Pers Ubiquit Comput (2017) 21:113–123
DOI 10.1007/s00779-016-0978-9

ORIGINAL ARTICLE

Evaluation of missing value imputation methods for wireless soil datasets

Jia Shao[1] · Wei Meng[1] · Guodong Sun[2]

Received: 1 September 2016 / Accepted: 30 September 2016 / Published online: 30 November 2016

Abstract Soil data are very important for hydrologists to model and predict the evolution of water–soil environments. In present, the soil data are often collected by unattended wireless sensing system and then inevitably involves continuous missing values due to the unreliability of system, which is different from the manually collected datasets with the data losses being sparsely distributed. This paper investigates seven typical methods that are used to infill soil missing data, and in particular we also attempt to employ the extreme learning machine in missing-data infilling. This work is aimed at answering such a question: Whether or not existing methods suit for wireless sensory soil dataset with continuous missing values, and how well they perform. With a real-world soil dataset involving complete samples as the benchmark, we evaluate and compare these methods, and analyze the possible reasons behind. This study provides insights for designing new methods that can effectively deal with the missing values in wireless sensory soil dataset.

✉ Wei Meng
mnancy@bjfu.edu.cn

Jia Shao
shaojia0822@icloud.com

Guodong Sun
sungd@bjfu.edu.cn

[1] Information School, Beijing Forestry University, 35 Tsinghua East Rd, Beijing 100083, China

[2] Rm 109, West Main Building, Information School, Beijing Forestry University, 35 Tsinghua East Rd, Beijing 100083, China

Keywords Wireless sensory data · Soil dataset · Missing value imputation · Performance evaluation · Extreme learning machine

1 Introduction

The existence of missing data makes it very difficult to realize accurate data analyzing and modeling. In fact, the data missing is not only common in industry, commerce, and scientific research fields [25, 30] but also inevitable in those scenarios. Generally data missing happens due to the errors or the failures of instrument or operation. Without careful considerations of missing data, domain experts cannot efficiently and precisely understand what their data really indicate [9].

For the hydrology, the agriculture, or other ecological fields, the ecological dataset is generally obtained by either human-operating devices or remotely automatic devices [19, 26, 43, 44]. Nowadays, a popular methodology of implementing large-scale, micro-level ecosystem monitoring is to deploy wireless sensor networks [4, 29] in the sites concerned by scientists. It benefits much—decreasing the costs of human resources and maintenance and realizing real-time observations across geographically distributed regions [5, 10, 31, 33, 35]. In practice, however, the use of wireless sensor network in ecosystem monitoring poses new challenges—the dataset collected by wireless sensing systems, often called wireless sensory dataset, often experiences more significant data losses. First, wireless ecology sensing systems are usually left unattended in outdoor environments, say, tropical forests, cold regions, wetlands, desserts, and riversides, and are expected to operate for a long term, say, a few weeks or even months. These systems are very prone to be accidentally damaged by extreme weather conditions, such as storms, rains, or lightening, and therefore cannot always record ecological events constantly. Moreover, limited labor resources or unpredictable harsh weathers sometimes lead to infeasible visits to remotely deployed devices, and consequently, the damage or the failure of devices often cannot be discovered until the next routine checking, which aggravates the data loss so much that the missing values or even records in the dataset occur one after another—forming large gaps in the dataset.

Second, the low-power low-rate wireless links used to form ecological sensors into a network are unreliable and rendered dynamics sometimes [20] and consequently cannot deliver all the obtained data to end-users—also leading to non-ignorable data missing. Different than traditional datasets in which missing values are very sparsely scattered, therefore, the wireless sensory dataset inevitably suffers missing values that occur continuously in a larger range and considerably undermine the completeness of dataset. Also, it is worth noting that strictly speaking, repeating the operations of obtaining ecological data does not make sense because of the ceaseless temporal–spatial dynamics of natural environment. Figure 1 indicates the incompleteness in the dataset obtained by a small-scale wireless sensor network we use to monitor the hydrology in forests. Clearly, we can see the continuous data missing due to the failed data communication through wireless links (at sensor 2) and the unattended battery depletion (at sensor 3). In summary, processing the data loss or determining desirable methods of infilling missing values is still the first step for scientists of hydrology or agriculture to well understand environmental evolution, even though they benefit from wireless ecology monitoring systems in terms of human resources and real-time data acquisition.

Until now, however, researchers have not yet paid attention to infill the continuous missing values in wireless sensory time-series datasets and have little knowledge about which existing methods are possibly effective under such a case. This study investigates several typical approaches of infilling missing data designed for traditional time-series dataset and Extreme Learning Machine (ELM), which has not been employed in missing-data infilling, and examines their performances in dealing with large-scale continuous data missing in the dataset of soil moisture. The purpose is to analyze and determine which approaches will be more potential for this new task and to give some insights for designing new data missing infilling policies for wireless sensory ecological dataset. Our work is based on the soil moisture dataset because as a critical environmental factor [21], the soil moisture data are a common input to hydrologic and agricultural models in the soil and water management activities [2, 3, 8, 12, 27].

The rest of this paper is organized as follows: Section 2 shows the related works about missing value imputation methods. Section 3 presents eight methods of infilling missing data and how to apply them in our dataset. Section 4 introduces the soil dataset we use and evaluates the performances of these eight methods in terms of accuracy. Finally, Sect. 5 concludes this study.

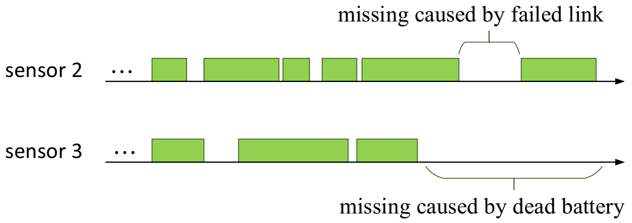

Fig. 1 Illustration of the continuous data missing in a dataset whose data points are returned by two wireless sensors

2 Related work

Various methods have been employed to infill the missing values appearing different scientific fields such as statistical methods, machine learning methods, and data mining methods. The authors in [13] apply hybrid methods, which combine the k-nearest neighbor and the dynamic time warping to infill the missing values in gene expression time-series data. [22] shows how genetic algorithms are used to develop locally weighted regressive models (LWR) and time delay neural network (TDNN) for estimating missing data and compare these two sophisticated methods on short-term hourly volumes of traffic missing counts. The results show that LWR outperforms TDNN. In [32], a piecewise interpolation method based on the cubic Ball and Bzier curves representation is presented to infill the missing value of solar radiation.

