International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 12 (2015) pp. 30845-30859
© Research India Publications
http://www.ripublication.com

Precipitation and Temperature based Spatial Data Preprocessing Using Machine Learning Techniques

Bhabani Shankar Das Mohapatra[1] and Dejenie Aynalem[2]

[1]Research Scholar, CSIT, JNT University, Hyderabad, India
dm_bhabani@yahoo.co.in
[2]Faculty of Computing, Institute of Technology, Bahir Dar University, Ethiopia
dejayn03@yahoo.com

Abstract

By its nature, spatial data collected from the real world are complex and highly susceptible to noisy and missing data. Noisy data is an unavoidable problem in dealing with most of the real world data sources.

Many machine learning methods have been used for handling missing spatial data. Each of them has its own limitations and they are suitable for some studies and unsuitable for others.

In this investigation, we discern that the imputation methods fill in missing data with more precise estimated values based on information available in the data set. We consider the results using the performance metrics of MAE, MSE, and correlation coefficient (R) for each climate variable. However, we observe that there is no existing model that is best in all performance measuring methods. In this work, we contemplate Holt Winter's method and ANN imputation technique that furnish an overall better result, because Holt Winter's method considers seasonal variation whereas ANN models are applied MLP as imputation methods if the time series data have sufficient available data before missing values occurred.

Keyword: Preprocessing, NMA, Holts Method, ANN, Imputation Techniques.

1. Introduction

Data preparation or data pre-processing is always the first step in the machine learning process. Preprocessing is required before one can apply machine learning to the dataset.

Data preparation comprises techniques concerned with analyzing raw data so as to produce quality data, mainly including data collecting, data integration, data transformation, data cleaning, data reduction, and data discretization [1].

Data can be generated using models or collected from the real world. In this study, we used real world data collected from the weather stations of National Meteorology agency (NMA) of Ethiopia.

By its nature, data collected from the real world are complex and highly susceptible to noisy and missing data.

The data can be preprocessed to improve the quality of data so as to improve the prediction results.

In this work, preprocessing has been done to replace the missing values [2].

Generally, the process for getting data ready for a machine learning algorithm can be shortened into three steps:

- **Select data**- Consider what data is available, what data is missing and what data can be removed.
- **Preprocess Data**- Organize your selected data by formatting, cleaning and sampling from it.
- **Transform Data**- Transform preprocessed data ready for machine learning by engineering features using scaling, attribute decomposition and attribute aggregation.

The rest of the paper is organized as follow:

In section 2, a brief literature review on missing data handling techniques is described. Section 3 introduces the Imputation Techniques and Section 4 illustrates the experiment of selected imputation methods. Evaluation of the models isdescribed in section 5 and conclusion is given in Section 6.

2. Missing Data Handling Mechanisms

Missing data is an unavoidable problem in dealing with most of the real world data sources.

Considering the fact that an inappropriate method can bias the information contained in the data set and further processing can result in incorrect models. It is clear that handling the missing values is a very important part of data preprocessing. When having to handle with a missing value problem we have to consider some basic things [5].To handle missing, it is helpful to know that the missing variables are whether related or not with the others that are not missing. If they are not related, it is impossible to predict the missing values based on the others consistently. If the data is missing at random, then there is no information in the absence of the data and it is possible to replace the missing values with estimates based on the observed data without losing any information. In our case, the data is MAR, there is no hidden information behind in the absence of the data since the reason for their absence is unavailability of physical measurements or the observer due to different reasons. Thus, it is possible to fill the missing values with estimates of the recorded data. On the other hand, if the missing data are related to non-missing data, there are different methods that can be used to handle them. In this study, we believe that there are strong relations between the variables and therefore it is possible to model the missing values based on the variables that are present [7].

Another aspect when dealing with a missing-value problem is the cost of the possible solution. There is a trade-off between the time required to compute estimates for the missing values and the quality of the estimates. We have to make a compromise between the two. The cost of the method used is determined from a variety of factors. First is the size of the data set. This can be very important for choosing a method to handle missing values [5]. But, in this study the dataset is not big and so ideally we can use a method that is computationally expensive and it is possible to select that produces sufficiently good results. Another issue used to select a method is the number of missing values in the data set. For this issue, we used some commonly used missing data handling methods using station data that have less missing data and high percentage missing data as well. By comparing the result, we select the method that performs better estimation.

Many methods have been used for handling missing data. Each of them has its own properties and they are suitable for some problems and unsuitable for others. Below, we present some of the most widely used methods for handling missing values and discuss their suitability to handle missing values of the data.

