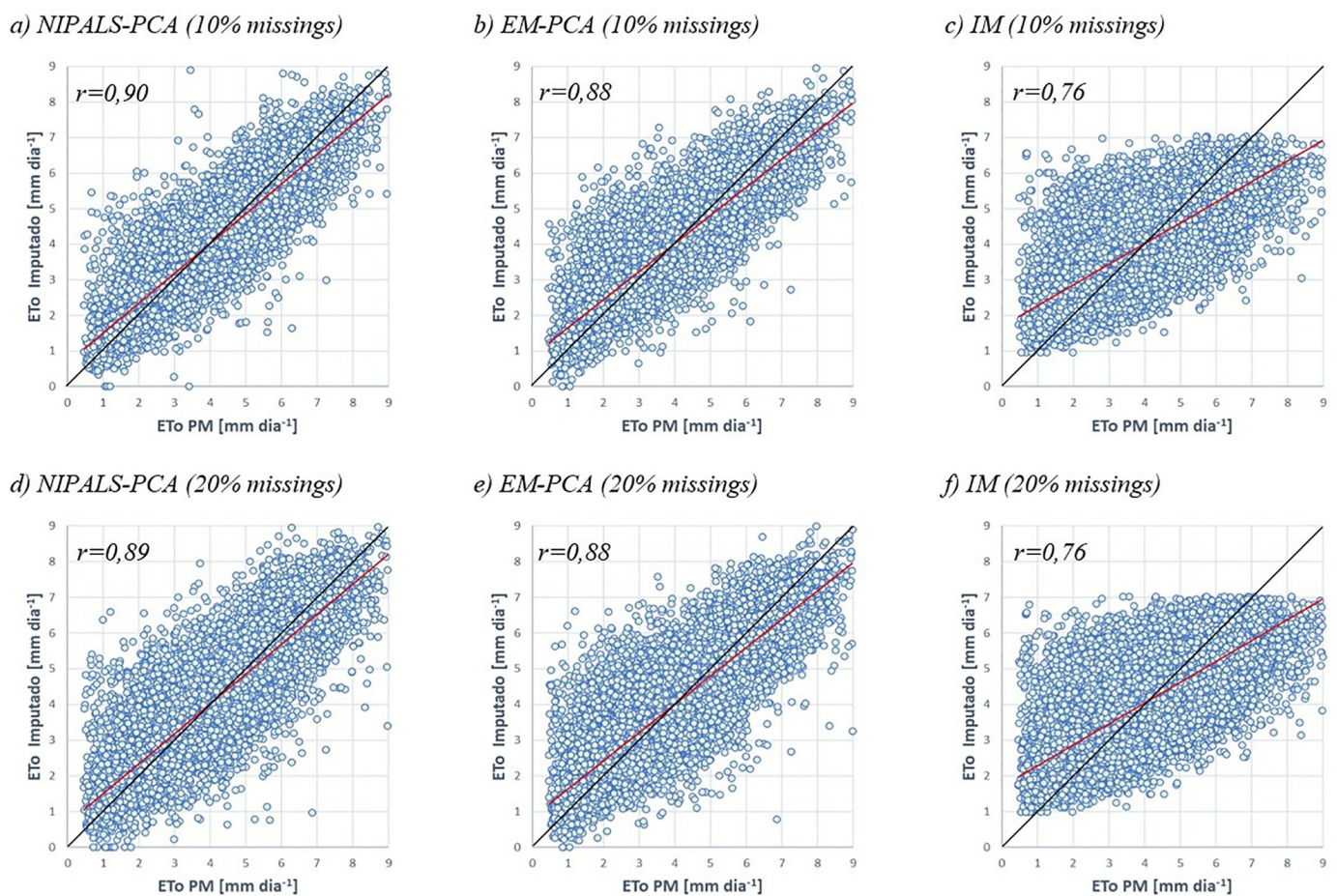


Fig 3. Dispersion between observed ET_o values and those imputed by the NIPALS-PCA, EM-PCA and IM methods (10% and 20% scenarios). Source: Own authorship.

