Imputation of missing data in time series by different computation methods in various data set applications

Dhiraj Magare[1,*], *Sushil Labde*[2], *Manoj Gofane*[3]*and Vishwesh Vyawahare*[4]

[1,2,3,4] Ramrao Adik Institute of Technology, D.Y. Patil Group, Nerul, Navi Mumbai-400 706, India

Abstract. In a modern technology generation, big volumes of data are evolved under numerous operations compared to an earlier era. However, collection of data without missing single value, is a great challenge ahead. In practice, there are many solutions suggested to avoid the missing values in time series applications. The existing methods used in imputation and their prediction with time series, varies with applications. The existing methods mostly available for imputation are least squares support vector machine (LSSVM), autoregressive integrated moving average models (ARIMA), Artificial Neural Network (ANN), Artificial Intelligence (AI) techniques, state space models, Kalman filtering and fuzzy model. The extensive experimental application data is used to analyze these methods. In addition, a synthetic set of data can also be used to forecast missing value, which improves performance of imputation methods in time series. In this paper, predominantly used imputation methods have been listed with their fundamental computational information along with their verification on set of data mentioned.

1 INTRODUCTION

A time series observations are the observations that have been taken successfully at an equal time interval. The main objective of time series forecast solely depends on the past recorded data. In case the historical data recorded includes few hours or days or months, of missing data, the parameter prediction increases the complexity. This missing data of various real time applications, influences the deviation in the actual output. Thus, this missing data plays an imperative task in many decision-making applications. The applications are monitoring the industrial activities, financial analysis, business resources, and power plant with grid control.

It would create a loss in prediction in different aspects. There are two different types of methods available for evaluation of missing data in multiple steps. Those are direct method and iterative method. The direct method evaluates the forecasting result in multiple step and reaches towards the final value while, iterative method evaluates the forecasted value iteratively until it matches the required step value. In multi-variable time series, various methods are proposed in the literature.

The various methods are such as linear interpolations, Stineman interpolations, Kalman filtering with structural model and smoothing, weighted moving average are used for estimation of missing values for solar irradiance data [1]. Further, time-series forecasting explained in [2]. In this, author had adopted wavelet model technique. The suggested methods might be infeasible to particular set of application or inefficient to predict missing data in real time. Thus, still there is a scope of improvement to exiting proposed methods, and their verification on various data sets.

As the forecasting of time series data plays a vital role, it is very important to find the accuracy of these different forecasting techniques. Different applications use different parameters for the measurement of the accuracy of these forecasting techniques. Shin-Fu Wu et al. [3] used normalized root mean squared error (NRMSE), performance measurement parameter for defining the accuracy of (LSSVM) forecasting technique of time series data. Root mean square error (RMSE) is used to evaluate the performance of the proposed algorithm for missing data estimation of synthetic multivariate time series data [4, 5, 6].

2 Methodology

In this section, most commonly used and advanced imputation methods are briefly described.

2.1. ARIMA (Autoregressive Integrated Moving Average)

ARIMA is Autoregressive Integrated Moving Average models. With the help of time-series data, statistical modeling technique predicts the future values, so this model is used in the field where short term forecasting is needed. It needs minimum 40 past data points. If the data is reasonably extended and the correlation between the past data points is steady, ARIMA model is more efficient compared to the exponential smoothing. Flowchart of Autoregressive Integrated Moving Average

[*] Corresponding author: dhiraj.magare@rait.ac.in

model shown in Fig. 1. With the help of ARIMA model, stationary time series data can be described as a function of auto regression and moving average parameters.
In flowchart p, d, q are order of AR model, order of MA model, and order of difference respectively (to make non stationary data to stationary data).
AR model is represented by:

$$Z(t) = C(1) * Z(t-1) + G(t) \qquad (1)$$

MA model is represented by:

$$Z(t) = D(1) * G(t-1) + G(t) \qquad (2)$$

where, G(t) = error parameter, Z(t) is time series data C and D are constants.
Vichaya L. et al. [7] proposed forecasting technique namely, seasonal autoregressive integrated moving average (SARIMA), for missing data imputation, which uses mean value of the global horizontal irradiance (GHI) averages over data. A comparative study of different forecasting models such as neural network (NN), autoregressive and moving average (ARMA), coupled autoregressive and dynamical system (CARDS) is done for forecasting of 1 to 6 hr ahead [8].