Recently, there have been more attempts that study the missing values infilling methods for soil moisture datasets. The authors in [41] present a three-dimensional method, based on the discrete transforms, for filling the missing values of the satellite images dataset of soil moisture. Dumedah and Coulibaly [7] treat the soil moisture dataset to be a time series and investigate the effectiveness of six methods, including the multiple linear regression, the weighted Pearson correlation coefficient, the station relative difference, the soil layer relative difference, the monthly average, and the merged method. In their subsequent work [8], they further evaluate nine neural network based infilling methods; they find that the nonlinear autoregressive neural network, the rough set method, and the monthly replacement can achieve better accuracy in comparison with the methods in their previous paper. Kornelson and Coulibaly [17] examine the effectiveness of the monthly average, the soil layer relative difference, the linear and cubic interpolation, the artificial neural networks, and the evolutionary polynomial regression infilling methods; the evaluation results show that the interpolation and the artificial neural network methods are more effective, yet only for infilling small gaps in dataset.

However, these methods all assume small gaps in the datasets and then are unable to be effectively applied to infill continuous missing data inherently existing in the wireless sensory datasets. In this paper we test the Extreme Learning Machine (ELM) and seven typical methods to evaluate their performance, which are the Linear Interpolation (LI), the Soil Layer Relative Difference (SLRD), the Autoregressive Integrated Moving Average (ARIMA), the Vertical Multiple Linear Regression (VMLR), the Horizontal Multiple Linear Regression (HMLR), the Weighted K-Nearest Neighbors (WKNN), and Radial Basis Function networks (RBFs). We conduct comprehensive numeric experiments based on a soil moisture dataset with various missing gaps, and we compare their imputation performances on a soil moisture dataset involving unsteady records.

3 Description of infilling methods

This section will introduce eight methods for infilling missing values in the soil moisture dataset. They are Linear Interpolation (LI), the Soil Layer Relative Difference (SLRD), the Autoregressive Integrated Moving Average (ARIMA), the Vertical Multiple Linear Regression (VMLR), the Horizontal Multiple Linear Regression (HMLR), the Weighted K-Nearest Neighbors (WKNN), the Radial Basis Function networks (RBF), and the Extreme Learning Machine (ELM). The reasons why we chose these eight methods are as follows. The LI, a simple but effective method, is commonly used in practice to infill missing values. The SLRD is commonly employed by field experts to infill the hydrological data. The ARIMA model always appears in the reconstruction of time-series data with missing values. The VMLR and HMLR are the multiple linear regressions to infill the missing soil moisture values; the difference is that they leverage different sensing attributes in modeling: The first uses the attributes from different layers of a given station, and the second uses the attributes from different stations at the same layer. The WKNN is a kind of typical machine learning method, which is also used to predict the missing values. ELM and RBF are both Single Layer Feed Forward Neural Networks (SLFNs). The RBF recently has widely used to impute the missing values and achieved the ideal results. However, the ELM shows better generalization performances and then has been applied to many fields recently, such as hydrology, pattern recognition, neuroscience, and consumer electronics; we want to know the potential of ELM in the scenario of the soil data infilling. We develop programs based on the R language and MATLAB to implement those methods.

3.1 Linear interpolation (LI)

Based on the curve fitting with linear polynomials, the linear interpolation (LI) is a simple but effective method in practice [23]. The LI fills the missing values of time series by Eq. (1), where y_0 and y_1 are the soil moisture values on time t_0 and $t_1 (t_1 > t_0)$, respectively, and then y will be the missing value on time t which ranges from t_0 to t_1.

$$y = y_0 + (y_1 - y_0)\frac{t - t_0}{t_1 - t_0} \tag{1}$$

3.2 Soil layer relative difference (SLRD)

For infilling missing values of soil data, field experts often resort to the SLRD method [40], which usually employs the parametric test of relative difference among soil moisture data. Equation (2) shows how to impute the missing soil moisture data. Suppose that there are n soil-monitoring stations in a given region, each of which reports a time-series soil records including samples returned by the probes of different depths (layers). For a given sampling depth j, in Eq. (2), $\theta_{i,j}(t)$ represents the soil moisture of depth j at station i at time t, and $\bar{\theta}_j(t)$ represents the average over the depth-j soil moisture values reported by all the n stations at time t. And, $\delta_{i,j}$, called the relative difference, is calculated by the first equation of Eq. (2).

$$\begin{aligned} \delta_{i,j}(t) &= \frac{\theta_{i,j}(t) - \bar{\theta}_j(t)}{\bar{\theta}_j(t)} \\ \bar{\theta}_j(t) &= \frac{1}{n}\sum_{i=1}^{n}\theta_{i,j}(t) \end{aligned} \tag{2}$$

Note that the SLRD method only takes into consideration the data with as the same depth as the missing data, because it assumes that across different stations, the soil moisture data with an identical depth are relatively correlated [7]. When soil moisture is missing at depth j of station i at time t, $\bar{\theta}_j(t)$ is computed by the available depth-j data of all the other stations, while $\bar{\delta}_{i,j}$ is estimated by the mean of all the values of the j-th depth at station i. The estimated soil moisture θ_{est} can be expressed with Eq. (3).

$$\theta_{\text{est}}(t) = \bar{\theta}_j(t) + \bar{\theta}_j(t) \times \bar{\delta}_{i,j} \tag{3}$$

3.3 Autoregressive integrated moving average (ARIMA)

Typical for statistics, the ARIMA model is widely used to analyze the time-series data [16]. In fact, ARIMA involves three types of models: the autoregressive model (AR), the moving average model (MA), and the model (ARMA) combining MA and AR. To process a non-stationary data series, like the soil data we use, ARIMA has to difference this data series to make it stationary for the const statistical properties. We do not consider the seasonal effect of our soil data because it is collected in winter; we then use the non-seasonal ARIMA(p, d, q) model [11] to predict (infill) the missing values, in which p is the number of autoregressive term, d, the number of non-seasonal differences for keep stationary, and q, the number of lagged forecast errors. The general ARIMA model is given together in Eqs. (4) and (5).

$$y_t = \begin{cases} Y_t & d = 0 \\ Y_t - Y_{t-1} & d = 1 \\ (Y_t - Y_{t-1}) - Y_{t-1} - Y_{t-2} & d = 2 \\ \cdots & \cdots \end{cases} \tag{4}$$

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \cdots \phi_p y_{t-p} - \theta_1 e_{t-1} - \cdots \theta_q t_{t-q} \tag{5}$$

In Eq. (4), Y_t is the observed data series until time t, y_t is d-th difference of Y_t, and generally, that $d \in [0, 4]$ is adequate to lead to a stationary series. For the general forecasting given by Eq. (5), $\phi_i (1 \le i \le p)$ and $\theta_i (1 \le i \le q)$ are model parameters, while p and q are the model orders. The parameters ϕ_i and θ_i are often estimated according to the least square or the maximum likelihood methods.