<https://doi.org/10.1371/journal.pone.0315574.g003>

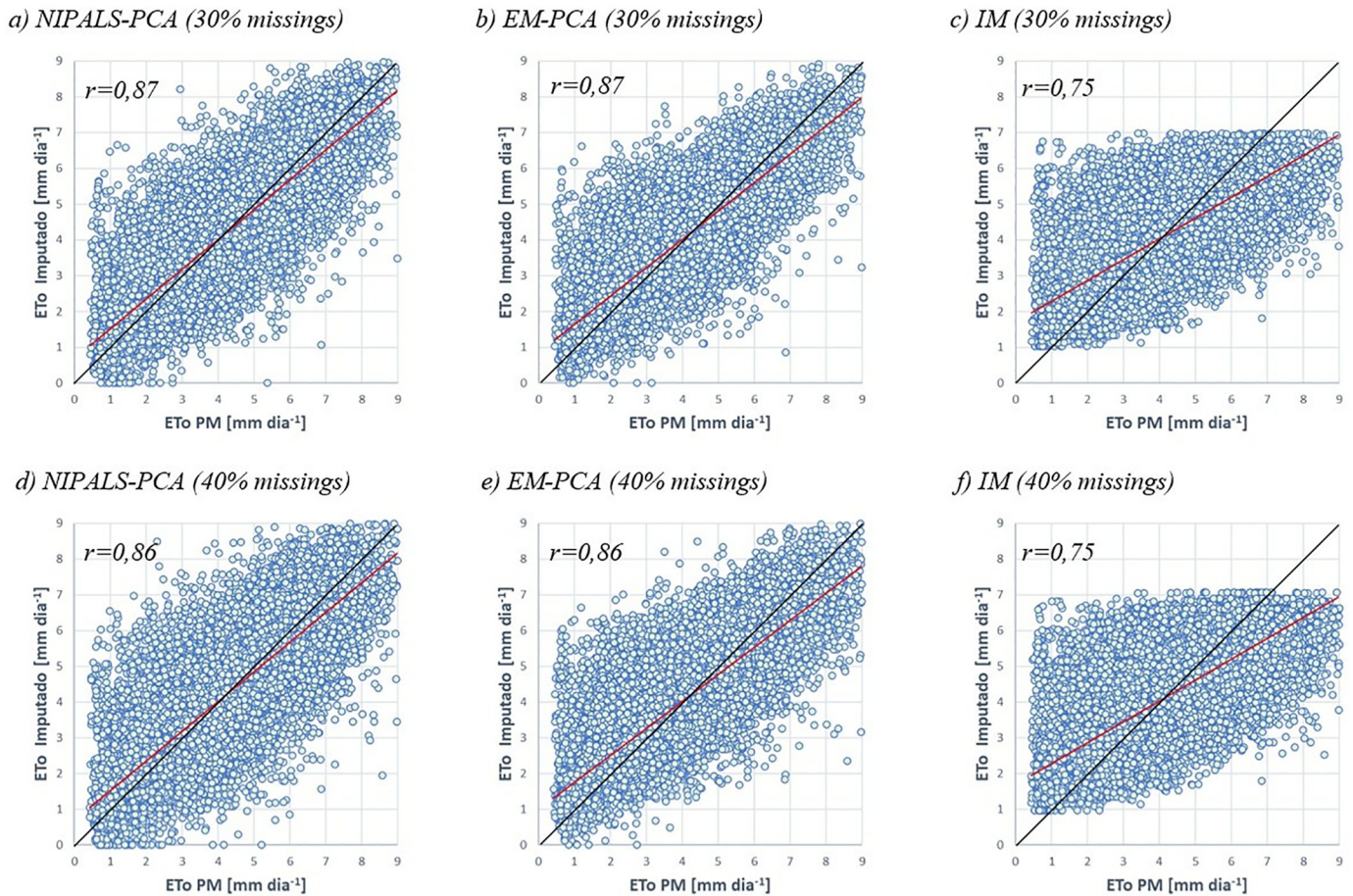


Fig 4. Dispersion between observed ET_0 values and those imputed by the NIPALS-PCA, EM-PCA and IM methods (30% and 40% scenarios). Source: Own authorship.

<https://doi.org/10.1371/journal.pone.0315574.g004>

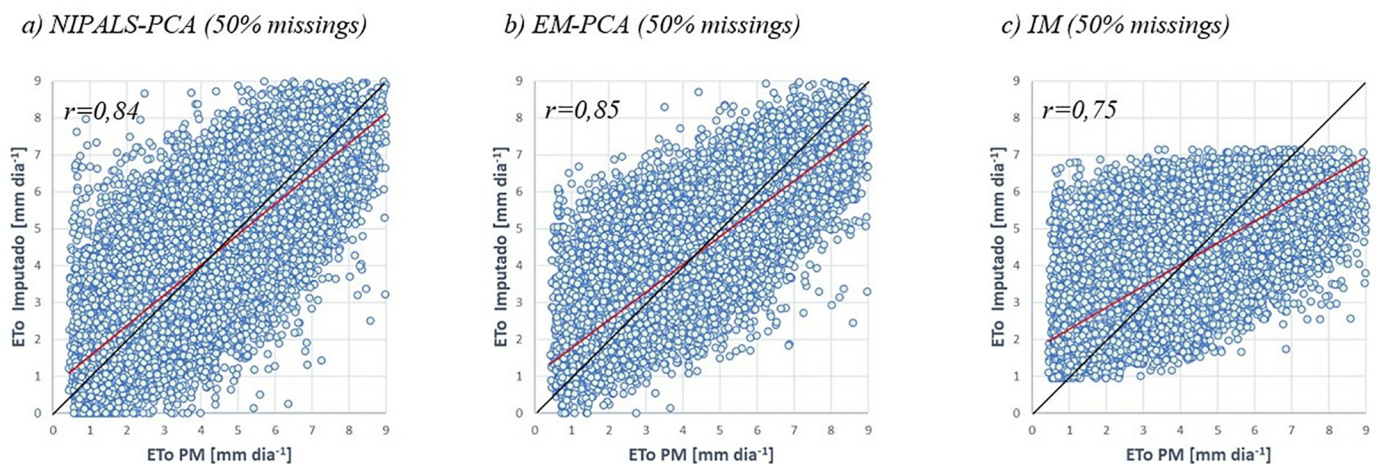


Fig 5. Dispersion between observed ET_0 values and those imputed by the NIPALS-PCA, EM-PCA and IM methods (50% scenario). Source: Own authorship.

<https://doi.org/10.1371/journal.pone.0315574.g005>

estimate and the observed value are corresponding. If the point is above this curve, it means that the imputed value is underestimated; when it is below, it means that it is overestimated. The curve in red represents a linear regression of the cloud of estimated versus observed points, and it is important to note that the deviation from this curve indicates the performance of the model in question. Like the points, when the red line is above the black line, it indicates underestimated estimates, and when it is below, it indicates overestimated estimates. As can be seen in Figs 3–5, the IM imputation method visually shows a greater distance between the ideal and observed lines compared to the NIPALS-PCA and EM-PCA methods.

In Fig 3, two scenarios of missing data are presented, with rates of 10% and 20%. It is observed that the correlation coefficients for the EM-PCA and IM methods remain unchanged, at 0.88 and 0.76, respectively. Conversely, the NIPALS method shows superior results, with correlation coefficients of 0.90 and 0.89 for the 10% and 20% cases, respectively.