2.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is effective Deep Learning method. The theory of statistics is the basis of machine learning. It is basically used for classification problems and forecasting of time series data. SVM was first developed at AT&T Bell Laboratory in 1970 by Vapnik and colleagues. Forecasting of time series data is difficult with the missing data. Replacing missing data with values obtained by imputation method and ignoring missing data, will affect the performance of forecasting of time series data. The principle of SVM is to discover function which establishes a relationship between input P and output Q.Q=f(P) from set of data points.
Consider input data points

$$F= \{(P_t, (Q_t)\} \in R \qquad i=1: n \ (3)$$

$$f(x) = w\phi(x) + b \qquad b \in R \ (4)$$

F:time series data
R: Real set
W: Noise Matrix
b: Constant
$\phi(x)$ is nonlinearly mapped higher dimensional feature space from input space

Shin-Fu Wu et al. [3] suggested a machine-learning tool LSSVM (Least square support vector machine). It is applied to time series forecasting with missing data. For forecasting, time series data and local time indexes (LTI) are fed to LSSVM. The results so obtained are compared with the forecasting performance of other imputation methods.

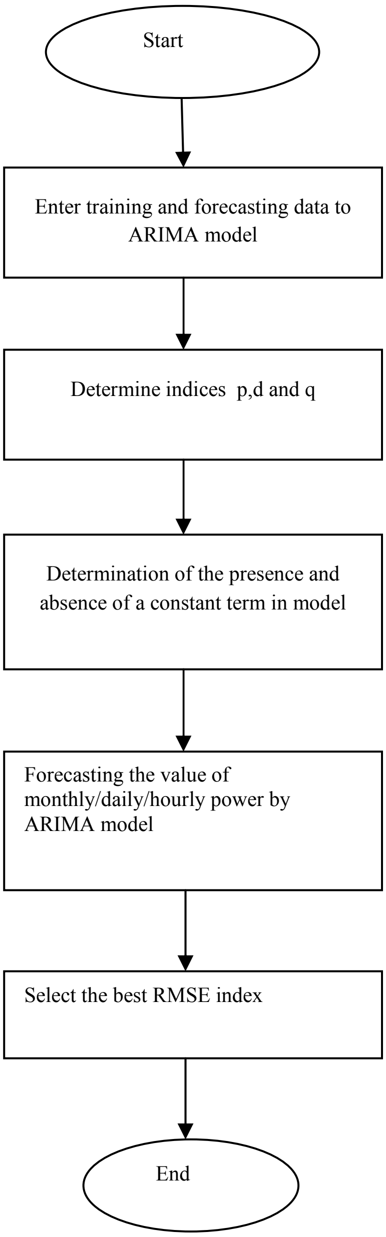
Start
Enter training and forecasting data to ARIMA model
Determine indices p,d and q
Determination of the presence and absence of a constant term in model
Forecasting the value of monthly/daily/hourly power by ARIMA model
Select the best RMSE index
End

Fig. 1 Flowchart of Autoregressive Integrated Moving Average model

2.3 ANN (artificial neural network)

Recently, ANN (artificial neural network) is most widely used machine learning algorithm that is based on biological nervous systems. Neural network consist of elements called neutron operating in parallel. Flowchart of artificial neural network shown in Fig. 3. Neuron is a fundamental information processing part of neural network. It consists of three basic elements; a set of weights, an adder for adding the inputs, and activation function for limiting the amplitude. The neural network is getting rapid importance because of its capability to offer solution to variety of problems. Flowchart of artificial neural network shown in Fig. 3. Badia Amrouche et al. [9] proposed novel approach, which is combination of artificial neural network and spatial modeling for forecasting global solar irradiance.