There is a number of missing data treatment methods [2] [3]:

a) **Ignoring or discarding data**: There are two main ways to discard data with missing values. The first one is known as complete case analysis. It is available in all statistical packages and is the default method in many programs. This method consists of discarding all instances with missing data. The second method is known as discarding instances and/or attributes. This method determines the extent of missing data on each instance and attribute, and deletes the instances and/or attributes with high levels of missing data. Before deleting any attribute, it is necessary to evaluate its relevance to the analysis. Unfortunately, relevant attributes should be kept even with a high degree of missing values [4].

b) **Parameter estimation**. Maximum likelihood estimation (MLE) procedures are used to estimate the parameters of a model defined for the complete data in the presence of missing data (Expectation-Maximization or EM algorithm). A disadvantage of EM algorithm is that its rate of convergence can be painfully slow when there is a large fraction of missing values. EM algorithm is an iterative algorithm that finds the parameters which maximizes the log likelihood when there are missing values. It capitalizes on the relationship between missing data and the unknown parameters of a data model. A disadvantage of EM algorithm is that its rate of convergence can be slow when there is a large fraction of missing values. Each of the iteration of EM consists of an E-step (expectation step) and M- step (maximization step). Given a set of parameter estimates, E-step calculates the conditional expectation of the complete data log likelihood given the observed data and the parameter estimates [2].

For computational convenience, the MLE estimate is obtained by maximizing the log-likelihood function, $\ln L(w/y)$: Assuming that the log-likelihood function, $\ln L(w/y)$ is differentiable, it must satisfy the following partial differential equation known as the likelihood equation:

$$\frac{\partial \ln L(w/y)}{\partial wi} = 0,$$

L(w/y) represents the likelihood of the parameter *w* given the observed data *y*; and as such is a function of *w*.

The likelihood equation represents a necessary condition for the existence of an MLE estimate. An additional condition must also be satisfied to ensure that *ln L(w/y)* is a maximum and not a minimum, since the first derivative cannot reveal this [17].

c) **Imputation Methods**: It is a class of procedures that aims to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values.

In this work, the interest is on imputation techniques to fill in missing data because there a number of alternatives methods.

3. Imputation Techniques

Imputation methods fill in missing data with estimated values based on information available in the data set.

Replacing missing values using imputation methods is a better means rather than ignoring or removing variables or observations with missing data.

There are many options varying from simplistic methods such as the mean imputation to more robust methods based on relationships among attributes.

Some commonly used imputation methods are [6]:

a) **Substitution**: In this case, one instance with missing data is substituted by another non sampled instance. This method is very simple: for nominal attributes, the missing data is replaced with the most common attribute value or the mode; numerical values are replaced with the average of all values of the corresponding attribute or the mean. And mostly used in sample surveys not employed in this work.

b) **Hot deck imputation**: Identify the most similar case to the case with a missing value and substitute the most similar case's Y value for the missing case's Y value.

The imputed values do not distort the distribution of the sampled values. It is common in survey practice and can involve very elaborate schemes for selecting units that are similar for imputation. The disadvantage is that it is difficult to find such similar responding units in the sample area

c) **Regression Substitution**: Here, we can replace the missing value with historical value from similar cases. Given a missing value for a variable X, suppose that q variables have been observed for that record.

The records where these q + 1 variables are available define a training set, and a regression model to predict X from the q predictors is fitted. Finally, the fitted model provides a prediction for the initial missing value of X.

A number p >1 of independent variables X_1, X_2,..., X_p is considered, so a population model, $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdot\cdot\cdot + \beta_pX_p + \varepsilon$, is assumed where Y denotes the dependent variable or response, X_1, X_2,..., X_p are the independent or predictor variables, ε is a random disturbance or error whose presence represents the absence of an accurate relationship.

And β0, β1... βp are unknown coefficients or parameters that define the regression hyper-plane $\beta_0 + \beta_1X_1 + \beta_2X_2 + \cdot\cdot\cdot + \beta_pX_p$.

If a qualitative variable is considered with c categories, c–1 dummy dichotomous variables are introduced into the model [10].

d) **Matching Imputation:** In such case, for each unit with a missing y, we find a unit with similar values of x in the observed data and take its y value [7].

e) **Mean or Mode Imputation**: It is a simple method where any missing value of a quantitative variable is replaced by the mean of the observed values for that variable. So, if a variable presents several missing values for different records, all of them are imputed with the same value. If the variable is qualitative, the missing values are replaced by the mode.

The disadvantage of this method may severely distort summary measures including underestimates of the standard deviation.

f) **Last Observation Carried Forward (LOCF):** One of the simple and widely used methods is Last Observation Carried Forward (LOCF). This method is for every missing value to be replaced by the last observed value from the same type. However, the value of the outcome remains unchanged after missing, which seems likely to be unrealistic.

g) **Prediction model imputations**: Prediction models are one of the methods used for handling missing data.

