Application of data analytics for information retrieval from a typical DSO's database

DOI:
10.1109/ISGTEurope.2016.7856314

Document Version
Accepted author manuscript

Application of Data Analytics for Information Retrieval from a Typical DSO's Database

J. Ponoćko, J. V. Milanović, R. Preece
Electrical energy and Power Systems Group
University of Manchester
Manchester, UK
jelena.ponocko@manchester.ac.uk

N. C. Woolley
National Grid
London, UK

***Abstract*—This paper introduces the reasons for big data analytics in distribution network studies and potential benefits it could give. Summary of the most common data mining methods used in power system studies is also given, followed by a comparative analysis. A use case is shown at the end in order to present some examples of extraction of useful information from raw data stored in a real distribution utility's database. This was done by using some of the basic data mining methods applied to different types of attributes describing distribution system feeders in 11 kV and 6.6 kV network. The initial results showed that the usefulness of information depends on the level of data aggregation, as well as the choice of data analytics method.**

***Index Terms*—Databases, data mining, power distribution.**

I. INTRODUCTION

Following the low carbon generation technologies (LCT) introduction to power systems, distribution networks are becoming more active in balancing generated and consumed electrical energy. This is done through participation of the end-users through demand side management (DSM). In order to maintain secure response of the end-users during DSM actions, it is necessary to provide high reliability of the distribution network (voltage levels not higher than 132 kV, in case of the UK).

One of the conditions for the network reliability is optimal maintenance of its main assets. Data about the condition of the assets is mainly stored in distribution system utilities' databases. These data can be static, i.e. data about the previous electrical measurements or replacement of an old asset with a new one. This type of data is usually stored in the form of reports, containing both numeric and text data, often given in tables. Databases also collect dynamic, mainly numeric data coming in data streams, such as on-line monitoring data for power transformers or load demand at a substation point. This type of data can be collected at various time steps, depending on the application. Distribution system utilities' servers are keeping large amount of static data and also constantly receiving and storing large number of real-time data, keeping enormous memory space busy with numbers and text. It has not been analysed yet, though, how thoroughly these data are being processed, i.e. how much knowledge has been retrieved from the existing collection of data.

With the increased involvement of information technologies (IT) and significant reliance on monitoring systems and processing of large number of data streams, a need for utilization of data mining techniques has been raised in distribution network analysis. Therefore, methodologies of big data analytics should be developed and applied on huge available data streams, as well as on static data, to investigate potential correlations between condition of the network and some other factors, such as equipment type, geographical position, weather conditions, socio-demographic profile of an area, etc. In this way, it would be possible to manipulate some of the critical factors in order to provide secure and stable operation of the distribution network.

This paper introduces reasons for big data analytics in distribution network and presents a case study showing results of the data analysis applied to a real distribution utility's database.

II. REASONS FOR BIG DATA ANALYTICS IN DISTRIBUTION SYSTEM ANALYSIS

A. Distribution Network Observability

An important source of uncertainty in distribution grid analysis is the actual observability of the distribution system, i.e. the ability to perceive the real state of the network reliability or quality of service (QoS) based on the data coming from monitoring devices. Distribution system operator (DSO) controls a much larger number of power lines and substations than a transmission system operator (TSO), which is why it is hardly feasible to monitor all these assets. LV network in the UK, for example, involves 230,000 HV/LV substations, including 580,000 transformers and 376,000 km of overhead lines and underground cables [1].

A lot of effort has been put to maintain the optimal control of the distribution system despite the reduced observability. Hence the use of data mining methods is taking the lead as a cost-effective means of gaining additional and useful information from raw data. For example, [2] proposes making

This research is partly supported by the EU Horizon 2020 project "Nobel Grid", contract number 646184

correlations between monitoring data and characteristics of feeders as a solution to the problem of estimating number of customers connected to a feeder and mixture (percentages) of customers' types. This way it would be feasible to assess more accurattely the current state of feeders and customer types connected to them by using monitoring data only.