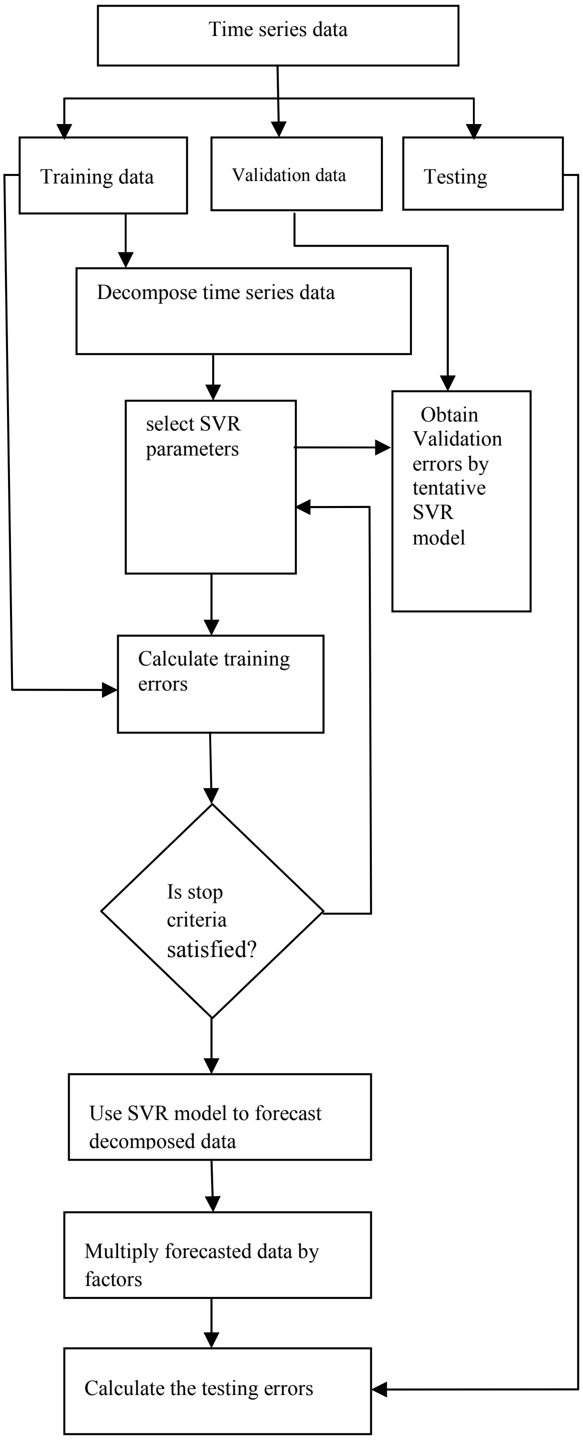
Time series data
Training data
Validation data
Testing
Decompose time series data
select SVR parameters
Obtain Validation errors by tentative SVR model
Calculate training errors
Is stop criteria satisfied?
Use SVR model to forecast decomposed data
Multiply forecasted data by factors
Calculate the testing errors

Fig. 2 Flowchart of support vector machine

The application area of ANN is very wide. MLP (Multi-layer-Perceptron) is one type of such neural technique. MLP with back propagation ANN model is mostly applied in various types of real time examples. MLP Model does not assume any prior information regarding data distribution as many statistical models do.

The model consists of a system of interconnected nodes or neurons and weights; it represents the non-linear relationship between input and output. A MLP may have one or more hidden output of a neuron. The model is trained to perform the particular function by adjusting the weights between the elements. Neural networks are trained or re-estimated in such a type that the specific input leads to specific output target. An ANN model consists of input, output layer, and hidden layers. The layers may vary for different applications. The model has capability to learn through training. Training needs a set of data. To minimize the error, training is done to find the different combination of weights. Adel Mellit a, et al. [2] proposed a practical method for forecasting the time series data of solar radiation using ANN. Further, ANN used in forecast irradiance [10].