In Fig 4, the NIPALS-PCA and EM-PCA methods show similar results, with correlation coefficients of 0.87 and 0.86 for the 30% and 40% scenarios, respectively, while the IM method shows a result of 0.75.

Fig 5 shows the scenario of 50% missing data, estimated by the NIPALS-PCA, EM-PCA and IM methods. In this case, there was an inversion between the methods in terms of correlation coefficient, with values of 0.85 and 0.84 for EM-PCA and NIPALS-PCA, respectively.

Analysis of performance indicators

A descriptive summary of the results obtained by performance indicators in simulations for each scenario and imputation procedure can be observed in Tables 2–5, which include the mean, minimum, and maximum values.

Regarding the correlation coefficient, in Table 2, no significant differences were observed between the missing data scenarios for the IM method. For all missing data scenarios, the correlation coefficient ranged from 0.75 to 0.76 for the IM method, except for 50% missing data, where it was observed that the value practically did not vary between the simulations conducted.

The NIPALS-PCA method has an amplitude of 0.06, which is greater than that of the EM-PCA (0.03). However, in terms of average correlation coefficient, there is practically no difference between the two methods.

The Table 3 presents the results for the MAE and MSE indicators. Once again, it is observed that for the IM method, the results of the MAE and MSE indicators practically do not undergo changes. As for the NIPALS-PCA and EM-PCA methods, there is a reversal in performance: the NIPALS-PCA shows better performance in the scenarios of 10% and 20%, while the EM-PCA performs better in the scenarios of 40% and 50%.

Table 2. Correlation coefficient (10%, 20%, 30%, 40% and 50% scenarios).

Indicators	% <i>missings</i>	NIPALS-PCA	EM-PCA	IM
<i>r</i>	10	0,90 [0,89; 0,91]	0,88 [0,88; 0,89]	0,76 [0,75; 0,76]
	20	0,89 [0,88; 0,89]	0,88 [0,88; 0,89]	0,76 [0,75; 0,76]
	30	0,87 [0,87; 0,88]	0,87 [0,86; 0,88]	0,75 [0,75; 0,76]
	40	0,86 [0,85; 0,86]	0,86 [0,86; 0,86]	0,75 [0,75; 0,76]
	50	0,84 [0,83; 0,84]	0,85 [0,85; 0,86]	0,75 [0,75; 0,75]

Source: Own authorship.

*Average; [Minimum value, Maximum value]

<https://doi.org/10.1371/journal.pone.0315574.t002>

Table 3. MAE and MSE indicators (10%, 20%, 30%, 40% and 50% scenarios).

Indicators	% missings	NIPALS-PCA	EM-PCA	IM
MAE [mm day ⁻¹]	10	0,50 [0,49; 0,51]	0,53 [0,52; 0,54]	0,75 [0,73; 0,76]
	20	0,53 [0,53; 0,54]	0,54 [0,54; 0,55]	0,75 [0,74; 0,76]
	30	0,57 [0,56; 0,57]	0,58 [0,55; 0,59]	0,75 [0,75; 0,76]
	40	0,61 [0,60; 0,61]	0,60 [0,59; 0,60]	0,76 [0,75; 0,76]
	50	0,65 [0,65; 0,66]	0,60 [0,60; 0,61]	0,76 [0,76; 0,77]
MSE [mm ² day ⁻²]	10	0,48 [0,45; 0,51]	0,54 [0,52; 0,56]	1,07 [1,03; 1,11]
	20	0,55 [0,53; 0,58]	0,56 [0,54; 0,58]	1,07 [1,05; 1,10]
	30	0,62 [0,60; 0,64]	0,62 [0,57; 0,66]	1,08 [1,06; 1,10]
	40	0,70 [0,68; 0,72]	0,66 [0,65; 0,67]	1,08 [1,06; 1,11]
	50	0,80 [0,78; 0,83]	0,69 [0,67; 0,69]	1,09 [1,08; 1,11]

Source: Own authorship.

*Average; [Minimum value, Maximum value]

<https://doi.org/10.1371/journal.pone.0315574.t003>

Table 4. MAPE and nRMSE indicators (10%, 20%, 30%, 40% and 50% scenarios).

Indicators	% missings	NIPALS-PCA	EM-PCA	IM
MAPE [%]	10	15,44 [14,95; 15,89]	16,96 [16,53; 17,43]	24,65 [23,90; 25,43]
	20	16,40 [16,04; 16,72]	17,15 [16,83; 17,54]	24,74 [24,26; 25,23]
	30	17,46 [17,16; 17,72]	18,29 [17,18; 18,98]	24,84 [24,42; 25,23]
	40	18,59 [18,29; 18,85]	18,89 [18,63; 19,15]	24,93 [24,60; 25,26]
	50	19,92 [19,65; 20,37]	19,13 [18,91; 19,29]	25,06 [24,76; 25,28]
nRMSE [%]	10	17,16 [16,60; 17,60]	18,18 [17,78; 18,55]	25,50 [25,03; 26,01]
	20	18,27 [17,96; 18,70]	18,41 [18,15; 18,76]	25,55 [25,24; 25,89]
	30	19,41 [19,18; 19,65]	19,46 [18,58; 20,01]	25,61 [25,41; 25,89]
	40	20,62 [20,31; 21,01]	20,03 [19,85; 20,19]	25,67 [25,45; 25,94]
	50	22,07 [21,77; 22,41]	20,30 [20,14; 20,45]	25,77 [25,58; 25,96]

Source: Own authorship.

*Average; [Minimum value, Maximum value]

<https://doi.org/10.1371/journal.pone.0315574.t004>

Table 5. Willmott and performance indicators (10%, 20%, 30%, 40% and 50% scenarios).

