SIMPLE APPROACHES TO MISSING DATA FOR ENERGY FORECASTING APPLICATIONS

Kasım Zor[1], Özgür Çelik[1], Oğuzhan Timur[2], Hatice Başak Yıldırım[3], Ahmet Teke[2]
[1]Department of Electrical and Electronic Engineering, Adana Science and Technology University
[2]Department of Electrical and Electronics Engineering, Çukurova University
[3]Department of Energy Systems Engineering, Adana Science and Technology University
Corresponding author: Kasım Zor, e-mail: kzor@adanabtu.edu.tr

REFERENCE NO

FORC-03

Keywords:
Missing data, energy forecasting, linear interpolation, marginal mean imputation.

ABSTRACT

Energy forecasting is not only a prominent sub-discipline in energetics, but also an arduous challenge to acquire electrical and climatological data that might have some missing values frequently due to power outages or equipment failures. There are several methods in the literature to treat missing data in datasets. Applying each method causes different accuracy of results in performance metrics. In this paper, a comparison of simple approaches named as linear interpolation method and marginal mean imputation method to missing data for energy forecasting applications using multilayer perceptron neural networks (MLPNN) and support vector machines (SVM) is presented through a case study of electrical energy consumption data and climatological data of a hospital in the Mediterranean Region.

1. INTRODUCTION

Datasets are inseparable parts of energy forecasting studies and tackling the presence of missing values in the datasets improves the quality of data pre-processing while enhancing the accuracy. Missing data are ubiquitous [1] that reveals not only in energy forecasting applications, but also appears in many real-world situations such as surveys [2], control-based applications [3-5], wireless sensor networks [6-7], financial and business applications [8-9], biological researches [10-11], and so on.

In data acquisition stage of energy forecasting, missing data can show up owing to a power outage or a malfunctioned sensor. Most decision-making tools such as the commonly used MLPNN, SVM, and many other machine learning techniques cannot be used for decision making if data are incomplete [12].

In this paper, a very short-term energy forecasting of a hospital in the Mediterranean Region is performed by using MLPNN and SVM, and employing both linear interpolation method and marginal mean imputation method separately for treating missing data belonging to electrical energy consumption, temperature, and humidity values. After this introductory section, this paper is organised as follows. Missing data mechanisms, linear interpolation and marginal mean imputation methods, and dataset contents are presented in the materials and methods. A comparison table of accuracy results according to mean absolute percentage error (MAPE) as performance metric is given in the results. The paper is concluded with conclusions section.

2. MATERIALS AND METHODS

In the literature, there are three missing data mechanisms named as missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). First of all, data are MCAR when the probability of a case having a missing value for a variable does not depend on either the known values or the missing data. Secondly, data are MAR when the probability of a case having a missing value for a variable may depend on the known values but not on the value of the missing data itself. Lastly, the pattern of missing data is non-random and depends on the missing variable. In this situation, the missing variable in the NMAR case cannot be predicted only from the available variables in the database [13].

Missing data in the context of this study is occurred due to power outage in the hospital. For this case, the actual variables where data are missing are not the cause of the

incomplete data. Instead, the cause of the missing data is due to some other external influence (power outage) which obviously states that missing mechanism for this case is MAR [14].

In the literature, conventional approaches except listwise deletion to missing data can be basically classified as linear interpolation method and marginal mean imputation method [15]. Firstly, the simplest form of interpolation is called as linear interpolation [16] which is used to connect two data points with a straight line. Using similar triangles,

$$\frac{f_1(x)-f(x_0)}{x-x_0} = \frac{f(x_1)-f(x_0)}{x_1-x_0}(x-x_0) \quad (1)$$

which can be arranged to yield

$$f_1(x) = f(x_0) + \frac{f(x_1)-f(x_0)}{x_1-x_0}(x-x_0) \quad (2)$$

which is the formula of linear interpolation. The notation $f_1(x)$ indicates that this is a first-order interpolating polynomial. In general, the smaller the interval between the data points result in the better the approximation. This is due to the fact that, as the interval decreases, a continuous function will be better approximated by a straight line [17]. In the latter case, missing values are imputed using the average of the observed values for that variable in marginal mean imputation. It is also sometimes referred to as simple mean imputation or just mean imputation [18]. The equation of marginal mean imputation is given as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (3)$$

where n is the number of available data and y_i is the data points [19].

Dataset used throughout this paper contains data between the time interval of 2 October and 29 October 2017. The averaging period of the dataset is 10-minute. The contents of the dataset are presented in Table 1. Missing electrical energy consumption data of the dataset belong to 2 October 2017 between 12:50-13:40 and 14 October 2017 between 07:50-09:10 is emphasised and illustrated in Figure 1. In Figure 2 and 3, real data of the hospital is demonstrated in comparison with the implementation of linear interpolation method and daily marginal mean imputation method in MATLAB environment.