When a missing value is of sequence k in the whole data, ARIMA first chooses a sub-series of length L_k before the k-th data point. In this paper, we plot the original soil moisture data and find its non-stationarity. After empirically differencing the non-stationary soil data of length L_k with a proper d, we can determine a desirable pair of p and q by examining the auto-correlation and the partial correlation of y_t. Finally we mainly use the `arima` function provided by the R language to complete the missing value imputation.

3.4 Vertical multiple linear regression (VMLR)

Each sensing probe attached to the station can sample not only the soil moisture but also the soil temperature and the electrical conductivity data. The VMLR method assumes that for a given depth k, the soil moisture data of depth k correlate both with the soil moisture values of other depths and with the temperature and the electrical conductivity of depth k.

$$\hat{y}_k = a_1 \times t_k + a_2 \times c_k + \sum_{i=1, i \neq k}^{m} b_i y_i \tag{6}$$

If there are m layers, the VMLR model is expressed in Eq. (6) where t_k and c_k represent the temperature and the electrical conductivity of depth k, respectively, and y_i, the soil moisture value of depth $i (i \neq k)$. Therefore, the task of the VMLR is to find parameters a_1, a_2, and b_i.

3.5 Horizontal multiple linear regression (HMLR)

Similar to the VMLR method, the HMLR method also uses the multiple linear regression to infill the missing soil moisture values. Yet the HMLR method focuses on the correlation of data points at the same depth from different stations; in other words, for a given station s, the soil moisture data of depth k at s correlate both with the soil moisture values of depth k of other stations and with the temperature and the electrical conductivity of depth k at station s. The correlation of sensing attributes from nearby sensors is often employed to predict the missing data due to faulty devices [42].

$$\hat{y}_{s,k} = a_1 \times t_{s,k} + a_2 \times c_{s,k} + \sum_{i=1, i \neq s}^{m} b_i \cdot y_{i,k} \tag{7}$$

The HMLR model is given by Eq. (7) where m denotes the number of stations, $t_{s,k}$ and $c_{s,k}$ are the temperature and the electrical conductivity values of depth k at station s.

3.6 Weighted K-nearest neighbors (WKNN)

The WKNN resorts to K similar observations to impute missing values. In practice, the Euclidean distance is commonly used to determine the similarity between two data points. For the simplicity, suppose that data point x has a missing value at attribute a, denoted by $x^{(a)}$, and that there are n data points, y_1, y_2,...y_n in the training dataset. The similarity between x and $y_i (1 \le i \le n)$ can be calculated by Eq. (8) where m is the number of attributes of x or y_i.

$$d(x, y_i) = \sqrt{\sum_{j=1, j \neq a}^{m} \left(x^{(j)} - y_i^{(j)} \right)^2} \tag{8}$$

After obtaining all the distances between x to y_i, we can determine the k-nearest neighbors. For instance, if the k-nearest neighbors of x are shown in Fig. 2 and the distance from x to y_i is equal to d_i, we can infill $x^{(a)}$ with $\hat{x}^{(a)}$ calculated by Eqs. (9) and (10), both of which together express an implementation of a K-nearest neighbors model with a weighted function.

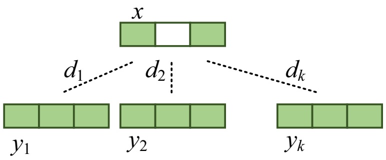

Fig. 2 Illustration for the WKNN method where the *white block* represents the attribute with missing value

$$\hat{x}^{(a)} = \sum_{i=1}^{k} y_i^{(a)} w(d_i) \tag{9}$$

$$w(d_i) = e^{-d_i} \tag{10}$$

3.7 Extreme learning machine (ELM)

Extreme learning machine (ELM) proposed by [14, 15] is a kind of machine learning method for Single Layer Feed Forward Neural Networks (SLFNs). It shows better generalization performances and higher speed of learning process, compared with the SVM and other traditional SLFNs trained by gradient-based algorithms. Therefore, the ELM has been applied to many fields, such as hydrology [6, 38, 45], pattern recognition [24], neuroscience [36], and consumer electronics [1].

Figure 3 illustrates the basic schematic topological structure of an ELM network. Briefly, the basic theory of the ELM model states that for N arbitrary distinct input samples $(\mathbf{x}_k, \mathbf{y}_k) \in R^n \times R^m$, the standard SLFNs with M hidden nodes and an activation g(.) function can be mathematically described as Eq. (11)

$$\sum_{i=1}^{M} \boldsymbol{\beta}_i g(\mathbf{x}_k; c_i, \mathbf{w}_i) = \mathbf{y}_k \quad k = 1, 2, \ldots N \tag{11}$$

where $c_i \in R$ is the bias of the ith hidden node, $\mathbf{w}_i \in R$ is the input weight vector connecting the ith hidden node and the input nodes, $\boldsymbol{\beta}_i$ is the weight vector connection the ith hidden node to the output node, and $g(\mathbf{x}_k; c_i, \mathbf{w}_i)$ is the output of the ith hidden node with respect to the input sample $\mathbf{x}_k$. In ELM, the input weights and hidden biases are randomly generated. By doing so, the nonlinear system can be converted to the following linear system:

$$\mathbf{H} \times \boldsymbol{\beta} = \mathbf{Y} \tag{12}$$

where $\mathbf{H}$, $\boldsymbol{\beta}$, and $\mathbf{Y}$ are expressed with Eqs. (13), (14), and (15) shown as follows, respectively.

$$\mathbf{H} = \begin{pmatrix} g(\mathbf{x}_1; c_1, \mathbf{w}_1) & \ldots & g(\mathbf{x}_1; c_M, \mathbf{w}_M) \\ \vdots & \vdots & \vdots \\ g(\mathbf{x}_N; c_1, \mathbf{w}_1) & \ldots & g(\mathbf{x}_N; c_M, \mathbf{w}_M) \end{pmatrix}_{N \times M} \tag{13}$$