Indicators	% missings	NIPALS-PCA	EM-PCA	IM
d	10	0,95 [0,94; 0,95]	0,94 [0,93; 0,94]	0,85 [0,84; 0,86]
	20	0,94 [0,94; 0,94]	0,94 [0,93; 0,94]	0,85 [0,85; 0,86]
	30	0,93 [0,93; 0,93]	0,93 [0,92; 0,93]	0,85 [0,85; 0,85]
	40	0,92 [0,92; 0,93]	0,92 [0,92; 0,92]	0,85 [0,85; 0,85]
	50	0,91 [0,91; 0,92]	0,92 [0,92; 0,92]	0,85 [0,85; 0,85]
c	10	0,85 [0,84; 0,86]	0,83 [0,82; 0,84]	0,65 [0,64; 0,66]
	20	0,83 [0,82; 0,84]	0,82 [0,82; 0,83]	0,64 [0,64; 0,65]
	30	0,81 [0,81; 0,82]	0,80 [0,79; 0,82]	0,64 [0,64; 0,65]
	40	0,79 [0,78; 0,80]	0,79 [0,79; 0,79]	0,64 [0,64; 0,65]
	50	0,77 [0,76; 0,77]	0,79 [0,78; 0,79]	0,64 [0,64; 0,64]

Source: Own authorship.

*Average; [Minimum value, Maximum value]

<https://doi.org/10.1371/journal.pone.0315574.t005>

In Table 4, the results for the MAPE and nRMSE indicators are presented. It is observed that the NIPALS-PCA method outperforms the EM-PCA for the scenarios of 10%, 20%, and 30%. However, there is a performance reversal in the 50% scenario. A gradual increase in MAPE and nRMSE was observed as a greater number of missing values were added, indicating a deterioration in the performance of the procedures considered. For the scenario with 10% of missing values, NIPALS-PCA obtained the lowest MAPE (15.44%), followed by EM-PCA (16.96%), while IM obtained a MAPE equal to 24.65%. In the scenario with 50% of missing values, there is a performance reversal, with a lower MAPE (19.13%) for EM-PCA, followed by NIPALS-PCA (19.92%).

Considering the classification scale for the different nRMSE intervals, the NIPALS-PCA and EM-PCA approaches present good results ($10\% \leq \text{nRMSE} < 20\%$) in the imputation of missing values. Particularly noteworthy is the NIPALS-PCA method for the scenarios of 10%, 20%, and 30%, and the EM-PCA for the scenarios of 40% and 50%.

In Table 5, the results for the performance indices (c) and Willmott's agreement index (d) are presented. It is observed that, for Willmott's agreement index (d), the NIPALS-PCA and EM-PCA methods stand out, on average, with an agreement index of 93%, compared to 85% for the IM method. Regarding the confidence coefficient (c), the NIPALS-PCA and EM-PCA methods present an average value of 0.81, classified as "very good" estimation models according to the classification provided by researchers Camargo and Sentelhas [77]. On the other hand, the IM method showed an average value of 0.64, classified as a "good" estimation model.

Fig 6 presents the results for the MAPE indicator for the scenarios of 10% and 50%. It can be observed that the EM-PCA and NIPALS-PCA methods are similar, while the IM method deviates, showing results with higher deviations.

Research comparison

Compared to the results obtained by researchers Martí and Zarzo [24] modeling 30 weather stations located in the Valencia region of Spain, from 2000 to 2007, we see lower results in the

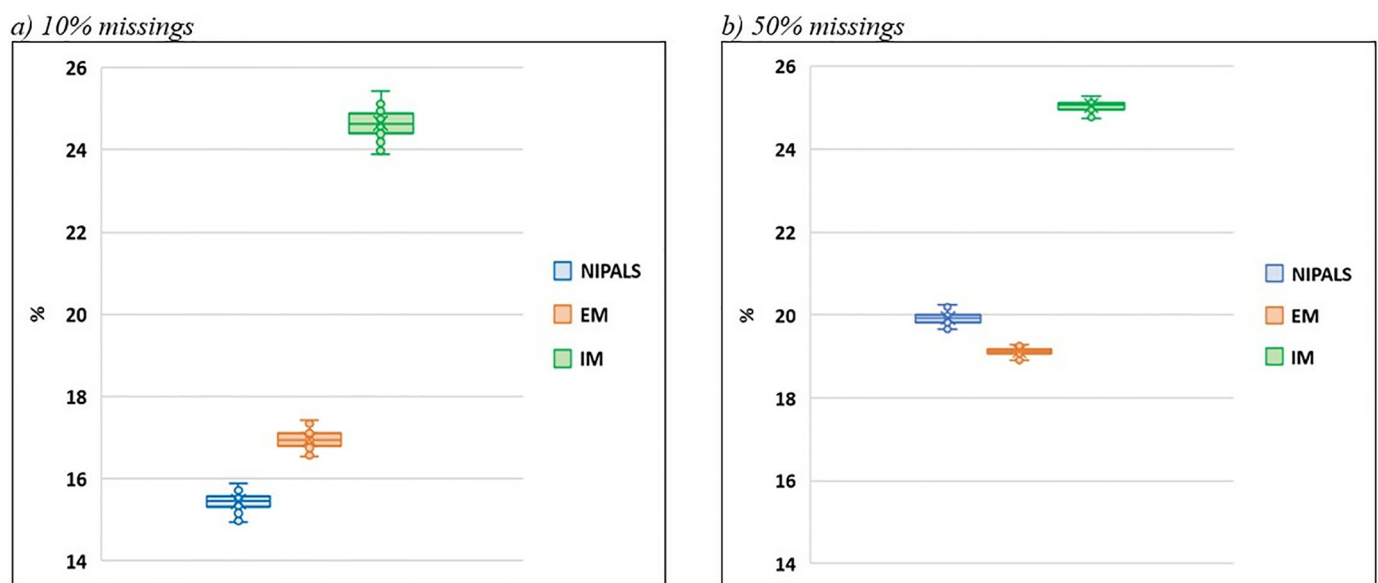


Fig 6. MAPE indicator (10% and 50% scenarios). Source: Own authorship.