B. Future Distribution Power Network and Big Data

One of the most important features of the future power grid operation will be the increased use of communication and information tecnologies. Hence, with the complexity of modern distribution power systems grows the size of their monitoring systems and databases accommodating variety of data coming from numerous monitors and sensors. Databases are increasing in two dimensions: in the number of objects (instances) and in the number of fields for attributes describing those objects. The most common types of attributes typical data mining methods are dealing with are numerical and nominal ones.

Power distribution utilities' databases are, as many other types of databases, characterised by several common features [3]:

- large size
- noisiness
- incompleteness or absence of records
- semi-random survey design (redundancy of records of one attribute but a lack of records of another)
- high heterogeneity of response variables (attributes) and large number of predicting variables.

Different departments in distribution utilities use different styles of record keeping, conventions, different time periods, levels of data aggregation, different primary keys (identifiers), and different types of errors. These are all aggravating factors when there is a need for merging different databases into one central system. Therefore data have to be assembled, integrated, and cleaned up, taking into account the importance of the right type and level of aggregation [2, 4].

C. Potential Benefits of Big Data Analytics

One of the aims of big data analytics is to reduce the size of big data. The need for information is ever growing, but the actual capacity of communication network and data processors is not. That is why the inclusion of new data sources, especially those sending streams of data in real-time, requires investments for enhancement of the communication network capacity. For instance, capital cost of the communication infrastructure needed for sending smart meter data in the UK is estimated to be around £1.15 billion [5]. In that sence, using data analytics in order to distinguish the types of data with higher importance for information retrieval might save significant expenditures predicted for upgrading servers and communication lines. Another way to make savings in investments is to analyse sampling time of data versus accuracy of extracted information, because results might show that acceptable accuracy of information can be obtained with lower sampling steps, which reduces data traffic.

Data analysis of static data stored in databases could also reveal some unexpected correlations between certain features and support actions such as asset management. An example could be high correlation between feeders of certain type and number of faults per feeder, which could support decision making regarding the investments in distribution network.

III. METHODOLOGIES OF BIG DATA ANALYTICS

Data mining techniques have already been widely used in power system security assessment, fault detection, power system control, power generation risk management and load and price forecasting [6]. The two basic outputs of data mining methods are prediction, based on observations on already existing instances with known response variables, and knowledge discovery in big databases. Number of these methods and their modifications depending on the application has been increasing constantly, but nonetheless they are all based on either correlation, regression or classification methods [6]. First part of any data mining process is pre-processing of raw data, which consists of several stages: extraction of useful data; removing data noise; statistical analysis for generating new useful variables; formatting data for the desired data mining method.

A. Correlation

Correlation is a widely used statistical tool for retrieving relationships between data. In the case of linear correlation, it gives the strength and direction (positive or negative) of relationship between random numerical variables. Also, as a means of data selection, it can be very useful for rejecting uncorrelated data, i.e. reducing data size. The typical measure of correlation is given with Pearson's coefficient r [6], calculated as follows:

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{\sqrt{[\sum_{i=1}^{n}(x_i-\bar{X})^2][\sum_{i=1}^{n}(y_i-\bar{Y})^2]}} \quad -1 \le r_{X,Y} \le 1\,, \qquad (1)$$

where $\bar{X}$ and $\bar{Y}$ are mean values of variables *X* and *Y*.