These methods consist of creating a predictive model to estimate values that will substitute the missing data such as ANN. Most of the time attributes have correlations among themselves. In this way, those correlations could be used to create a predictive model for classification or regression for, respectively, qualitative and quantitative attributes with missing data. Some of these relationships among the attributes may be maintained if they are captured by the predictive model. The disadvantage of this approach is that the model estimated values are likely to be more consistent with this set of attributes than the true (observed) values it would be. A second disadvantage is that there must be a correlation among the attributes. If there are no relationships among attributes in the data set and the attribute with missing data, then the model will not be precise for estimating missing values. The disadvantage of this approach is that the model estimated values are likely to be more consistent with this set of attributes than the observed values it would be [5]. The basic unit of the neural network is the neuron. The simplest model of a neuron is the perceptron.

A perceptron simply computes a weighted sum of its inputs and computes a non-linear function of it [14, 15]:

$$y = f(\sum_{i=1}^{n} w_i x_i + b),$$

where, w_i are the weights of the inputs, x_i are the i^{th} components of an input x to the neuron and b is a bias parameter. Here, f is a non-linear activation function (transfer function) and y is the output.

The non-linear function f is called activation function or transfer function and some of the most commonly used choices are the sigmoid, hyperbolic tangent and Gaussian functions.

The most frequently used neural network model is the so-called multilayer perceptron (MLP), which is a fully connected network of neurons organized in several layers [11].

The network, which solves practical nonlinear problems, has hidden layers between the input and output layers. Analysis of the neurons in the hidden layers is a valuable technique for understanding what has been learnt by the network. Multilayer hierarchical networks are powerful because they can generate their own internal representation in the hidden units, which can be used for interpreting the results.

An MLP can learn with a supervised learning rule using the back-propagation algorithm. The backward error propagation algorithm for ANN learning or training caused an advance in the application of multilayer perceptrons.

The back-propagation algorithm gave rise to the iterative gradient algorithms designed to minimize the error measure between the actual output of the neural network and the desired output using a pre-computed error on the forward pass of information through the network.

A trained neural network as a computational model can be represented with a simple formula for computing predictions based on learned/tuned weights and the inputs, i.e. for a two-layer perceptron.

From the single perceptron, it is possible to modify the formula for multilayer perceptron. A trained neural network for computing predictions can be based on learned/tuned weights and the inputs. For a two-layer perceptron, the trained neural networks can be the following model:

$$y\,(x_1, x_2, \ldots x_K) = f^{out}\left\{\sum_{h_2=1}^{H_2} w_{h_2,m}^{(out)} f^{h_2}\left[\sum_{h_1=1}^{H_1} w_{h_2,h_1}^{(2)} f^{h_1}\left(\sum_{k=1}^{K} w_{k,h_1}^{(1)} x_k + b_k\right) + b_{h_1}\right] + b_{h_2}\right\}$$

The weight $w_{h_q,h_p}^{(H)}$ is the weight of the link from the neuron h_p of the previous layer to the neuron h_q in the layer H. This layer can be the output layer that is the weights for this are denoted as $w_{h_q,m}^{(out)}$ meaning the link between the neuron h_q in the last hidden layer and the output m. Transfer functions for the hidden layers are denoted as $f^{hs}(.)$ and for the output layer $f^{out}\,(\cdot)$ correspondingly.

Other variables are: m is the index of an output, H_1, H_2 are the number of hidden units in the first and second layers, K is a number of inputs, and b_k, b_{h1} and b_{h2} are the biases of the layers.

Back propagation [15] is an algorithm to compute the gradients of the error function with respect to the network weights. The error to be minimized is by a mean squared error (MSE). The outputs of the MLP trained with an MSE error function can be explained as the conditional average of the target data. By using a single output t, for an inputs-outputs pair ($\boldsymbol{x}$, t), the error is simply:

$$E_{\mathrm{MSE}}(\mathrm{w}) = ½ \;.E[t - z^{out}(x, w)]^2,$$

where, Z^{out} is an output of the MLP estimated for the desired value t.

The basic back propagation algorithm follows these steps [13]:

1) The weights are initialized. At first, we initialize all the weights and bias to be small random values.

2) A pair of inputs -target ($\boldsymbol{x}$, t) is provided to the network.

3) The derivatives of E_{MSE} for a single pair ($\boldsymbol{x}$, t) are computed with respect to the weights in each layer, starting at the output layer with the backward move to the inputs.

h) **K-Nearest Neighbor Imputation**: This method also is used to estimate and substitute missing data. The main benefits of this method are that it can predict both the mode and the mean among the k nearest neighbors. There is no necessity for creating a predictive model for each attribute with missing data. Actually, the k-nearest neighbor algorithm does not create explicit models since the data set is used as a “lazy” model. Thus, the k-nearest neighbor algorithm can be easily adapted to work with any attribute as class, by just modifying the attributes to be considered in the distance metric. Also, this approach can easily treat examples with multiple missing values. The main drawback of the k-nearest neighbor approach is that, whenever the k-nearest neighbor looks for the most similar instances, the algorithm searches through all the data set. We consider what it might be like in a time series problem. In this case the input data is just a long series of time series over time without any particular record that could be considered to be an object. The value to be predicted is just the next value of the time series. The way that this problem is solved for both nearest neighbor techniques and for some other types of prediction algorithms is to create training records by taking, for instance, 10 consecutive stock prices and using the first 9 as predictor values and the 10th as the prediction value. Doing things this way, if there are 100 data points in time series we could create 10 different training records.