<https://doi.org/10.1371/journal.pone.0315574.g006>

Table 6. Values of some statistical performance indicators.

Researchers	Location	MSE mm ² day ⁻²	MAE mm day ⁻¹	pMAE %	r
Martí and Zarzo (2012)	Valencia—Spain	0,11	0,24	9,20	0,98
Present research	São Paulo—Brazil	0,48	0,50	15,44	0,90

Source: Own authorship.

<https://doi.org/10.1371/journal.pone.0315574.t006>

performance indicators than those found in this research, as shown in [Table 6](#), for the 10% *missings* scenario.

Researchers Dray and Josse [44] review some PCA imputation methods applied to data in the field of ecology. They suggest using EM-PCA rather than NIPALS-PCA, due to the difficulty of convergence. This was not observed in this study.

Research limitations

It is important to note that the simulations carried out in this study used the *Missing Completely at Random* (MCAR) mechanism and, therefore, the results presented may not be generalizable to situations in which the missing values occurred in a non-random or biased manner. In addition, this study used the mean to complete the initial base, favoring this method in the simulations carried out and, therefore, the differences between the performance of the multivariate methods via PCA in relation to imputation by the mean may be greater.

Conclusions

This study examined the performance of alternative multivariate principal component analysis procedures using NIPALS and EM algorithms, along with simple mean imputation (IM), for reconstructing a high-dimensional, small-sample reference evapotranspiration database. Consequently, estimated values for simulated missing data were obtained under scenarios of 10%, 20%, 30%, 40%, and 50% missing data. The study spanned from 2012 to 2021 and focused on automatic weather stations in the São Paulo region, Brazil. Results underscored the importance of choosing the right imputation approach, with significant implications for the accuracy of climate estimates. PCA proved to be a useful tool for estimating missing values, particularly when the sample size was small relative to the number of variables. This study focused on imputing missing data in an evapotranspiration database, considering ET_o measurement days as correlated variables (3653 columns) measured across 45 automatic weather stations (rows). Statistical performance comparison among the techniques revealed that NIPALS-PCA and EM-PCA outperformed IM, depending on the percentage of missing data. Based on the statistical indicator classification of the validation base for NIPALS-PCA, EM-PCA, and IM imputation models, there are indications that they are suitable for estimating missing reference evapotranspiration values, with particular emphasis on PCA imputation models in the NIPALS and EM algorithms. For future work, exploring the effectiveness of different imputation methods across various climatic and geographic contexts is recommended. Investigations into the development of new imputation techniques, especially those considering the temporal and spatial structure of meteorological data, are essential for advancing understanding and forecasting capacity in climatology. In summary, this study provides a solid foundation for future research on imputation strategies for missing meteorological data, with the potential to significantly improve the accuracy and utility of climate estimates in various applications.

Author Contributions

Conceptualization: Valter Cesar de Souza.

Data curation: Valter Cesar de Souza.

Formal analysis: Valter Cesar de Souza.

Funding acquisition: Valter Cesar de Souza.

Investigation: Valter Cesar de Souza.

Methodology: Valter Cesar de Souza.

Project administration: Sergio Augusto Rodrigues, Luís Roberto Almeida Gabriel Filho.

Resources: Valter Cesar de Souza.

Software: Valter Cesar de Souza.

Supervision: Sergio Augusto Rodrigues, Luís Roberto Almeida Gabriel Filho.

Validation: Valter Cesar de Souza.

Visualization: Valter Cesar de Souza.

Writing – original draft: Valter Cesar de Souza.

Writing – review & editing: Sergio Augusto Rodrigues, Luís Roberto Almeida Gabriel Filho.

References

1. Mikaeili O, Shourian M. Improving Evapotranspiration Estimation in SWAT-Based Hydrologic Simulation through Data Assimilation in the SEBAL Algorithm. *Water Resources Management*. 2024; 1–22. <https://doi.org/10.1007/S11269-024-03854-4/METRICS>
2. Abbaspour K, Rouholahnejad E, . . . SV-J of, 2015 undefined. A continental-scale hydrology and water quality model for Europe: Calibration and uncertainty of a high-resolution large-scale SWAT model. ElsevierKC Abbaspour, E Rouholahnejad, S Vaghefi, R Srinivasan, H Yang, B Kløve *Journal of hydrology*, 2015•Elsevier. [cited 18 Apr 2024]. <https://www.sciencedirect.com/science/article/pii/S0022169415001985>
3. Rana G, Katerji N. Measurement and estimation of actual evapotranspiration in the field under Mediterranean climate: a review. *European Journal of Agronomy*. 2000; 13: 125–153. [https://doi.org/10.1016/S1161-0301\(00\)00070-8](https://doi.org/10.1016/S1161-0301(00)00070-8)
4. Allen RG, Pereira LS, Howell TA, Jensen ME. Evapotranspiration information reporting: I. Factors governing measurement accuracy. *Agric Water Manag*. 2011; 98: 899–920. <https://doi.org/10.1016/J.AGWAT.2010.12.015>
5. Onnabi Milani A, Hossein Zad Derakhshan A, Khodaverdizadeh gahramani M, Chitsaz Moghaddam S, Pashaei S. Evaluating direct and indirect estimation methods of reference evapotranspiration (ET_o). 2007 [cited 26 Dec 2022].
6. Faseyiku OO, Obinna Obiora-Okeke A, Ayodeji Olowoselu S, Oluwatosin , et al. Validation of selected gridded potential evapotranspiration datasets with ground-based observations over the Ogun-Osun River Basin, Nigeria. *Arabian Journal of Geosciences* 2024 17:5. 2024; 17: 1–16. <https://doi.org/10.1007/S12517-024-11962-Z>
7. Islam S, Heliyon AA-, 2021 undefined. Performance evaluation of FAO Penman-Monteith and best alternative models for estimating reference evapotranspiration in Bangladesh. *cell.comS Islam, AKMR AlamHeliyon*, 2021•cell.com. 2017; e07487.
8. Abeyisiriwardana H, Muttill N, Hydrology UR-, 2022 undefined. A comparative study of potential evapotranspiration estimation by three methods with FAO Penman—Monteith method across Sri Lanka. *mdpi.comHD Abeyisiriwardana, N Muttill, U RathnayakeHydrology*, 2022•mdpi.com. [cited 18 Apr 2024]. <https://www.mdpi.com/2306-5338/9/11/206>
9. Satpathi A, Danodia A, Nain AS, Dhyani M, Vishwakarma DK, Dewidar AZ, et al. Estimation of crop evapotranspiration using statistical and machine learning techniques with limited meteorological data: a case study in Udham Singh Nagar, India. *Theor Appl Climatol*. 2024; 1–18. <https://doi.org/10.1007/S00704-024-04953-3/FIGURES/7>