B. Regression

Linear regression is a well-known technique for numeric prediction- therefore it is often used in forecasting applications. The response (y) is presented as a linear combination of predictors $(x_1, x_2, \dots, x_n)$ and weights $(w_0, w_1, \dots, w_n)$ given in the following form [7]:

$$y = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n \qquad (2)$$

Weights are calculated based on training data, i.e. given set of examples of response values and corresponding predictors' values. These numerical weights can be used as predictors of the unknown response if the predictor attributes are known. In cases of data with nonlinear dependency, where linear regression gives only a rough estimation of the prediction

function, more accurate estimation is made using non-linear regression model given in the following form [8]:

$$y_i = f(x_i, \theta) + \varepsilon_i \,, \tag{3}$$

where y_i and x_i are vectors of response and predictor attributes in the *i*-th instance, respectively, θ is the vector of parameters, while ε_i is a random error. The parameter vector that is unknown can be estimated from training set using least squares method, i.e. minimization of the expression:

$$\sum_{i=1}^{n} (y_i - f(x_i, \theta))^2 \,. \tag{4}$$

C. Classification

Classification is a general term for all data mining methods that form groups (classes) of data based on some categorical rules [7]. It is a two-step process: first, in the training step, a model, i.e. a number of classes with defined attributes is formed based on available observations (patterns, data items or feature vectors). Second step is to classify unseen examples based on their attributes. In supervised classification methods, given set of labelled patterns (training data) is used to learn description of each class (group). Some of the classification methods are:

1) *Decision trees*

Decision trees are one of the widely used classification methods. They have been applied in power system stability studies using data from phasor measurement units (PMUs) [9]. This data mining method is based on generating comprehensive rules for dealing with both continuous and discrete data. The tree structure consists of if-then rules, i.e. tests on attributes given in nodes of the tree. Branches represent results of these tests and leaves contain class labels [10]. Class labels can be nominal, in case of classification trees, or numerical, in case of regression trees [11].

2) *Clustering*

Clustering is a common name for the group of unsupervised data mining methods, which can also be seen as a subgroup of classification methods. Classification of measurements is based on either (i) goodness-of-fit to a postulated model, or (ii) natural groupings (clusters) revealed through analysis. While classification models assign new data to previously-defined classes which are specified as a target, clustering models do not use a target. Among numerous types of clustering methods, k-means and artificial neural networks (ANN) have been most frequently used in big data analytics.

k-means is a classical clustering method [4], where initial centres of *k* clusters are randomly chosen from the data set. All other data objects (instances) are assigned to their closest cluster centre according to the ordinary Euclidean distance metric ($\|x_j - x_k\|$, in case of two vectors or patterns x_j and x_k). Next the centroid, or mean, of the instances in each cluster is calculated [12]. These centroids are taken to be new centre values for their respective clusters. The whole process is iteratively repeated with the new cluster centres, until the same points are assigned to each cluster in consecutive rounds, at which stage the cluster centres have stabilized and do not change any more. One of the applications of k-means clustering is in studies of electricity customers' daily load profiles, where profiles from the same cluster get tariffed the same way.

Artificial neural networks (ANN) present an upgrade of logistic (nonlinear) regression [4]. In power system analysis, they are mostly used for load forecasting, stability and security analysis, power system control, fault diagnosis, reactive power planning and control and for state estimation [13-15]. Disadvanatages of ANN are the empirical design of network structures and parameters and the need for numerous training instances [16]. As stated in [15], ANN are useful in cases where:

- no direct algorithmic solutions exist, but examples of predictive and response variables are available,
- problems change over time, i.e. the solution has to be adapted to the change
- only complicated algorithms can be derived.

In order to further analyse performance of data mining methods applied specifically to big data systems, comparison of the observed methods is given in Fig.1, in a form of a "radar" diagram. Good performance of the method is marked as 3 and bad performance is marked as 1, based on the comparative analysis given in Table I. Mark 2 was given if quality of performance wasn't strictly defined.

TABLE I. COMPARISON OF METHODS IN BIG DATA MINING

Method	Advantages	Disadvantages
k-means clustering	Deals with large data sets; Low computational complexity	Sensitive to initial cluster centres; Deals only with numerical data; Sensitive to outliers
ANN	Deals with large data sets; Estimates non-linear relationships; Adaptable to changes (new data) in dataset	Requires numerous samples
Decision trees	Deal with heterogenous data; Deal with multidimensional data	Sensitive to outliers; Computational complexity
Linear regression	Deals with large data sets; Simple to interpret	Discovers only linear relation between numerical variables; Sensitive to outliers

As seen from the Fig. 1, decision trees and ANN have the best performance in handling big and heterogeneous data, while k-means and linear regression handle big data sets, but only the numerical ones. K-means also showed the highest computational speed, which justifies its frequent use.