We can create even more training records than 10 by creating a new record starting at every data point. For instance, we can take the first 10 data points and create a record. Then we can take the 10 consecutive data points starting at the second data point, then the 10 consecutive data point starting at the third data point. Even though some of the data points would overlap from one record to the next the prediction value would always be different.

The k-nearest neighbor algorithm assigns the classification of the most similar record or records based on the distance.

The most common distance function is *Euclidean distance*, which represents the usual manner in which humans think of distance in the real world:

$$d(x,y) = \surd \sum(x_i - y_i)^2$$

where **x** = x_1, x_2...x_m, and **y** = y_1, y_2,..., y_m represent the *m* attribute values of two records[11].

i) Double Exponential Smoothing (Holt's Method):

This method works best when the time series has a positive or negative trend (i.e. upward or downward). This method uses two constants: β, which is the trend component, which must be chosen in conjunction with α, the mean component [9].

It is defined as

$$Y_{i+1} = E_i + T_i,\ i = 1, 2, \ldots, n$$

where Y_{i+1} = fore-casted value, $E_i = \alpha y_i + (1-\alpha)(E_{i-1} + T_{i-1})$, $Ti = \beta(Ei - E_{i-1}) + (1-\beta)T_{i-1}$, α is the mean constant ($0<\alpha \leq 1$), β is trend constant ($0<=\beta<=1$), and y_i is observed value.

After observing the value of the time series y_i at period i, this method computes an estimate of the base, or expected level of the time series (E_i) and the expected rate of increase or decrease per period (T_i). It is customary to assume that $E_1 = y_1$ and unless told otherwise we assume $T_1 = 0$.

To use the method, first we calculate the *base level* E_i for time i. Then we compute the expected trend value Ti for time period i. Finally, we compute the forecast y_{i+1} [12].

j) Triple Exponential Smoothing (Holt Winters Method):

This method is appropriate when trend and seasonality are present in the time series. When an actual observation is divided by its corresponding seasonal factor, it is said to be de-seasonalized. This allows us to make meaningful comparisons across time periods.

Let $\mathbf{S_i}$ = seasonal factor for period i, **c** = the number of periods in a cycle (for example 12 if months of year, 7 if days of week, etc.), **Li** = seasonal values in one cycle.

The relevant formulas for this method follow:

$$E_i = \alpha(y_i / S_{i-c}) + (1-\alpha)(E_{i-1} + T_{i-1})$$
$$T_i = \beta(E_i - L_{i-1}) + (1-\beta)T_{i-1}$$
$$S_i = \gamma(y_i / L_i) + (1-\gamma)S_{i-c}$$
$$Y_{i+1} = (E_i + T_i)S_{i+1-c} = \text{forecasted value}$$

where α is the mean constant ($0<\alpha <= 1$), β is trend constant ($0<=\beta<=1$), and y_i is observed value. γis another smoothing constant between 0 and 1.

4. Empirical Experiment in Handling Missing/Noisy Data

The aim of this subsection is to investigate the climate dataset to explore the appropriate techniques for handling missing values of the temperature and rainfall dataset.

Different imputation models are compared from simple to complex such as mean/mode, Last Observation Carried Forward (LOCF), ANN imputation, K-Nearest

Neighbor Imputation, double exponential smoothing (Holt’s Method), triple exponential smoothing (Holt Winters Method).

Two types of sample climate variable dataset with missing values were taken and aforementioned imputation techniques were applied. The maximum temperature and mean rainfall data having 5-20 percent missing data were investigated and the models were compared to select the best model and were applied on whole weather station to replace the missing value.

For the imputation model selection, we considered two weather stations having high missing data and less missing data.

Bahir Dar station has less missing data and Zegie station that has high amount of missing data. The figure below shows the missing data taking maximum temperature and mean rain fall of the two stations.