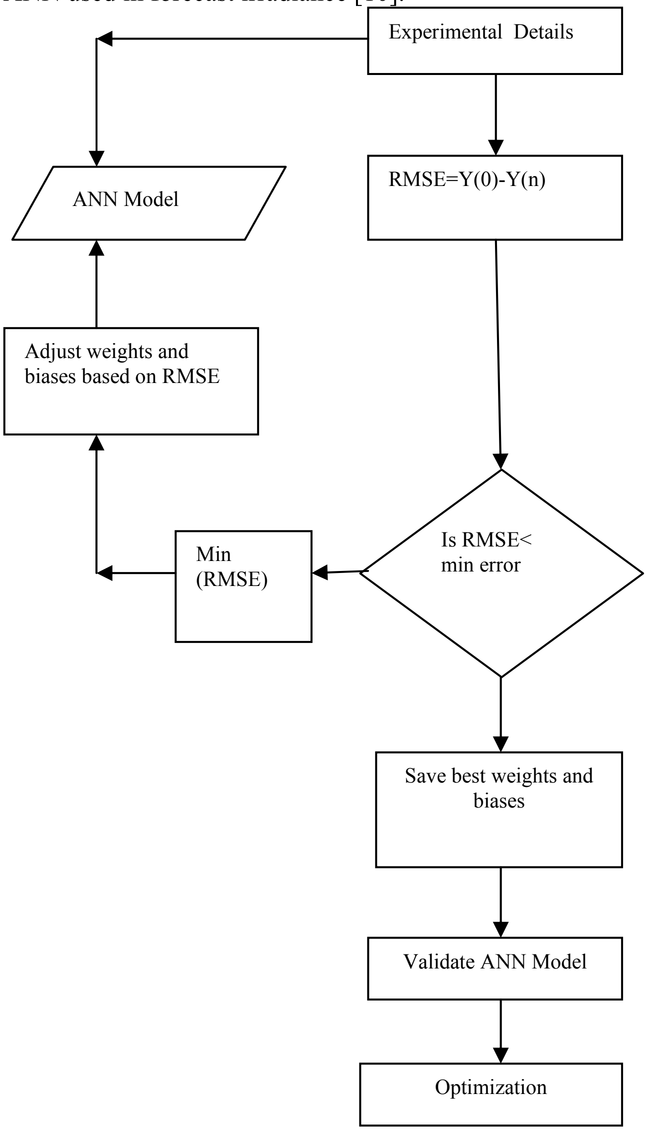

Fig. 3 Flowchart of artificial neural network.

2.4 KALMAN FILTER

The optimal estimation of states of the system can be implemented by kalman filtering algorithm, with the help of least squares approximation or maximum likelihood estimation. Due to this, the accuracy of time series forecasting improves. Flowchart of generalised kalman filter shown in Fig. 4.

The algorithm is based on the discretization of linear dynamic system. It has capability to predict the current state with the help of measurement and with predicted state from previous time step. The model can be mathematically represented as :

$$X_t = A_t X_{t-1} + W_t \qquad (5)$$

$$Y_t = H_t X_t + V_t \qquad (6)$$

where X_t : State vector at time t.

A_t: State transition matrix.

Y_t: The vector of measurement.

H_t: Transformation matrix.