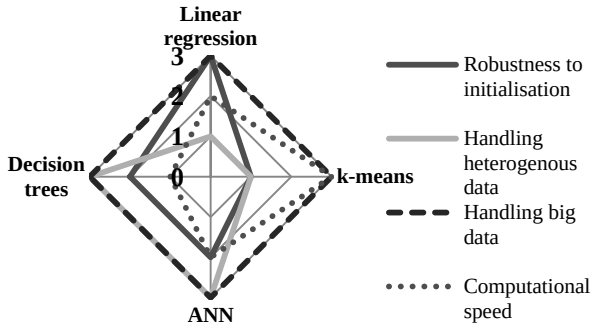

Figure 1. Performance comparison of different data mining methods

IV. CASE STUDY: DATA ANYLITICS METHODS APPLIED TO A DISTRIBUTION UTILITY'S DATABASE

As a presentation of information retrieval from databases, a real distribution utility's SQL database was given with static data about faults on feeders in HV (6.6 kV and 11 kV) network. The database consisted of numerous tables showing feeders characteristics, i.e. types of feeders, districts, exact locations of primary substations (33 kV/11 kV and 33 kV/6.6 kV), number of connected customers, etc. Also, given were exact dates and times of faults followed by the number of interruptions and cumulative duration of customer interruptions per fault. Number of customer interruptions (CI) and customer minutes lost per customer (CML), as key indicators of QoS of the distribution network, could then be calculated as follows [17]:

$$CI = \frac{number\ of\ interruptions}{total\ number\ of\ supplied\ customers} \cdot 100\ \% \qquad (6)$$

$$CML = \frac{cumulative\ interruption\ duration}{total\ number\ of\ supplied\ customers} \qquad (7)$$

SQL database data was presented over a five-year period, each instance referring to a fault causing interruption of supply longer than 3 minutes. Datasets were further analyzed using Matlab and Weka tool. The analysis presented in this paper was performed over HV (6.6 kV and 11 kV) feeders' data. Data was aggregated to feeder class level (according to classification of HV feeders' types) and district level (in order to compare network performance among geographical districts). Following the results of the analysis given in Table I and the fact that the considered database is not big enough to justify the use of ANN or k-means clustering, a linear regression and decision trees were applied in this case study. Feeder Class Analysis

All HV feeders were classified into 11 classes based on the percentage of the overhead line (OHL) part in respect to the total length (in km) of a feeder, as well as the number of customers supplied by the feeder (Table II).

TABLE II. CHARACTERISATION OF FEEDER CLASSES

	UG1A	UG1B	UG2A	UG2B	MA1	MA2	MB1	MB2	MC1	MC2	OH
% OHL	0				<20		20-50		50-80		>80
Length (km)	<4		>4		<8	>8	<11	>11	<19	>19	All
Number Of Customers	<1000	>1000	<2000	>2000	All	All	All	All	All	All	All

Fig. 2 shows some QoS indicators performance over a five-year period, together with cost of compensation for energy not supplied (ENS) to domestic users, correlated to feeder classes. Calculation of the cost of compensation was done as follows:

1) Number of CML was calculated for every fault and multiplied by number of interrupted (affected) customers;
2) Value calculated in 1) was then multiplied by the average domestic consumption, standing for 1.1 kW [18], giving the energy not supplied (ENS) per fault. This value was summarised for all the faults happening on each of the feeder classes;
3) The cumulative ENS was multiplied by the value of lost load (VoLL) for domestic sector (16.94 £/kWh [19]), giving the compensation cost for each feeder class during the given period.