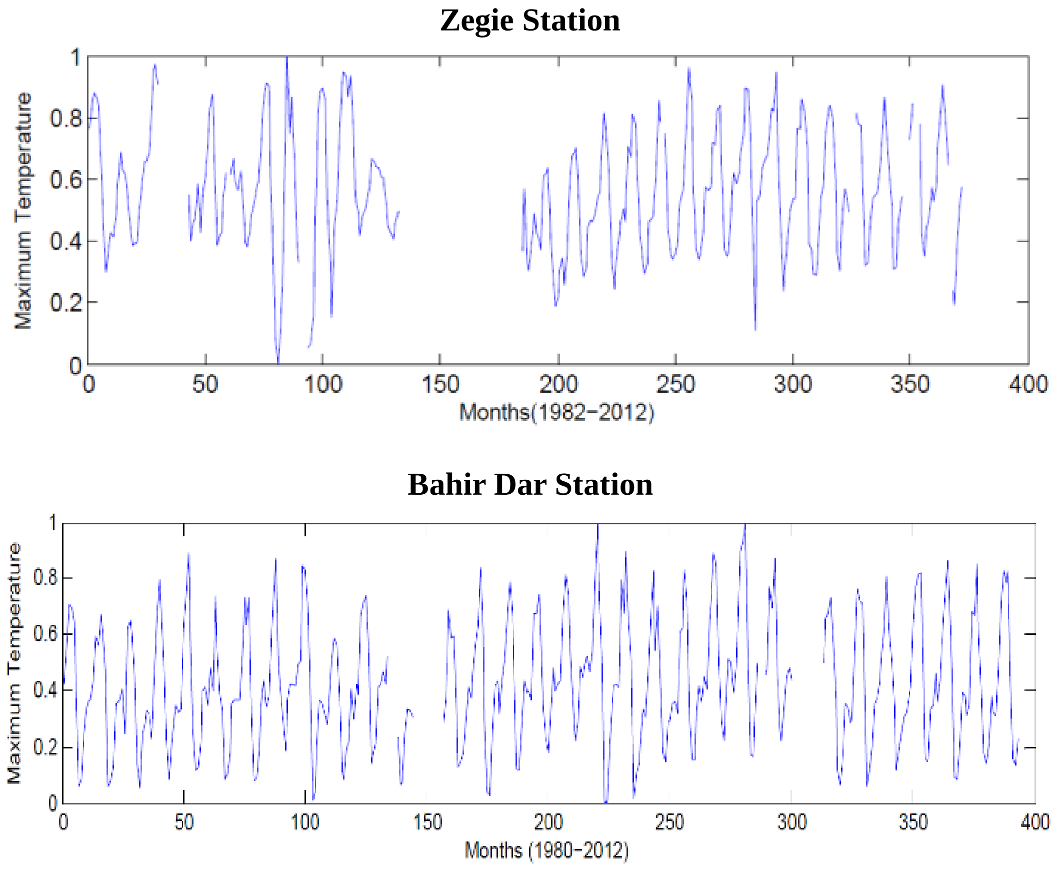

Fig.1 Zegie and Bahir Dar Station Maximum temperature missing data

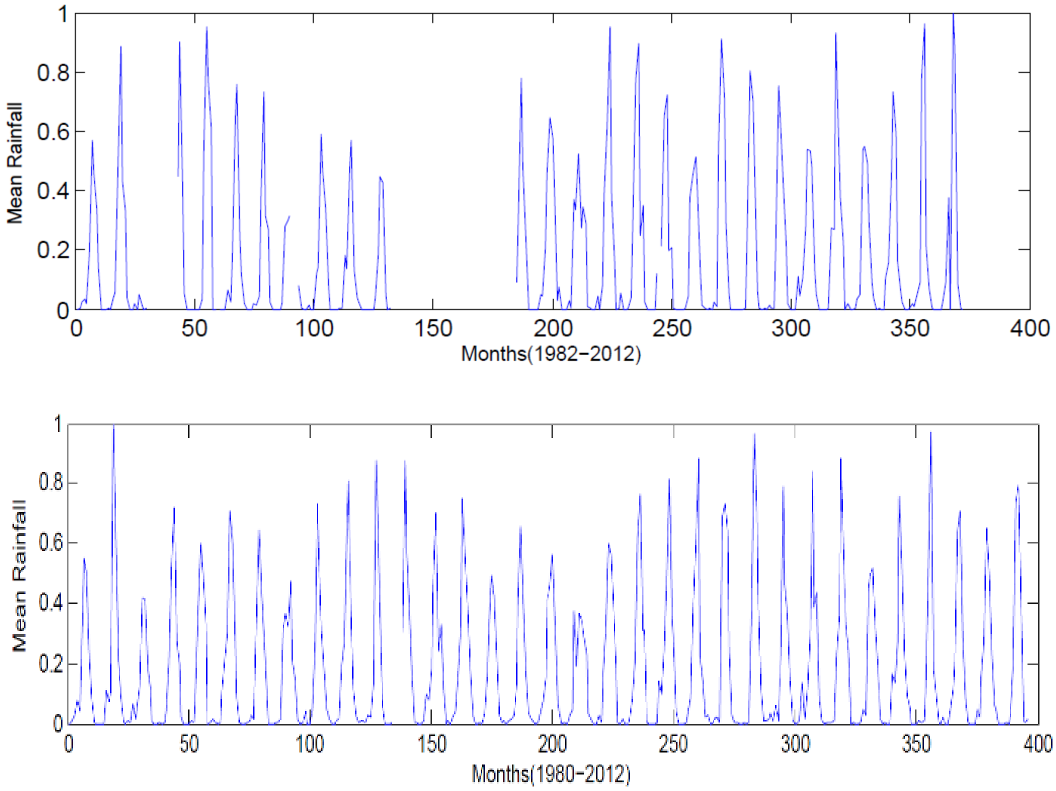

Fig.2. Zegie and Bahir Dar Station mean Rainfall missing data

5. Comparison of imputation Techniques.

In this experiment, we tested different imputation methods selected in the previous sections using the dataset of maximum temperature, minimum temperature and mean rain fall of the two weather stations, Zegie and Bahir Dar. The models were compared using different performance evaluation methods like MAPE, RMSE, Correlation coefficient and others.

These three measures are commonly used to assess the performance of machine learning for numerical values [16].

(a) **Mean absolute error (MAE)** is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error is given by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|f_i - y_i\right| = \frac{1}{n}\sum_{i=1}^{n}\left|e_i\right|.$$

As the name suggests, the mean absolute error is an average of the absolute errors $e_i = |f_i - y_i|$, where f_i is the prediction and y_i the true value. We note that the alternative formulations may include relative frequencies as weight factors.

The mean absolute error is a common measure of forecast error in time series analysis.

(b) **Mean Square Error (MSE):**This measures the average of the squares of the errors, that is, the difference between the estimator and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss or quadratic loss. The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate.

The Root Mean Square Error (**RMSE**) (also called the root mean square deviation, RMSD) is a frequently used measure of the difference between the values predicted by a model and the values actually observed from the environment that is being modelled. These individual differences are also called residuals, and the RMSE serves to aggregate them into a single measure of predictive power.

The RMSE of a model prediction with respect to the estimated variable X_{model} is defined as the square root of the mean squared error:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i}-X_{model,i})^2}{n}}$$

WhereX_{obs} is observed values and X_{model} is modelled values at time/place i.

(c) **Correlation coefficient** is a measure of association between two variables, and it ranges between –1 and 1. If the two variables are in perfect linear relationship, the correlation coefficient will be either 1 or –1.The sign depends on whether the variables are positively or negatively related. The correlation coefficient is 0 if there is no linear relationship between the variables.

Two different types of correlation coefficients are in use. One is called the Pearson product moment correlation coefficient, and the other is called the Spearman rank correlation coefficient, which is based on the rank relationship between variables.

The Pearson product-moment correlation coefficient is more widely used in measuring the association between two variables. Given paired measurements (*X*1, *Y1*), (*X*2, *Y2*)... (*Xn*, *Yn*), the Pearson product moment correlation coefficient is a measure of association given by