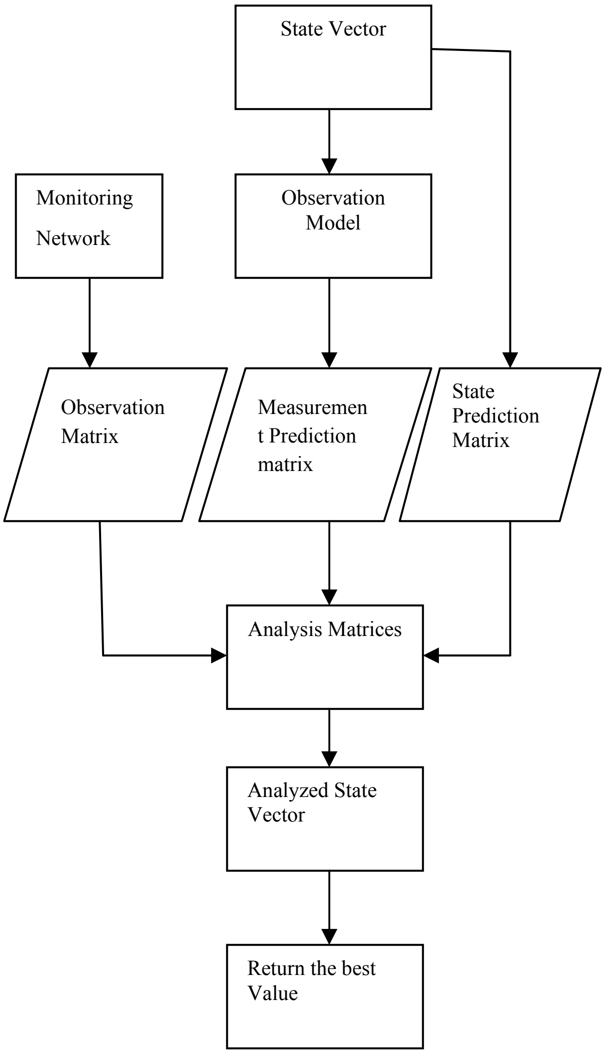

Fig. 4 Flowchart of generalised kalman filter.

3 Results and Discussion

In this section, mostly used important imputation methods have been briefly explained. Imputation methods along with data used in that particular method have been list out in Table 1. The performance parameter also is incorporated in Table 1.

Wu et. al. [3] discussed that LTI offers comparable or even better NRMSE values. Furthermore, LTI runs much faster, compared with other imputation methods, including mean, hot deck and auto regression (AR). Shi. et. al. [4] infers that even for high (90%) missing ratio the RMSE calculated would be small using proposed methods. Hence, this proposed method can be an alternative model to find the missing values in time series under large scaled multivariable system. The suggested method can be used in various applications such as, electric equipment monitoring, climate or financial forecasting. The environment state monitoring, security inspection etc. are some of the applications. Mellit et. al. [2] proposed prediction of 24 hour in advance of solar irradiance is possible with Multi layer Perceptron model with consideration of mean value of solar irradiance. The air temperature, as well as day of month also been included in forecasting. The MLP-forecaster finds application in GCPV plant, solar irradiance forecasting of (24 hour in advance), renewable systems etc. Gao et. al. [11] discussed methods for estimating missing data in case of sensor failures. Guo et. al. [12] discussed that the published algorithm proved with considerable performance. The verification had been applied on the missing data, which is at random fashion. This algorithm is more appropriate for applications with high computational error requirements. Demirhan et. al. [1] showed in case when the rate of missingness is enlarged to 50 %, the special forecasting algorithm needed to follow. For solar radiation forecasting, non-seasonal Stineman interpolation, Kalman filtering in addion of ARIMA, and smoothing work with good agreements. ARIMA can also been used in marine [11, 13]. Anava Oren et. al. [14] discussed model which is used in biomedical application. Application of adaptive wavelet network architecture for forecasting solar radiation on daily basis explained in [2, 15]. Akarslan et. al. [16] showed the predictions possibility with linear prediction filters over the measured values. The prediction of in-plane irradiation with a neural network approach has been most common [17, 18].

Table 1 Imputation methods along with data set and performance parameter.

Ref. no.	Methods	Data set used	Performance parameter
[1]	linear and Stineman interpolations and Kalman filtering with structural model and smoothing, weighted moving average	short to mid-term horizontal solar irradiance data	Not Applicable
[2]	Multilayer Perceptron MLP-model	Solar irradiance. Air temperature. These data covers July 1st 2008 to May 23rd 2009. Also, November 23rd 2009 to January 24th 2010. Both have been used.	MAE: < 5% verified on the correlation coefficients arranged from 90-92%