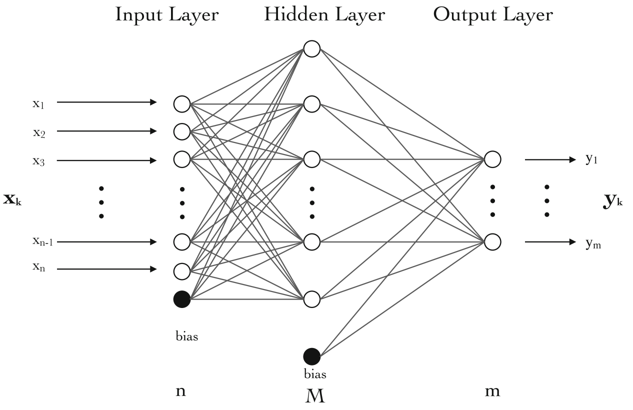

Fig. 3 General topology of the ELM model

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \ldots \boldsymbol{\beta}_M^{\mathrm{T}})^{\mathrm{T}}_{m \times M} \tag{14}$$

$$\mathbf{Y} = (\mathbf{y}_1^{\mathrm{T}}, \mathbf{y}_2^{\mathrm{T}}, \ldots \mathbf{y}_N^{\mathrm{T}})^{\mathrm{T}}_{m \times N} \tag{15}$$

Thus, determining the output weights $\boldsymbol{\beta}$ is as simple as finding the minimum norm least-square (LS) solution to the linear system, described as Eq. (16). As been analyzed by [14], by using such a MP inverse method, ELM tends to obtain good generalization performance and increase the learning speed dramatically.

$$\hat{\boldsymbol{\beta}} = \mathbf{H}^{\dagger}\mathbf{Y} \tag{16}$$

In this paper, we denote, by $\mathbf{Y}$, the attributes with missing values, and by $\mathbf{X}$, the other attributes. All complete records were used to train the ELM. $\mathbf{w}_i$ and c_i are randomly generated within $[-1, 1]$. In order to ensure the statistical significance of the result, in this paper, we repeat 100 times training and predicting processes and use average predicting values to infill the missing values. It is worth noting that the number of hidden neutrons has great influence on the accuracy of the prediction. Through a large number of experiments, shown in Fig. 4 we found that the more the number of neurons, the greater the accuracy, but too many neutrons do not improve the prediction accuracy significantly. Therefore, we empirically set a topology of ELM with 60 neurons in the hidden layer.

3.8 Radial basis function networks (RBF)

Radial basis function networks are also a kind of SLFNs, so the topology of ELM is the same with the RBFs. Different than the other SLFNs, for the RBF model, the weights between input layer and hidden layer are set to be one. In addition, for a given input $\mathbf{x}$, each hidden node of the RBF model will employ a radial basis function to quantify the degree of activation, both of which significantly reduce the SLFN parameters and make the SLFNs easy to be implemented. The general topology of the radial basis function

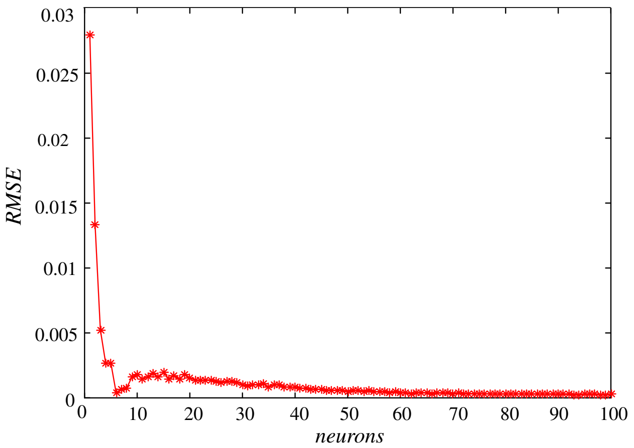

Fig. 4 Effect of the number of neutrons on the accuracy

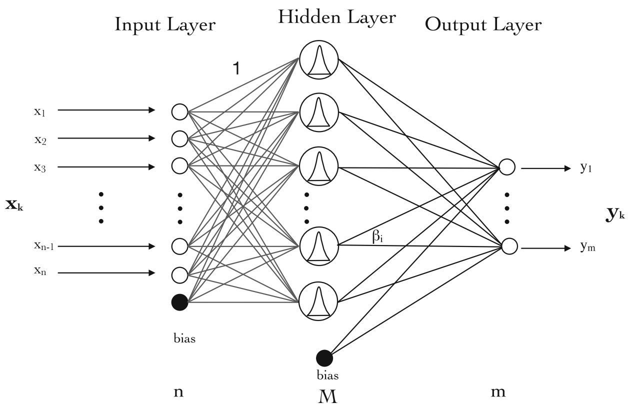

Fig. 5 General topology of the RBF model

neural networks (RBF NNs) is shown in Fig. 5 where $\mathbf{y}_k$ is a weighted linear combination of the activation degrees of incoming input $\mathbf{x}_k$:

$$\sum_{i=1}^{M} \boldsymbol{\beta}_i \Theta_i(\mathbf{x}_k) = \mathbf{y}_k \quad k = 1, 2, \ldots N \tag{17}$$

In the case with the Gaussian type of RBFs, we have

$$\Theta_i(\mathbf{x}_k) = \exp(-\sigma_i \|\mathbf{x}_k - \boldsymbol{v}_i\|^2) \tag{18}$$

where $\mathbf{x}_k = [x_1, \ldots x_n]^{\mathrm{T}}$, represents the n-dimensional input vector, and $\boldsymbol{v}_i = [v_1, \ldots v_n]^{\mathrm{T}}$and σ_i represent center vector of the ith nodes of the hidden layer and the spread parameter of the ith nodes of the hidden layer, respectively. The notation $\| \cdot \|$ in Eq. (18) is the function calculating the Euclidean distance. There are many algorithms to train RBF models [18, 28, 34, 37, 39]. In this paper, we use the standard RBF training algorithm of MATLAB neural network toolbox to infill the missing values in the soil moistures.

4 Analysis

4.1 Soil dataset

The dataset used in this paper is collected by a soil-monitoring system deployed in the Jiufeng National Forestry Park, Beijing, China; this system is shown in Fig. 6, and it locates at 115.7E° and 39.4°N (marked with a red point). Beijing is of dry and monsoon-influenced humid continental climate, where the daily average temperature is only −3.7 °C in January and the precipitation from June to August is about three-fourths of the total yearly precipitation. In this system, there are three soil-monitoring stations around ten meters away from each other; they report their data to a data logger which buffers the collected the data in local SD card. Every week, an operator manually pulled out the soil data file from the SD card. Each station is equipped with five soil sensing probes arranged at five top-bottom layers (depths): 2, 5, 10, 15, and 20 cm, respectively, and each probe simultaneously captures three attributes with an interval of 15 minutes: the soil moisture, the soil temperature, and the soil electrical conductivity. Also, the logger associates a timestamp with each record.