Table 1. Contents of the dataset.

Input Parameters	Description
Electrical Variable	✓ Historical Electrical Energy Consumption Data for the last 10-minute
Calendar Variables	✓ Week ✓ Day ✓ Time Sample
Climatological Variables	✓ Outdoor Temperature ✓ Relative Humidity

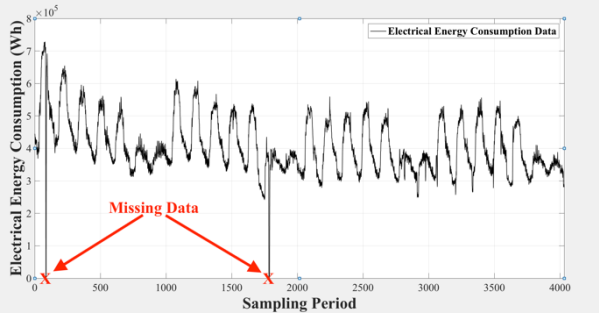

Fig. 1. Missing electrical energy consumption data

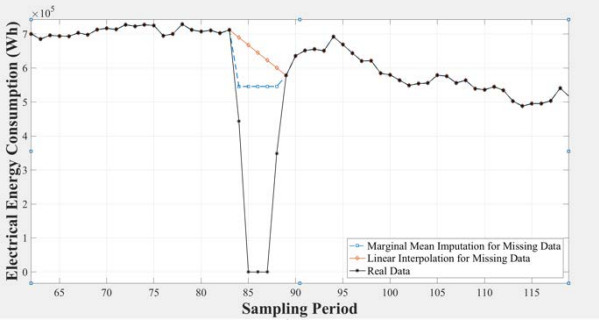

Fig. 2. Missing data belong to 2 October 2017

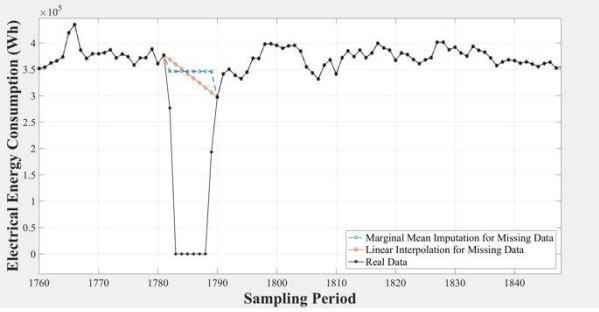

Fig. 3. Missing data belong to 14 October 2017

Before performing MLPNN and SVM based models, the original dataset involving real data is

turned into two different datasets named as interpolation and imputation datasets for the application of linear interpolation and marginal mean imputation methods.

For MLPNN based model, 10-fold cross validation technique is used as validation method. A search is carried out for the number of neurons in the hidden layer from 2 to 20 in order to find the model topology having the best performance and the optimal size for the neuron number in the hidden layer is found as 11 because of having the lowest residual variance at that neuron number. Logistic and linear activation functions are used for the hidden layer and output layer respectively. Scaled conjugate gradient backpropagation algorithm is utilised for the training method of MLPNN to optimise the weight values.

For SVM based model, epsilon type support vector regression (ε-SVR) is used with 10-fold cross validation technique. Gaussian radial basis function (RBF) is employed as the kernel function for the model. A grid search of the model parameters is performed for cost parameter *C*, regularisation parameter gamma (γ), and ε parameter in order to minimise mean square error (MSE) and overcome the generalisation problem.

3. RESULTS

Table 2. Performance comparison of different approaches to missing data for MLPNN and SVM.

Missing Data Approach	AI Technique	MAPE
Marginal Mean Imputation	SVM	2.5128
Linear Interpolation	SVM	2.5288
Linear Interpolation	MLPNN	2.8775
Marginal Mean Imputation	MLPNN	2.9025

Evaluating the performances of different artificial intelligence methods on datasets, the the most commonly used performance metric for error calculation in the energy forecasting literature is MAPE which can be formulated as [20]

$$\mathrm{MAPE}(\%) = \frac{\sum_1^n |(X_t - X_t')/X_t|}{n} \times 100 \qquad (4)$$

where X_t represents actual or measured output, X_t' shows predicted output, and *n* indicates the number of observations.

The results of the different approaches to missing data for MLPNN and SVM is given in Table 2.

4. CONCLUSIONS

Missing data is crucial especially for the recent big data phenomena. To handle with missing data, linear interpolation and marginal mean imputation methods are conventionally utilised in the literature.