10. Fang S-L, Lin Y-S, Chang S-C, Chang Y-L, Tsai B-Y, Kuo B-J. Using Artificial Intelligence Algorithms to Estimate and Short-Term Forecast the Daily Reference Evapotranspiration with Limited Meteorological Variables. *Agriculture* 2024, Vol 14, Page 510. 2024; 14: 510. <https://doi.org/10.3390/AGRICULTURE14040510>
11. Hasan MK, Alam MA, Roy S, Dutta A, Jawad MT, Das S. Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Inform Med Unlocked*. 2021; 27: 1–23. <https://doi.org/10.1016/J.IMU.2021.100799>
12. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Networks*. 1989; 2: 359–366. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)
13. Santos JEO, Cunha FF da, Filgueiras R, Silva GH da, Castro Teixeira AH de, Santos Silva FC dos, et al. Performance of SAFER evapotranspiration using missing meteorological data. *Agric Water Manag*. 2020; 233: 106076. <https://doi.org/10.1016/J.AGWAT.2020.106076>
14. Abrishami N, Sepaskhah AR, Shahrokhnia MH. Estimating wheat and maize daily evapotranspiration using artificial neural network. *Theor Appl Climatol*. 2019; 135: 945–958. <https://doi.org/10.1007/S00704-018-2418-4>
15. Hashemi M, Sepaskhah AR. Evaluation of artificial neural network and Penman—Monteith equation for the prediction of barley standard evapotranspiration in a semi-arid region. *Theor Appl Climatol*. 2020; 139: 275–285. <https://doi.org/10.1007/S00704-019-02966-X>
16. Pagano A, Amato F, Ippolito M, De Caro D, Croce D, Motisi A, et al. Machine learning models to predict daily actual evapotranspiration of citrus orchards under regulated deficit irrigation. *Ecol Inform*. 2023; 76: 102133. <https://doi.org/10.1016/J.ECOINF.2023.102133>
17. De Caro D, Ippolito M, Cannarozzo M, Provenzano G, Ciraolo G. Assessing the performance of the Gaussian Process Regression algorithm to fill gaps in the time-series of daily actual evapotranspiration of different crops in temperate and continental zones using ground and remotely sensed data. *Agric Water Manag*. 2023; 290: 108596. <https://doi.org/10.1016/J.AGWAT.2023.108596>
18. Huang H, Song Y, Fan Z, Xu G, Yuan R, Zhao J. Estimation of walnut crop evapotranspiration under different micro-irrigation techniques in arid zones based on deep learning sequence models. *Results in Applied Mathematics*. 2023; 20: 100412. <https://doi.org/10.1016/J.RINAM.2023.100412>
19. Sentelhas PC, Gillespie TJ, Santos EA. Evaluation of FAO Penman—Monteith and alternative methods for estimating reference evapotranspiration with missing data in Southern Ontario, Canada. *Agric Water Manag*. 2010; 97: 635–644. <https://doi.org/10.1016/J.AGWAT.2009.12.001>
20. Mhaweji M, Caiserman A, Nasrallah A, Dawi A, Bachour R, Faour G. Automated evapotranspiration retrieval model with missing soil-related datasets: The proposal of SEBALI. *Agric Water Manag*. 2020; 229: 105938. <https://doi.org/10.1016/J.AGWAT.2019.105938>
21. Chen S, He C, Huang Z, Xu X, Jiang T, He Z, et al. Using support vector machine to deal with the missing of solar radiation data in daily reference evapotranspiration estimation in China. *Agric For Meteorol*. 2022; 316: 108864. <https://doi.org/10.1016/J.AGRFORMET.2022.108864>
22. Karimi S, Shiri J, Marti P. Supplanting missing climatic inputs in classical and random forest models for estimating reference evapotranspiration in humid coastal areas of Iran. *Comput Electron Agric*. 2020; 176: 105633. <https://doi.org/10.1016/J.COMPAG.2020.105633>
23. Josse J, Husson F. Selecting the number of components in principal component analysis using cross-validation approximations. *Comput Stat Data Anal*. 2012; 56: 1869–1879. <https://doi.org/10.1016/j.csda.2011.11.012>
24. Martí P, Zarzo M. Multivariate statistical monitoring of ETo: A new approach for estimation in nearby locations using geographical inputs. *Agric For Meteorol*. 2012; 152: 125–134. <https://doi.org/10.1016/j.agrformet.2011.08.008>
25. García-Diego FJ, Zarzo M. Microclimate monitoring by multivariate statistical control: The renaissance frescoes of the Cathedral of Valencia (Spain). *J Cult Herit*. 2010; 11: 339–344. <https://doi.org/10.1016/j.culher.2009.06.002>
26. De Ketelaere B, Hubert M, Schmitt E. Overview of PCA-Based Statistical Process-Monitoring Methods for Time-Dependent, High-Dimensional Data. 2017; 47: 318–335. <https://doi.org/10.1080/00224065.2015.11918137>
27. Howley T, Madden MG, O'Connell ML, Ryder AG. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowl Based Syst*. 2006; 19: 363–370. <https://doi.org/10.1016/J.KNOSYS.2005.11.014>
28. de la Fuente RLN, García-Muñoz S, Biegler LT. An efficient nonlinear programming strategy for PCA models with incomplete data sets. *J Chemom*. 2010; 24: 301–311. <https://doi.org/10.1002/CEM.1306>