4.2 Setup

The monitoring system in our study site operated from October 2010 to October 2012. In the whole dataset of two years, there are a large amount of irregularly distributed data losses. We elaborately find that the set of records obtained from October 2010 to January 2011 involves only one missing soil moisture value; therefore, this set of records can be reasonably reckoned to be a complete dataset; specifically, we choose the data—returned by the sensing probe of depth 5 cm at a station—as the benchmark dataset to evaluate the eight infilling methods. The benchmark has 6060 records of three soil attributes (the soil moisture, the soil temperature, and the soil conductivity). Figure 7 shows the distribution of all the soil moisture values in the benchmark dataset.

To simulate the continuous missing characteristics of soil dataset returned by wireless sensing system, we artificially specify various missing segments with different ratios. We first remove the missing segment from the benchmark dataset and then apply the eight imputation methods to infill the values in this missing segment. Table 1 gives the missing ratios used in this paper. The choices of five missing ratios are determined by the inspection (physical visit) period in practice, which is usually half a day, 1 day, 1 week, or 1 month (4 weeks). It is clear, in Fig. 7, that the moisture varies steadily before the 2500-th data point, but drastically after the 3000-th data

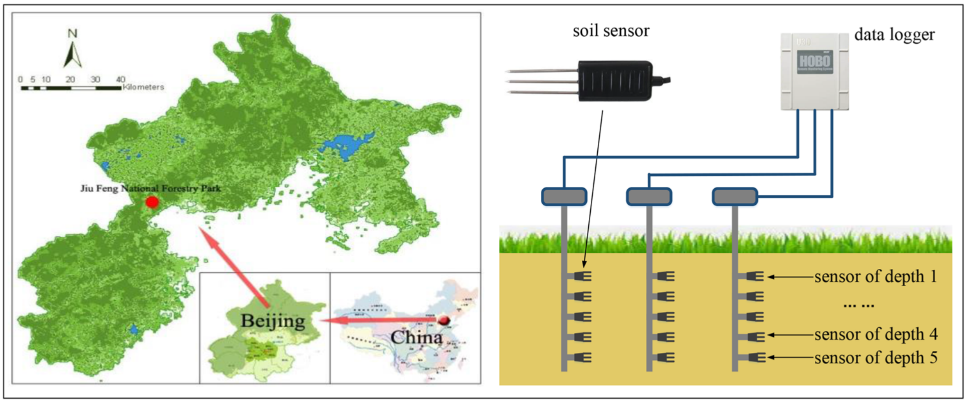

Fig. 6 Deployment of the soil-monitoring site at Jiufeng National Forestry Park, Beijing, China

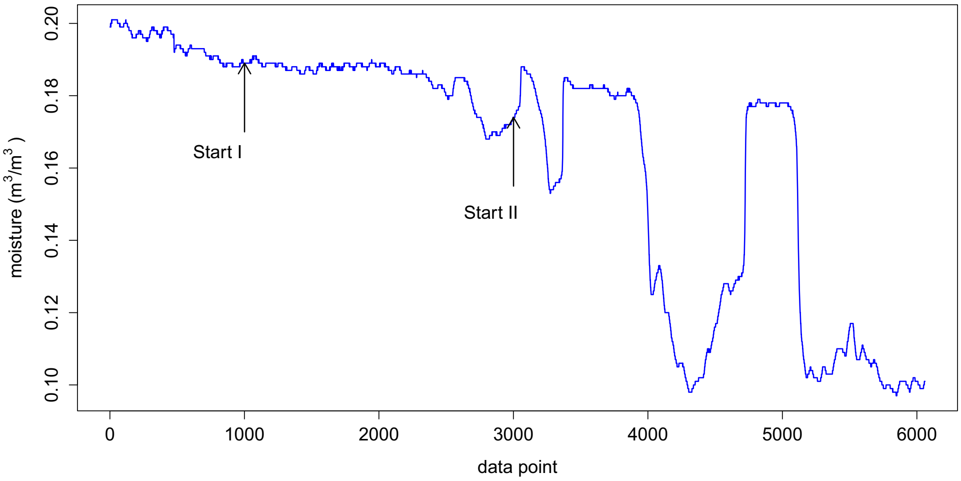

Fig. 7 Soil moisture data from the sensing probe of depth 5 cm at a station

Table 1 Configuration of the missing records in evaluation

Missing segment	Missing range	Missing ratio (%)	Duration (days)
Seg. 1	1001–1048	0.79	0.5
Seg. 2	1001–1192	3.17	2
Seg. 3	1001–1672	11.09	7
Seg. 4	1001–2344	22.18	14
Seg. 5	1001–3688	44.36	28
Seg. 6	3001–3048	0.79	0.5
Seg. 7	3001–3192	3.17	2
Seg. 8	3001–3672	11.09	7
Seg. 9	3001–4344	22.18	14
Seg. 10	3001–5688	44.36	28

point. So, to evaluate the performance of the eight methods under steady and dynamic time-series data, we specify two data points in the benchmark dataset: Start I, the 1001-th data point, and Start II, the 3001-th data point, as labeled in Fig. 7. According to Table 1, missing segments 1–5 all start from Start I and missing segments 6–10 all start from Start II.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{19}$$

In this study, we use the root-mean-square error (RMSE), widely adopted in the community [8], to evaluate the eight methods of infilling the missing soil moisture data. In detail, as shown in Eq. (19), RMSE is the root of the