All calculated values were normalized to maximum calculated values given in Table III.

TABLE III. BASE VALUES FOR NORMALIZATION IN FIG. 2

Measures	Base values
Average CML per feeder	18,860 min
Cumulative ENS/Cost of compensation	720 MWh/£ 12.2 million
Domestic share	100 %
Average number of interruptions per fault	1,445
Number of faults per feeder class	2297

As it can be seen from the radar diagram in Fig. 2, the highest cumulative amount of ENS was calculated across HV feeders of class MC2 – feeders with high share of domestic users and also a high share of overhead lines. Following this, MC2 showed the highest compensation cost rate for domestic users during the period (around 1.8 % of the network DSO's five-year profit). Feeder class UG1A (underground cable) showed the highest rate in average number of CML per feeder, probably due to the reduced accessibility for fault removal. This is also the most common type of feeder used in the observed HV distribution network, with a contribution of

around 27 %. Feeder class UG2B showed the highest number of interruptions per fault, which is justified by the fact that this feeder class supplies more than 2,000 customers on average.

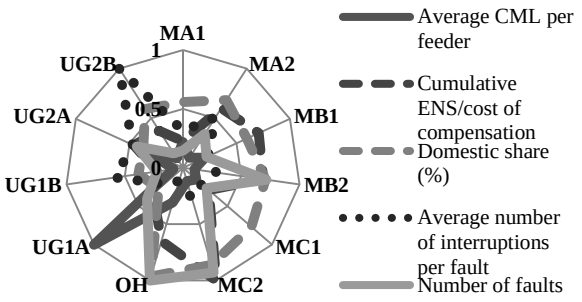

Figure 2. QoS performance with normalized values per feeder class

QoS indicators were then correlated to feeder characteristics (number of customers supplied by the feeder, length (km) of the overhead-line part and underground part and total length (km) of the feeder). Linear regression analysis showed that correlation coefficients for the same indicators were higher for higher aggregation level. When accumulated to primary substation level, QoS indicators showed high correlation with feeder parameters, mainly with total length of the feeders and number of customers supplied from the substation. Correlation coefficients for no aggregation and different levels of aggregation are given in Table IV.

TABLE IV. CORRELATION COEFFICIENTS

QoS indicator	Correlation coefficient (r) to feeder characteristics		
	No aggregation	Aggregation per feeder level	Aggregation per primary substation level
Number of faults	/	0.78	0.84
Cumulative number of interruptions during faults	0.36	0.60	0.74
Cumulative duration of interruptions (in minutes) during faults	0.26	0.60	0.72

In order to justify this, a presentation of learning based on data is given with a regression tree, i.e. an M5 pruned (M5P) model tree chosen in Weka tool, where number of faults accumulated per primary substation is assessed based on underground (UG), over-head line (OHL) and total (Tot) length of HV feeders, and number of connected customers (Num Cust). The first number in each bracket in Fig. 3. presents the number of instances reaching the leaf and the second number presents the percentage of misclassified instances. Correlation between feeder characteristics and number of faults is quite high in this case, too (r=0.86). Since the classifier showed low correlation with the number of customers, this attribute did not participate in the tree formation.

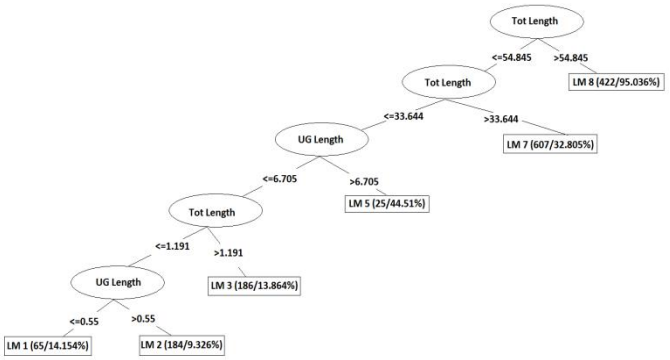

Figure 3. Regression tree result

B. *District Analysis*

QoS was analysed in seven geographical districts, all operated by the same DSO, in order to investigate possible connection between network perfomance and some specific characteristics of individual districts. Shares of domestic and non-domestic users per district (upper part of Fig. 3) were compared to estimated ENS in the domestic sector for all the observed districts during the five-year period (lower part of Fig. 3).