$$r_P = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}},$$

where$X$ and Y are the sample mean of $X1$, $X2$,..., $Xn$and $Y1$, $Y2$,..., Yn, respectively.

In this Experiment, we tested different imputation methods selected in the previous sections using the dataset of maximum temperature mean rain fall of the two weather stations: Zegie and Bahir Dar. The models were compared using different performance evaluation methods, MAE, MSE and Correlation coefficient.

To identify the model using different data (temperature and rainfall), each of the imputation methods was considered;we used a set of 12 or less lag variables as inputs from the time series data.

The best results using performance metrics of MAE, MSE, and correlation coefficient (R) for each climate variable wereinvestigated. However, there is no model that is best in all performance measuring methods. In this work, the method that best in one climate variable may not be good in another variable. In addition to this, from the time series data if one value is missed at a time, one method may be better than the method that is used to fill in values having two or more series of missing values. Holt Winter's method and ANN imputation methods are observed to be better because the Holt Winter's method considers seasonal variation.

ANN models are applied as imputation methods if the time series data has sufficient available data before missing values occurred. The results of ANNmodels are relatively better than the other methods using single step ahead and multistep ahead prediction in replacing missing values in time series data.

In summary, it can be noted that the machine learning methods specially the neural network for producing individual imputations tested are more successful in estimating the original data than the classical statistical procedures according to the testing experiment done aboveand listed inTable 1.

From the table, LOCF (Last Observation Carried Forward) method shows that its MSE, MSE error is high when compared to others but it is 100 percent correlated. In temperature or rainfall climate condition it is unlikely that the current month's temperature or rainfall is exactly similar to next month's whether condition. Therefore, we do not use this method to replace the missing value of the temperature and rainfall.

Double exponential smoothing (Holt's method) and K-NN show that their correlation is high but their error is not when compared to other models.

The last models that were used as imputation methods is triple exponential method or Holts Winter's method and ANN used to replace missing values. Missing values that occurred in the climate time series data before 130 items or instances is filled by the Holt Winter's method which records that has the required amount of length and that can sufficient for neural network training handled by the MLP neural network model. These two methods are found to be the appropriate methods to fill in missing time series data.Though, the correlation of the Holt winter's method is poor, and it is good especially in rainfall data relative to the other method.