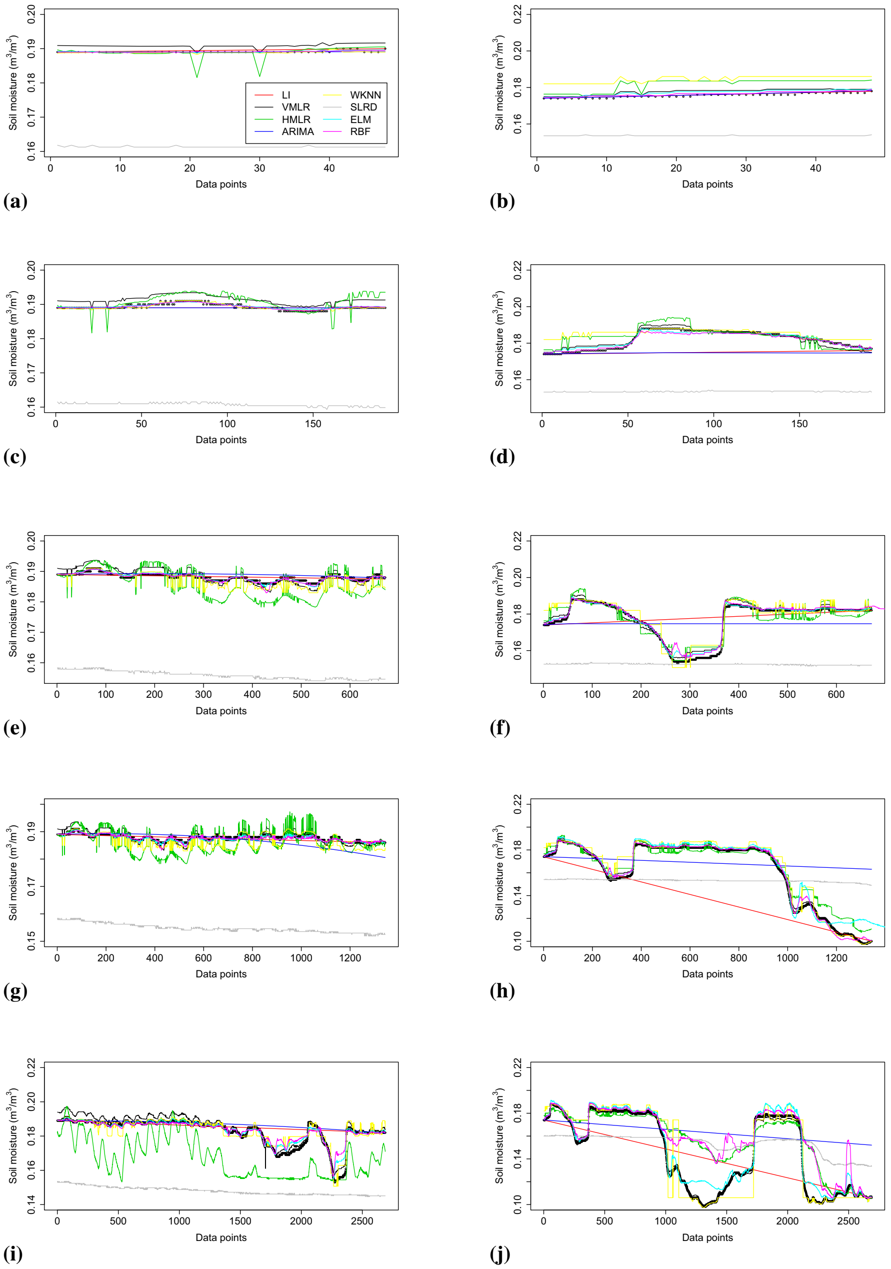

Fig. 8 Comparisons of the eight methods with five different missing ratios. The legend for **a** works for all the other sub-figures. **a** 0.79% missing from the 1001-th data point. **b** 0.79% missing from the 3001-th data point. **c** 3.17% missing from the 1001-th data point. **d** 3.17% missing from the 3001-th data point. **e** 11.09% missing from the 1001-th data point. **f** 11.09% missing from the 3001-th data point. **g** 22.18% missing from the 1001-th data point. **h** 22.18% missing from the 3001-th data point. **i** 44.36% missing from the 1001-th data point. **j** 44.36% missing from the 3001-th data point

average squared differences between the predicted value ($\hat{y}_i$) and the original one (y_i). In general, the smaller the RSME derived by a method is, the better the effectiveness of this method is.

4.3 Results

This section compares the performance of the eight infilling methods with different missing scales and different fluctuations. Figure 8 plots the data points infilled by the eight methods and the real data points. Note that the observed soil moisture values are marked by black circles, and the predicted values of the eight methods are marked by different colors shown as the legend in Fig. 8a. For the shortest missing segment over the steady dataset (Fig. 8a), these infilling methods except for the SLRD all work well; for the longest missing segment over the fluctuating dataset (Fig. 8j), the eight methods differentiate much in performance. From Fig. 8, it can be seen that as the missing ratio increases, the LI and the ARIMA both give straight lines to fit the observed data, regardless of the beginning points of missing segment (Start I or Start II), while the SLRD always performs poorly. However, the VLMR, the ELM, the WKNN, the RBF, and the HMLR all can predict the variation trend of the datasets even with different accuracies. Furthermore, we can see that the fitting performance

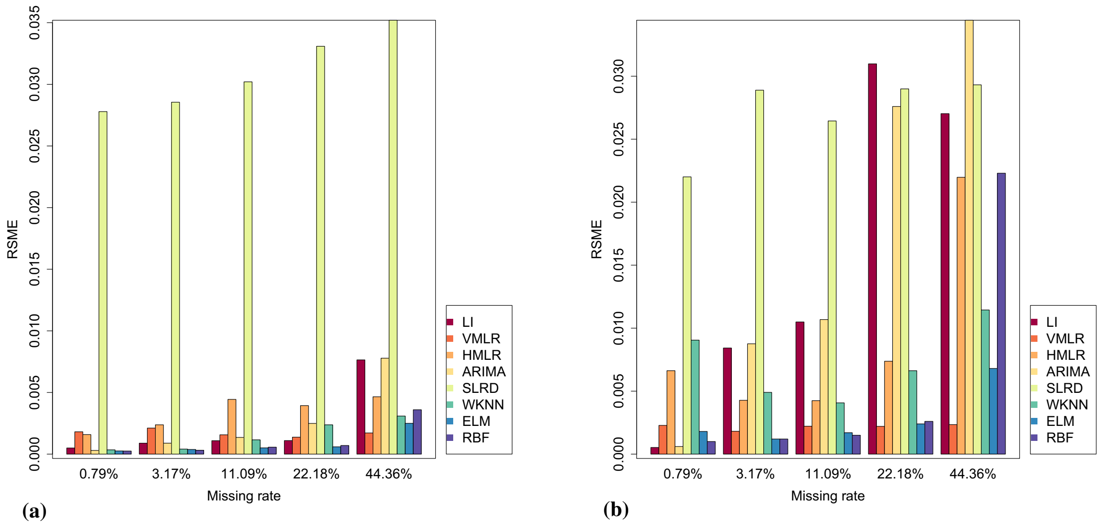

Fig. 9 Comparisons of the eight methods with different missing ratios. **a** Missing segments starting from the 1001-th data point. **b** Missing segments starting from the 3001-th data point