29. Eshghi P. Dimensionality choice in principal components analysis via cross-validators. *Chemometrics and Intelligent Laboratory Systems*. 2014; 130: 6–13. <https://doi.org/10.1016/J.CHEMOLAB.2013.09.004>
30. Yang Q, Zhang L, Wang L, Xiao H. MultiDA: Chemometric software for multivariate data analysis based on Matlab. *Chemometrics and Intelligent Laboratory Systems*. 2012; 116: 1–8. <https://doi.org/10.1016/J.CHEMOLAB.2012.03.019>
31. Patel N, Sivanathan K, Mhaskar P. Polymethyl Methacrylate Quality Modeling with Missing Data Using Subspace Based Model Identification. *Processes* 2021, Vol 9, Page 1691. 2021; 9: 1691. <https://doi.org/10.3390/PR9101691>
32. Vyas M, Pareek K, Spare S, Garg A, Gao L. State-of-charge prediction of lithium ion battery through multivariate adaptive recursive spline and principal component analysis. *Energy Storage*. 2021; 3: e147. <https://doi.org/10.1002/EST2.147>
33. Wright K. The NIPALS algorithm. 27 Oct 2017 [cited 13 Dec 2022]. https://cran.r-project.org/web/packages/nipals/vignettes/nipals_algorithm.html
34. Nilashi M, Abumalloh RA, Yusuf SYM, Thi HH, Alsulami M, Abosaq H, et al. Early diagnosis of Parkinson's disease: A combined method using deep learning and neuro-fuzzy techniques. *Comput Biol Chem*. 2023; 102: 107788. <https://doi.org/10.1016/j.compbiolchem.2022.107788> PMID: 36410240
35. Malan L, Smuts CM, Baumgartner J, Ricci C. Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns. *Nutrition Research*. 2020; 75: 67–76. <https://doi.org/10.1016/j.nutres.2020.01.001> PMID: 32035304
36. Bucior-Kwaczyńska A. The Possibility of Applying the EM-PCA Procedure to Lake Water. *Pol J Environ Stud*. 2018; 27: 19–30. <https://doi.org/10.15244/PJOES/74367>
37. Xie C, Bi S, Dong M, Li Y. Recovery Method for Missing Sensor Data in Multi-Sensor Based Walking Recognition System. 8th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, CYBER 2018. 2019; 558–563.
38. Dempster A. P.; Laird N. M.; Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*. 1977; 39: 1–38. <https://doi.org/10.1111/1.3424485>
39. Josse J, Husson F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *J Stat Softw*. 2016; 70: 1–31. <https://doi.org/10.18637/JSS.V070.I01>
40. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. <https://www.r-project.org/>
41. Josse J, Pagès J, Husson F. Multiple imputation in principal component analysis. *Adv Data Anal Classif*. 2011; 5: 231–246. <https://doi.org/10.1007/S11634-011-0086-7>
42. Josse J, Husson F, Pagès J. Gestion des données manquantes en Analyse en Composantes Principales. *Journal de la Société Française de Statistique*. 2009; 150: 2.
43. Podani J, Kalapos T, Barta B, Schmera D. Principal component analysis of incomplete data—A simple solution to an old problem. *Ecol Inform*. 2021; 61: 101235. <https://doi.org/10.1016/J.ECOINF.2021.101235>
44. Dray S, Josse J. Principal component analysis with missing values: a comparative survey of methods. *Plant Ecol*. 2014; 216: 657–667. <https://doi.org/10.1007/S11258-014-0406-Z>
45. INMET. Instituto Nacional de Meteorologia. In: Ministério da Agricultura, Pecuária e Abastecimento [Internet]. 2022 [cited 26 Dec 2022]. <https://portal.inmet.gov.br/servicos/bdmep-dados-historicos>
46. Allen RG, Pereira LS, Raes D, Smith M. Crop Evapotranspiration - Guidelines for computing crop water requirements. FAO Irrigation and drainage. 1998. <https://doi.org/10.3390/agronomy9100614>
47. Andrecut M. Parallel GPU implementation of iterative PCA algorithms. *J Comput Biol*. 2009; 16: 1593–1599. <https://doi.org/10.1089/cmb.2008.0221> PMID: 19772385
48. Pandey P. A Guide to Handling Missing values in Python. In: kaggle [Internet]. 2020 [cited 2 Feb 2023]. <https://www.kaggle.com/code/parulpandey/a-guide-to-handling-missing-values-in-python>
49. van Buuren S. Flexible Imputation of Missing Data. 2nd Editio. New York: Chapman and Hall/CRC; 2018. <https://www.routledge.com/Flexible-Imputation-of-Missing-Data-Second-Edition/Buuren/p/book/9781032178639>
50. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, et al. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports* 2018 8:1. 2018; 8: 1–10. <https://doi.org/10.1038/s41598-017-19120-0> PMID: 29330539
51. Magnani M. Techniques for dealing with missing data in knowledge discovery tasks. 2004 [cited 2 Feb 2023]. <https://www.researchgate.net/publication/228748415>