[3]	Least squares support vector machine (LSSVM).	They are tested on 3 generated datasets (sine function, sinc function, and Mackey-Glass chaotic time series) and 4 benchmark datasets Poland electricity load, Sunspot, Jenkins-Box, and EUNITE competition).	Not Applicable
[4]	Improved matrix factorization techniques. : 1) MFS: Matrix Factorization with Smoothness constraints; 2) CSM: Correlated Sensors based Matrix factorization; 3) USM: Uncorrelated Sensors based Matrix factorization; 4) CSMS: Correlated Sensors based Matrix factorization with Smoothness constraints; 5) USMS: Uncorrelated Sensors based Matrix factorization with Smoothness constraints;	Synthetic multivariate time series data.	NRMSE:0.01-0.2 for LTI & hot deck
[6]	Multivariate Imputation by Chained Equations (MICE)	Central Marine Fisheries Research Institute (CMFRI) is the source of past data. Data as old as 20 years may be collected.	MSE: 0.042 verified on ARIMA model.
[7]	seasonal autoregressive integrated moving average models (SARIMA)	Not Applicable	RMSE:0.86 Verified on proposed method II, MAE: 0.57
[8]	MTSDI(multi variate time series data imputation)	The daily concentrations of particulate matter with an aerodynamic volume up to 10 mg/m3 was collected for 366 days (PM10), which were measured at ten monitoring stations in the city of Brazil.	Not Applicable
[9]	Permutation of spatial modelling. Also, Artificial neural networks (ANNs) method	GHI data	highest MSE: 53.97 W/m2 ,lowest MSE: 16.46 W/m2 VERIFIED on GHI
[10]	fuzzy and neural networks, stochastic models: : AR, ARMA, ARIMA and Markov chain , (AI) techniques (ANN), fuzzy model	solar radiation data	MAPE: 6.03%-9.65%. These were checked on proposed method. Those included fuzzy and neural network scheme.
[11]	Gaussian function as well as Cosine function used. ANN with feed-forward.	Solar radiation and indoor temperature of tertiary buildings	MRE(%): 8 Gaussian approach
[12]	Kronecker Compressive Sensing theory.	MOTES, GSA and SST	RMSE: 0.005 Verified on gas sensor array dataset, ARTS: 0.05 verified on MOTES dataset (back propagation method)

[13]	LMQR, weighted quantile regression. (WQR) . Quantile regression neural network (QRNN). Recursive generalized AR conditional hetero-skedasticity (GARCHrls).	Not Applicable	Not Applicable
[14]	Yule-Walker, EM (Kalman filter),	Not Applicable	MSE:0.0979for 10% missing data verified on proposed algorithm
[15]	adaptive wavelet-network architecture	Solar-radiation data from Algeria. meteorological station 1981 - 2001	MSE:0.05 structure 5 (for wavelet 5x12x1)
[16]	Linear prediction filters & empirical model method used.	solar radiation data	RMSE (%) :34.86 Verified on Proposed Adaptive Approach
[17]	Neural networks.	Datasets available from meteorological lab.	MBE(%): 0.8, and energy output 0.9 verified on In-plane irradiation low as 0.9
[18]	Back-filling of missing data.	Synthetic dataset	RMSE (%): 2.8, MBE (%):1.5 verified on horizontal irradiation.

4 Conclusions

LSSVM is used in different fields of applications, such as time series prediction and financial forecasting [3]. Mellit et. al. [2] proposed MLP forecaster which finds application in GCPV plant, solar irradiance forecasting of (24h ahead) and, renewable systems etc. Junger et. al. [8] discussed an imputation based method. This method, when compared with GAM method (GAM is based on use of ARIMA) gave good results. ARIMA can also been used in marine [11, 13]. Anava Oren et. al. [14] discussed model that applied to various applications. Such as, in DNA microarray, market analysis as well as noise uncertainty reduction. Amrouche et. al. [9] proposed a novel technique. This technique applied to predict daywise global horizontal radiation. Thus, there is still a scope of improvement in the existing proposed methods. A new proposed method when verified on existing data sets may give better results for a particular application.