Table 2 RSMEs of eight methods with different missing ratios. Given a method, avg. I represents the average RSME over segments from 1 to 5 and avg. II, the average RSME over segments from 6 to 10

Method	RMSE at different missing segments (%)											
	Seg. 1	Seg. 2	Seg. 3	Seg. 4	Seg. 5	Avg. I	Seg.6	Seg. 7	Seg. 8	Seg. 9	Seg. 10	Avg. II
LI	0.050	0.089	0.109	0.110	0.765	0.225	**0.053**	0.843	1.049	3.098	2.703	1.549
SLRD	2.779	2.855	3.020	3.309	3.520	3.097	2.201	2.889	2.645	2.899	2.932	2.713
VMLR	0.181	0.211	0.157	0.138	**0.172**	0.172	0.228	0.181	0.222	**0.220**	**0.234**	**0.217**
HMLR	0.159	0.238	0.444	0.393	0.465	0.340	0.662	0.427	0.424	0.738	2.198	0.890
ARIMA	0.030	0.089	0.136	0.249	0.779	0.257	0.060	0.876	1.068	2.760	3.444	1.642
WKNN	0.035	0.041	0.116	0.238	0.309	0.148	0.905	0.490	0.407	0.662	1.144	0.722
ELM	0.026	0.038	**0.051**	**0.059**	0.25	**0.085**	0.180	**0.120**	0.170	0.240	0.680	0.278
RBF	**0.025**	**0.032**	0.057	0.070	0.360	0.109	0.100	**0.120**	**0.150**	0.260	2.23	0.572

Bold values represent the best performance in the column it belongs to

of the HMLR and the RBF both experience a significant degradation in the case with 44.36% missing ratios. Nevertheless, estimates from the VLMR, the ELM, and the WKNN are very well consistent with the trend of observed soil moisture, especially the VLMR in larger missing ratios.

The eight methods are further compared in Fig. 9. It is obvious that the imputation performances of these eight methods all degrade in different degrees for the missing segments that start at Start II and involve drastically varying data points. It is worth mentioning that the imputation performances of the LI and the ARIMA become very poor, when the missing segments are chosen from here. The LI method uses only two reference points; therefore, it does not work well for fluctuating datasets, especially when the missing gap is larger. The ARIMA just uses a segment of steady data before the missing values (Start II), which does not contain sufficient information (large or periodic dataset is preferable for ARIMA) and consequently results in lower performance. The RBF has a significant degradation in the 44.36% missing ratios from the Start II. The reason is that there are no sufficient various training samples to build the RBF model. And compared with the ELM, the RBF has the risk of overfitting or underfitting as a result of the restriction of the standard RBF training algorithm in MATLAB neural network toolbox. Interestingly, the VMLR demonstrates the most steady and precise prediction as the dataset becomes unsteady and the missing ratio is larger. Both the VMLR and the HMLR employ the multiple linear regression to infill the missing soil moisture values, but the VMLR is preferred to the HMLR—suggesting that for a given station, the different soil layers (depths) for the VMLR can profile the temporal correlation of soil moisture with higher accuracy, i.e., the data from vertically arranged layers at the same station render closer correlation, in comparison with the same layers at different stations.

A comprehensive numeric comparison in terms of RMSE is given in Table 2. For the missing segments beginning at Start I, in average, the ELM is the best predictor, followed by the RBF, the WKNN, the VMLR, all of which are not significantly different, and the worst is the SLRD. For the missing segments beginning at Start II, in average, the VMLR performs the best, followed by the ELM, both of which are similar, and the SLRD still is the worst. We can conclude that when being applied to infill the dataset with the shortest missing ratio, the LI is recommended, considering that this method is simple and has a relatively high infilling accuracy. As missing ratio of the dataset is increasing, the ELM and the RBF are suggested to infill the missing values. However, it is noticeable that the VMLR, with the average RMSE of 0.217% over the missing segments beginning at Start II, outperforms other methods and seems suitable to infill unsteady dataset with the largest missing ratio. Based on the time-series dataset used in this paper, the evaluation results show that the VMLR, the ELM, the RBF, the WKNN, the LI, the ARIMA, and the HMLR are all preferred to the SLRD, which is commonly used by field experts.

5 Conclusions

Ecological time-series dataset collected by wireless sensing systems often experiences continuous data losses which pose new challenges for missing-data processing. Researchers now have little knowledge about effective approaches to addressing this issue. This paper has investigated seven typical methods that are used to infill missing data in a soil time-series dataset and ELM which has not been employed in this task. We find that totally, the VMLR, the ELM, and the RBF can achieve a better accuracy in infilling continuous missing soil moisture data. In detail, to infill short missing segments, the ELM, the RBF perform desirably. The reason is that the ELM and RBF are Single Layer Feed Forward Neural Networks (SLFNs). They both have the ability to approximate arbitrary function, but the ELM shows better generalization. To infill missing values in unsteady soil dataset with a larger continuous missing segments, the VMLR overwhelms all the other methods, and the accuracy of the ELM is slightly lower than that of the VMLR. Therefore, we can see the ELM has a promising potential to infill the missing values in different missing segments. For all the specified missing segments, the VMLR is almost always preferred to the HMLR, indicating that the data from different layers of a given station are more strongly correlated than the data from different stations at the same layer. Thus, taking into account the correlation among multiple factors will be a promising start to design effective approaches of infilling the missing values in wireless soil datasets.

Acknowledgements This study was supported, in part, by the NSF of China with Grant No. 61300180 and by the Fundamental Research Funds for the Central Universities of China with Grant No. TD2014-01.