A further investigation of the imputation methods can obtain better results to the dataset used in this work.

Table.1 Performance evaluation of Imputation methods

No	Imputation Method	Maximum Temperature			Minimum Temperature			Mean Rainfall		
		MAE	MSE	R	MAE	MSE	R	MAE	MSE	R
1	LOCF	1.1216	2.1665	1.000	1.3329	3.0963	1.00	75.1727	14849.9	1.00
2	K-NN	1.7450	4.9490	0.5905	1.1533	2.1219	0.944	129	52356	0.2935
3	Holt's Method	1.3622	2.9862	0.9903	1.2940	2.9136	0.992	70.745	14120.2	0.97677
4	Holt Winter's Method	0.7691	0.9885	0.7507	1.1782	2.8132	0.730	32.567	2952.50	0.6777
5	ANN Imputation	0.7585	1.0945	0.8278	0.7552	1.1774	0.903	29.739	884.416	0.9525

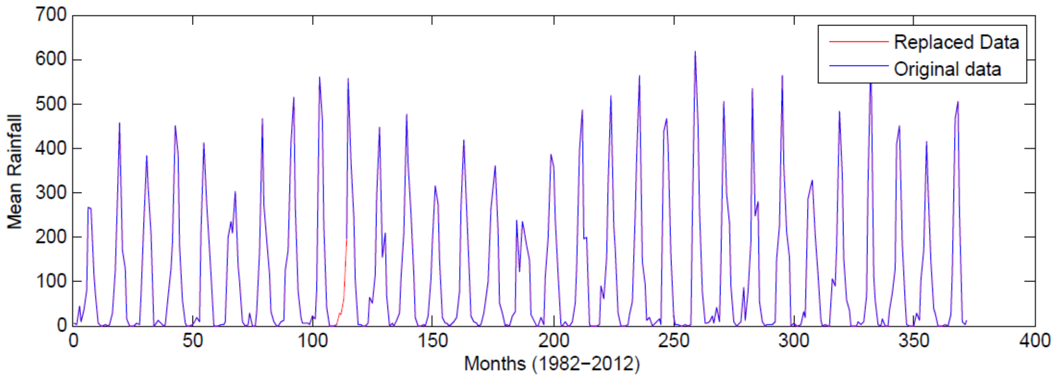

Fig. 3Bahir Dar Station maximum temperature missing data replaced

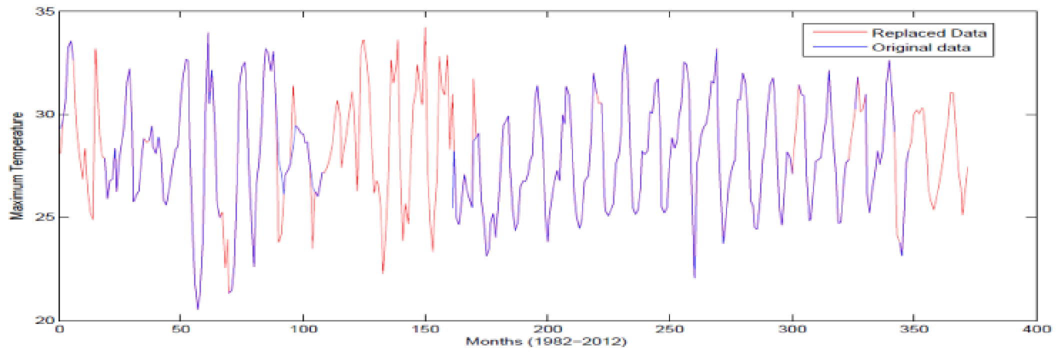

Fig. 4 Bahir Dar Station Mean rainfall missing data replaced

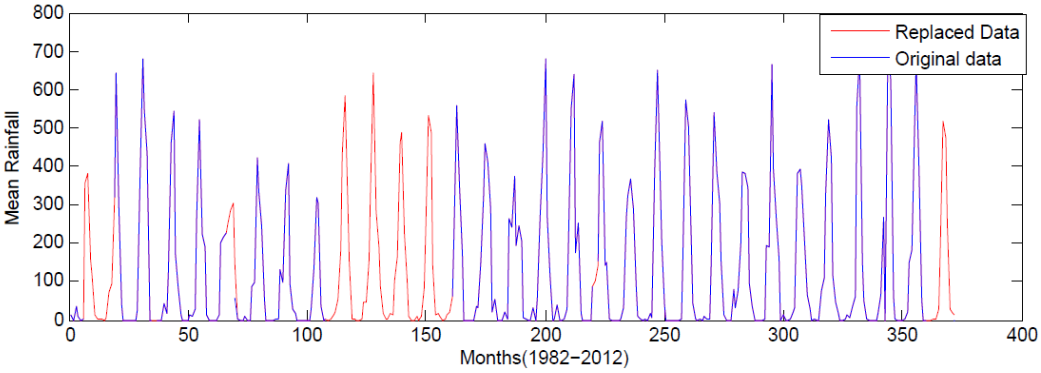

Fig.5Zegie Station Maximum temperature missing data replaced.