52. Nguyen M, He T, An L, Alexander DC, Feng J, Yeo BTT. Predicting Alzheimer's disease progression using deep recurrent neural networks. *Neuroimage*. 2020; 222: 117203. <https://doi.org/10.1016/j.neuroimage.2020.117203> PMID: 32763427
53. Patil BM, Joshi RC, Toshniwal D. Missing value imputation based on k-mean clustering with weighted distance. *Communications in Computer and Information Science*. 2010; 94 CCIS: 600–609. https://doi.org/10.1007/978-3-642-14834-7_56/COVER
54. Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*. 1901; 2: 559–572.
55. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol*. 1933; 24: 417–441. <https://doi.org/10.1037/h0071325>
56. Mingoti SA. Análise de dados através de Métodos de Estatística Multivariada: Uma abordagem Aplicada. Belo Horizonte: Editora UFMG; 2005.
57. Kiers HAL. Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*. 1997; 62: 251–266. <https://doi.org/10.1007/BF02295279/METRICS>
58. Josse J, Husson F. Handling missing values in exploratory multivariate data analysis methods. *Journal de la société française de statistique*. 2012; 153: 79–99.
59. Schafer JL. *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall/CRC; 1997.
60. Bro R, Kjeldahl K, Smilde AK, Kiers HAL. Cross-validation of component models: A critical look at current methods. *Anal Bioanal Chem*. 2008; 390: 1241–1251. <https://doi.org/10.1007/s00216-007-1790-1> PMID: 18214448
61. Fushiki T. Estimation of prediction error by using K-fold cross-validation. *Stat Comput*. 2011; 21: 137–146. <https://doi.org/10.1007/S11222-009-9153-8/METRICS>
62. Moreno-Torres JG, Saez JA, Herrera F. Study on the impact of partition-induced dataset shift on k-fold cross-validation. *IEEE Trans Neural Netw Learn Syst*. 2012; 23: 1304–1312. <https://doi.org/10.1109/TNNLS.2012.2199516> PMID: 24807526
63. Jung Y. Multiple predicting K-fold cross-validation for model selection. 2017; 30: 197–215. <https://doi.org/10.1080/10485252.2017.1404598>
64. Little RJA, Rubin DB. *Single Imputation Methods*. John Wiley & Sons, Ltd; 2002.
65. Pearson K. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character*. 1896; 187: 253–318. <https://doi.org/10.1098/RSTA.1896.0007>
66. Bhattacharjee A Das, Chowdhury AR. Short-Term Solar Irradiance Forecasting Using Long Short Term Memory Variants. *Lecture Notes in Networks and Systems*. 2022; 288: 227–243. https://doi.org/10.1007/978-981-16-5120-5_18
67. Samantaray S, Sahoo A, Deba Satapathy P. Prediction of groundwater-level using novel SVM-ALO, SVM-FOA, and SVM-FFA algorithms at Purba-Medinipur, India. *Arabian Journal of Geosciences* 2022 15:8. 2022; 15: 1–22. <https://doi.org/10.1007/S12517-022-09900-Y>
68. El-Azab HAI, Swief RA, El-Amary NH, Temraz HK. Machine and deep learning approaches for forecasting electricity price and energy load assessment on real datasets. *Ain Shams Engineering Journal*. 2024; 15: 102613. <https://doi.org/10.1016/J.ASEJ.2023.102613>
69. Sridharam S, Sahoo A, Samantaray S, Ghose DK. Estimation of Water Table Depth Using Wavelet-ANFIS: A Case Study. *Lecture Notes in Networks and Systems*. 2021; 134: 747–754. https://doi.org/10.1007/978-981-15-5397-4_76
70. Ghordoyee Milan S, Roobahani A, Arya Azar N, Javadi S. Development of adaptive neuro fuzzy inference system—Evolutionary algorithms hybrid models (ANFIS-EA) for prediction of optimal groundwater exploitation. *J Hydrol (Amst)*. 2021; 598: 126258. <https://doi.org/10.1016/J.JHYDROL.2021.126258>
71. Santhusitha D, Karunasingha K. Root mean square error or mean absolute error? Use their ratio as well. 2021 [cited 24 Apr 2024].
72. Sarkar BN, Samantaray S, Kumar U, Ghose DK. Runoff is a Key Constraint Toward Water Table Fluctuation Using Neural Networks: A Case Study. *Lecture Notes in Networks and Systems*. 2021; 134: 737–745. https://doi.org/10.1007/978-981-15-5397-4_75
73. Samantaray S, Sahoo A, Satapathy DP, Mishra SS. Prophecy of groundwater fluctuation through SVM-FFA hybrid approaches in arid watershed, India. 2022; 7: 341–365. <https://doi.org/10.1016/B978-0-323-91910-4.00020-0>
74. Ghose DK, Samantaray S. Integrated Sensor Networking for Estimating Ground Water Potential in Scanty Rainfall Region: Challenges and Evaluation. *Studies in Computational Intelligence*. 2019; 776: 335–352. https://doi.org/10.1007/978-3-662-57277-1_14

75. Samantaray S, Sumaan P, Surin P, Mohanta NR, Sahoo A. Prophecy of Groundwater Level Using Hybrid ANFIS-BBO Approach. *Lecture Notes in Networks and Systems*. 2022; 288: 273–283. https://doi.org/10.1007/978-981-16-5120-5_21
76. Willmott CJ, Ackleson SG, Davis RE, Feddema JJ, Klink KM, Legates DR, et al. Statistics for the evaluation and comparison of models. *J Geophys Res*. 1985; 90: 8995. <https://doi.org/10.1029/jc090ic05p08995>
77. Camargo ÂP de, Sentelhas PC. Avaliação do Desempenho de Diferentes Métodos de Estimativa da Evapotranspiração Potencial no Estado de São Paulo no Brasil. *Revista Brasileira de Agrometeorologia*. 1997; 5: 89–97.