In this paper, both of linear interpolation and marginal mean imputation methods are applied to the dataset for MLPNN and SVM methodologies. The results show that marginal mean imputation approach to missing data gives the best results operating with support vectors. In contrast, linear interpolation approach to missing data indicates better performance metric results for one hidden layer feed-forward MLPNN.

The authors believe that this paper will fill the gap in the energy forecasting literature for approaches to missing data, and will guide and help prospective researchers in the field.

Acknowledgements

The authors would like to acknowledge the Scientific Research Project Unit of Çukurova University owing to financial support for the individual research project named as “Electric Demand Prediction: Data Acquisition, Implementation of ANN and User Interface Design” and numbered as “FBA-2017-8252”.

References

[1] Zhang, Z., Missing data imputation: focusing on single imputation, *Annals of Translational Medicine*, Vol. 4, Iss. 1, 2016, pp. 1-8.

[2] Wang, L. and Fan, X., Missing Data in Disguise and Implications for Survey Data Analysis, *Field Methods*, Vol. 16, No. 3, 2004, pp. 332-351.

[3] Lakshminarayan, K., Harp, S. A., and Samad, T., Imputation of missing data in industrial databases, *Applied Intelligence*, Vol. 11, 2004, pp. 259-275.

[4] Ji, C. and Elwalid, A., Measurement-based network monitoring: missing data formulation and scalability analysis, IEEE International Symposium on Information Theory, Sorrento, Italy, 2000, pp. 78.
[5] Nguyen, L. N. and Scherer, W. T., *Imputation Techniques to Account for Missing Data for in Support of Intelligent Transportation Systems Applications*, Research Report No. UVACTS-13-0-78, University of Virginia, 2003.
[6] Halatchev, M. and Gruenwald, L., Estimating Missing Values in Related Sensor Data Streams, 11th International Conference on Management of Data, Goa, India, 2005, pp. 83-94.
[7] Mohammed, H. S., Stepenosky, N., and Polikar, R., An Ensemble Technique to Handle Missing Data from Sensors, IEEE Sensors Applications Symposium, Houston, Texas, USA, 2006, pp. 101-105.
[8] DiCesare G., *Imputation, Estimation and Missing Data in Finance*, Ph.D. Thesis, University of Waterloo, Ontario, Canada, 2006.
[9] Kofman, P. and Sharpe, I. G., Using Multiple Imputation in the Analysis of Incomplete Observations in Finance, *Journal of Financial Econometrics*, Vol. 1, Iss. 2, 2003, pp. 216-249.
[10] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B., Missing value estimation methods for DNA microarrays, *Bioinformatics*, Vol. 17, No. 6, 2001, pp. 520-525.
[11] Kim, H., Golubi G. H., and Park, H., Imputation of missing values in DNA microarray gene expression data, IEEE Computational Systems Bioinformatics Conference, Stanford, CA, USA, 2004.
[12] Olivas, E. S., Guerrero, J. D. M., Sober, M. M., Benedito, J. R. M. and Lopez, A. J. S., *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, 1st Ed., Information Science Reference, 2010.
[13] Schmitt, P., Mandel, J., and Guedj, M., A Comparison of Six Methods for Missing Data Imputation, *Journal of Biometrics & Biostatistics*, Vol. 6, Iss. 1, 2015, pp. 1-6.
[14] Little, R. J. A. and Rubin, D. B., *Statistical Analysis with Missing Data*, 2nd Ed., John Wiley & Sons, 2002.
[15] Soley-Bori, M., *Dealing with missing data: Key assumptions and methods for applied analysis*, Technical Report No. 4, Boston University, 2013.
[16] Kasam, A. A., Lee, B. D., and Paredis, C. J. J., Statistical methods for interpolating missing meteorological data for use in building simulation, *Building Simulation*, Vol. 7, 2014, pp. 455-465.
[17] Chapra, S. C. and Canale, R. P., *Numerical Methods for Engineers*, 6th Ed., McGraw-Hill, 2010.
[18] Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A. and Verbeke, G., *Handbook of Missing Data Methodology*, 1th Ed., CRC, 2015.
[19] Noor, N. M., Abdullah, M. M. A. B., Yahaya, A. S. and Ramli, N. A., Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Dataset, *Materials Science Forum*, Vol. 803, 2014, pp. 278-281.
[20] Zor, K., Timur, O., Çelik, Ö., Yıldırım, H. B., and Teke, A., Interpretation of Error Calculation Methods in the Context of Energy Forecasting, 12th Conference on Sustainable Development of Energy, Water and Environment Systems, Dubrovnik, Croatia, 2017, Vol. 0722, pp. 1-9.