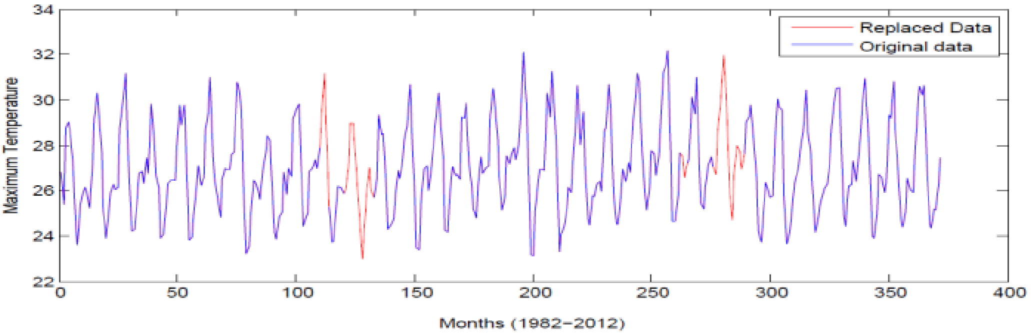

Fig. 6Zegie Station Mean rainfall missing data replaced.

Conclusion

In this paper, we investigate the climate dataset to explore the appropriate techniques for handling missing values of the temperature and rainfall dataset.

To handle missing values, we tested different imputation methods using the dataset of maximum temperature and mean rainfall of the two weather stations.

Different imputation models are compared from simple to complex such as mean/mode, Last Observation Carried Forward (LOCF), ANN imputation, K-Nearest Neighbour Imputation, Double Exponential Smoothing (Holt's Method), Triple Exponential Smoothing (Holt Winters Method).

In summary, it can be noted that the machine learning methods especially the neural network for producing individual imputations tested are more successful in estimating the original data than the classical statistical procedures according to the testing experimental done.

A further investigation of the imputation methods can achieve better results to the dataset used in this work.

References

1. S. Zhang, et al (2003): Data preparation for data mining: Applied Artificial Intelligence, 17:375–381, 2003.
2. Gustavo E. A. P. A. Batista et al (2008): An Analysis of Four Missing Data Treatment Methods for Supervised Learning, Viewed on August, 2014.
3. S. B. Kotsiantis, et al (2006), Data Preprocessing for Supervised Leaning: international journal of computer science volume 1 number 2, issn 1306-4428
4. R. Maier, C. Dandy, (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling and Software, 15:101–124.
5. T. Paraskevas, et al (2010), Use of Artificial Neural Network for Spatial Rainfall Analysis. Viewed on Jun, 2014, avaliableon, http://www.ias.ac.in/jess/forthcoming/JESS-D-13-00144.pdf
6. M.H. Beale, et al (2014), Neural Network Toolbox: User's Guide, March 2014, Viewed on September 2014, Available on; https://www.mathworks.com/help/pdf_doc/nnet/nnet_ug.pdf.
7. G. Batista et al (2003), An Analysis of Four Missing Data Treatment Methods for Supervised Learning, University of S.P.USP.
8. Daniel T. Larose (2005), Discovering Knowledge in Data: An Introduction to Data Mining, ISBN 0-471-66657-2 Copyright C_ 2005 John Wiley & Sons, Inc.
9. Joe Choong (2003), Powerful forecasting with MS-Excel.
10. E.-L. Silva-Ramírez et al (2011), Missing value imputation on missing completely at random data using multilayer perceptrons/ Neural Networks 24 (2011) 121–129.
11. R. Nkoana (2011), Artificial neural network modelling of flood prediction and early warning, Master dissertation, university of the Free State, Bloemfontein, South Africa, 2011
12. D. E. Rumelhart, et al (1986), learning representations by back-propagatingerrors, Nature, vol.323, no.6088, pp.533-536, October, 1986.
13. Mikhail Kanevski, et al (2009), Machine learning for spatial environmental data: Theory, applications and software; EPFL Press, Lausanne, Switzerland.
14. O. Antonic´ et al (2001), Spatio-temporal interpolation of climatic variables over large region of complex terrain using neural networks; Ecological Modelling138 (2001) 255–263.
15. G.Petkos (2003), Applying machine learning techniques to ecological data, master Thesis School of Informatics, University of Edinburgh, 2003.
16. Gustavo E. A. P. A. Batista et al (2008): An Analysis of Four Missing Data Treatment Methods for Supervised Learning, Viewed on August, 2014.
17. In Jae Myung(2002), Tutorial on maximum likelihood estimation, Journal of Mathematical Psychology 47 (2003) 90–100.