References

1. Budiman A, Fanany MI (2013) Pose-based 3d human motion analysis using extreme learning machine. In: 2013 IEEE 2nd global conference on consumer electronics (GCCE), pp 3–7
2. Charoenhirunyingyosa S, Hondaa K, Kamthonkiatb D, Inesc A (2011) Soil moisture estimation from inverse modeling using multiple criteria functions. Comput Electron Agric 75(2):278–287
3. Coopersmith E, Minsker B, Wenzel C, Gilmore B (2014) Machine learning assessments of soil drying for agricultural planning. Comput Electron Agric 104:93–104
4. Culler D, Estrin D, Srivastava M (2004) Introduction: overview of sensor networks. Computer 37(8):41–49
5. Dan L, Sun L, Dai W (2014) Wireless sensor networks system of forest habitat factors collection. J Harbin Inst Technol 46(7):123–128
6. Deo RC, Şahin M (2015) Application of the extreme learning machine algorithm for the prediction of monthly effective drought index in eastern Australia. Atmos Res 153:512–525
7. Dumedah G, Coulibaly P (2011) Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. J Hydrol 400:95–102
8. Dumedah G, Walker J, Chik L (2014) Assessing artificial neural networks and statistical methods for infilling missing soil moisture records. J Hydrol 515(16):330–344
9. Farhangfar A, Kurgan L, Dy J (2008) Impact of imputation of missing values on classification error for discrete data. Pattern Recogn 41(12):3692–3705
10. Gong J, Geng J, Chen Z (2015) Real-time GIS data model and sensor web service platform for environmental data management. Int J Health Geograph 14(2):1–13
11. Han P, Wang P, Zhang S, Zhu D (2010) Drought forecasting based on the remote sensing data using arima models. Math Comput Model 51(11–12):1398–1403
12. Hardy A, Barr S, Mills J, Miller P (2012) Characterising soil moisture in transport corridor environments using airborne lidar and casi data. Hydrol Process 26(13):1925–1936
13. Hsu HH, Yang AC, Lu MD (2011) KNN-DTW based missing value imputation for microarray time series data. J Comput 6(3):418–425
14. Huang GB, Zhu QY, Siew CK (2004) Extreme learning machine: a new learning scheme of feedforward neural networks. In: Proceedings of international joint conference on neural networks, vol 2, pp 985–990
15. Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing 70(1–3):489–501
16. Kohn R, Ansley C (1986) Estimation, prediction, and interpolation for arima models with missing data. J Am Stat Assoc 81(395):751–761
17. Kornelsen K, Coulibaly P (2014) Comparison of interpolation, statistical, and data-driven methods for imputation of missing values in a distributed soil moisture dataset. J Hydrol Eng 19(1):26–43

18. Kurban T, Beşdok E (2009) A comparison of RBF neural network training algorithms for inertial sensor based terrain classification. Sensors 9(8):6312–6329
19. Lee W, Alchanatis V, Yang C, Hirafuji M, Moshou D, Li C (2010) Sensing technologies for precision specialty crop production. Comput Electron Agric 74(1):2–33
20. Li J, Gao H (2008) Survey on sensor network research. J Comput Res Develop 45(1):1–15
21. Lindenmayer D, Likens G (2010) The science and application of ecological monitoring. Biol Conserv 143(6):1317–1328
22. Lingras P, Zhong M, Sharma S (1970) Evolutionary regression and neural imputations of missing values. Stud Fuzziness Soft Comput 226:151–163
23. Meijering E, Falk H (2012) A chronology of interpolation: from ancient astronomy to modern signal and image processing. Proc IEEE 90(3):319–342
24. Mohammed AA, Minhas R, Wu QMJ, Sid-Ahmed MA (2011) Human face recognition based on multidimensional PCA and extreme learning machine. Pattern Recogn 44(44):2588–2597
25. Moorthy K, Mohamad MS, Deris S (2014) A review on missing value imputation algorithms for microarray gene expression data. Curr Bioinform 9(1):18–22
26. Mukhopadhyay S, Jiang J (eds) (2013) Wireless sensor networks and ecological monitoring (smart sensors, measurement and instrumentation). Springer, Berlin
27. Nemes A, Wosten J, Varallyay G, Bouma J (2006) Soil water balance scenario studies using predicted soil hydraulic parameters. Hydrol Process 20(5):1075–1094
28. Neruda R, Kudová P (2005) Learning methods for radial basis function networks. Future Gener Comput Syst 21(7):1131–1142
29. Ojha T, Misraa S, Raghuwanshib N (2015) Wireless sensor networks for agriculture: the state-of-the-art in practice and future challenges. Comput Electron Agric 118:66–84
30. Pigott TD (2001) A review of methods for missing data. Educ Res Eval 7(4):353–383
31. Pomati F, Jokela J, Simora M, Veronesi M, Ibelings B (2011) An automated platform for phytoplankton ecology and aquatic ecosystem monitoring. Environ Sci Technol 45(22):9658–9665
32. Saaban A, Zainudin L, Bakar MNA (2014) On piecewise interpolation techniques for estimating solar radiation missing values in Kedah. J Immunol 160(6):2824–2830
33. Schneider A (2012) Monitoring land cover change in urban and peri-urban areas using dense time stacks of landsat satellite data and a data mining approach. Remote Sens Environ 124:689–704
34. Schwenker F, Kestler HA, Palm G (2001) Three learning phases for radial-basis-function networks. Neural Netw 14(4–5):439–458
35. Shi-Chang D, Li-Feng X, Jian-Jun S (2006) Distributed sensor system for fault detection and isolation in multistage manufacturing systems. Int J Comput Appl Technol 25(4):1
36. Song Y, Crowcroft J, Zhang J (2012) Automatic epileptic seizure detection in EEGs based on optimized sample entropy and extreme learning machine. J Neurosci Methods 210(2):132–146
37. Sultan Noman Qasem SMS (2011) Radial basis function network based on time variant multi-objective particle swarm optimization for medical diseases diagnosis. Appl Soft Comput 11(1):1427–1438
38. Taormina R, Chau KW (2015) Data-driven input variable selection for rainfall-runoff modeling using binary-coded particle swarm optimization and extreme learning machines. J Hydrol 529:1617–1632
39. Tsekouras GE, Tsimikas J (2013) On training RBF neural networks using input-output fuzzy clustering and particle swarm optimization. Fuzzy Sets Syst 221:65–89
40. Vachaud G, Silans APD, Balabanis P, Vauclin M (1985) Temporal stability of spatially measured soil water probability density function. Soil Sci Soc Am J 49(49):822–828
41. Wang G, Garciab D, Liu Y, Jeua R, Dolmana A (2012) A three-dimensional gap filling method for large geophysical datasets: application to global satellite soil moisture observations. Eviron Modell Softw 30:139–142
42. Wang J, Damevski K, Chen H (2015) Sensor data modeling and validating for wireless soil sensor network. Comput Electron Agric 112:75–82
43. Wang N, Zhang N, Wang M (2006) Wireless sensors in agriculture and food industry$^{\circ \mathrm{TM}}$recent development and future perspective. Comput Electron Agric 50(1):1–14
44. Yang J, Zhang C, Li X (2010) Integration of wireless sensor networks in environmental monitoring cyber infrastructure. Wireless Netw 16(4):1091–1108
45. Yue L, Long M, Su KO (2014) Prediction of soil moisture based on extreme learning machine for an apple orchard. In: IEEE international conference on cloud computing and intelligence systems