References

1. H. Demirhan, Z. Renwick, Missing value imputation for short to mid-term horizontal solar irradiance data. Appl Energy, vol. **225**, pp. 998–1012, (2018). 10.1016/j.apenergy.2018.05.054.
2. A. Mellit, A.M. Pavan, A 24-h forecast of solar irradiance using artificial neural network: application for performance prediction of a grid-connected PV plant at Trieste, Italy. Solar Energy, vol. **84(5)**, pp. 807-821, (2010). 10.1016/j.solener.2010.02.006.
3. S. Wu, C. Chang, S. Lee, Time Series Forecasting with Missing Values, *1st Int. Conf. Ind. Networks Intell. Syst.*, pp. 151-156, (*2015).* 10.4108/icst.iniscom.2015.258269.
4. W. Shi et al., Effective prediction of missing data on Apache Spark over multivariable time series", *IEEE Trans. Big Data, vol. **6**, pp. 57239-27248,* (2017). 10.1109/TBDATA.2017.2719703.
5. Yanjie Wei et al., Any-time Methods For Time-series Prediction With Missing Observations, *IEEE International Congress on Big Data (BigData Congress)*, (2017). 10.1109/BigDataCongress.2017.62.
6. M.O.D. Rizwan et al., A Novel Approach For Time Series Data Forecasting Based On Arima Model For Marine Fishes, International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), (2017). 10.1109/ICAMMAET.2017.8186707.
7. V. Layanun, S. Suksamosorn, J. Songsiri, Missing-data Imputation for Solar Irradiance Forecasting in Thailand. *In: Proceedings of the SICE Annual Conference 2017 September 19–22, Kanazawa University, Kanazawa, Japan*, (2017). 10.23919/SICE.2017.8105472.
8. W.L. Junger, A.P. de Leon, Imputation of missing data in time series for air pollutants, Atmospheric Environment, vol. **102**, pp. 96-104, (2015). 10.1016/j.atmosenv.2014.11.049.
9. B. Amrouche, X. Pivert Le, Artificial neural network based daily local forecasting for global solar radiation. Appl Energy vol. **130**, pp. 333–341, (2014) 10.1016/j.apenergy.2014.05.055.
10. S.X. Chen, H.B. Gooi, M. Wang. Solar radiation forecast based on fuzzy logic and neural networks. Renew Energy, vol. **60**, pp. 195–201, (2013). 10.1016/j.renene.2013.05.011.
11. Z. Gao, W. Cheng, X. Qiu, L. Meng, A Missing Sensor Data Estimation Algorithm Based on Temporal and Spatial Correlation, *International Journal of Distributed Sensor Networks*, vol. **2015**, pp. 1-10, (2015). 10.1155/2015/435391.

12. Y. Guo , X. Song , D. Fang, An Efficient Missing Data Prediction Method Based on Kronecker Compressive Sensing in Multivariable Time Series.IEEE Trans, pp. 57239 – 57248, (2018). 0.1109/ACCESS.2018.2873414.

13. M. David, L. Mazorra, P. Lauret, Comparison of intraday probabilistic forecasting of solar irradiance using only endogenous data. Int J Forecast, vol. **19(4)**, pp. 299–311, (2018). 10.1016/j.ijforecast.2018.02.003 .

14. A. Oren, H.. Elad, Z Assaf, Online time series prediction with missing data. *In Proceedings of the 32nd International Conference on Machine Learning*, vol. **37**, pp. 2191–2199, (2015).

15. S.A. Kalogirou, Artificial neural networks in renewable energy systems applications: a review. Renew Sustain Energy Rev, vol. **5(4)**, pp. 373–401, (2001). 10.1016/S1364-0321(01)00006-5.

16. E. Akarslan, FO Hocaoglu. A novel adaptive approach for hourly solar radiation forecasting. Renew Energy, vol. **87**, pp. 628–633, (2016). 10.1016/j.renene.2015.10.063.

17. E. Koubli, D. Palmer, T. Betts, P. Rowley, R. Gottschalg, Inference of missing PV monitoring data using neural networks*, 43rd IEEE Photovoltaic Specialists Conference, PVSC, Portland*, pp. 1-6, (2016). 10.1109/PVSC.2016.7750305 .

18. E. Koubli, D. Palmer, P. Rowley, R. Gottschalg. Inference of missing data in photovoltaic monitoring datasets, IET Renew. Power Gener., vol. **10(4)**, pp. 434-439, (2016). 10.1049/iet-rpg.2015.0355