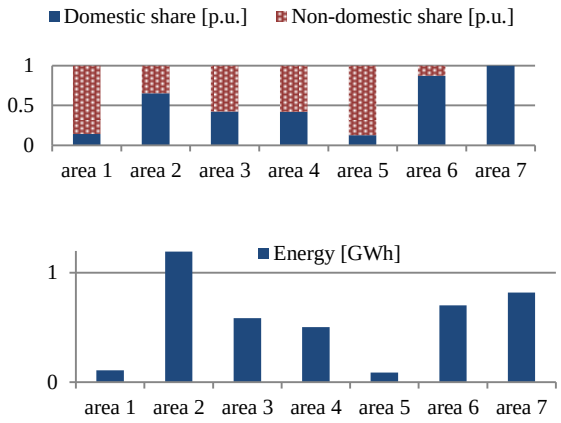

Figure 4. District analysis of estimated revenues for energy not supplied in the domestic sector

The highest estimated expenditures for the compensation for the ENS to the domestic customers were made in area 2 by supply interruptions and calculated to be around 3 % of the DSO's five-year profit. This area, with domestic share of more than 60 %, also showed the highest rate of ENS in domestic district, with around one million kWh of energy not supplied. Network in this area mainly consists of MA2 and MB2 feeders, which belong to medium length feeders with less than 50 % of overhead line part and with ageing deterioration as the main

cause of fault occurence. Distribution network in areas 3, 6 and 7, with total compensation costs almost two times higher than those in area 2 alone, consists of OH and MC2 feeders, which showed the highest fault tendency, especially since weather (wind and gale) was stated as one of the main causes of faults in these areas.

When the ENS values over the most critical districts (2, 3, 6 and 7) get disaggregated down to a year level, as in Fig. 4, it can be seen that excessive ENS in district 2 acumulated in only one year. Therefore further analysis should be made to investigate possible reasons for this.

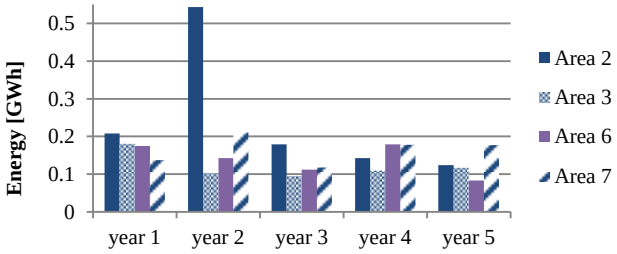

Figure 5. Amount of energy not supplied during five-year period in some districts

V. CONCLUSION

This paper discussed some of the most commonly used data mining methods in power system studies. Among the observed data mining methods, ANN, decision trees and k-means have until now showed good performance in big data applications. The database considered in this paper is not particularly large, therefore only some of the described methods were applied, namely linear regression and decision trees.

Results of the corelation analysis presented in this paper have shown that QoS parameters very much depend on feeder characteristics. Therefore, prediction methods, such as decision trees or linear regression, could be used to form a model for QoS indicators estimation based on some asset characteristics. At this point, this kind of a model would show significant errors due to a relatively small number of instances in the training data set. That means that prediction models bring more usefulness in case of effectively larger sample size, in this case larger number of feeders observed.

ACKNOWLEDGMENT

This research is partly supported by the EU Horizon 2020 project “Nobel Grid”, contract number 646184.

REFERENCES

[1] C. G. Zhao, C.; Li, F., "Classification of Low Voltage Distribution Networks Based on Fixed Data," presented at the 23rd Inernational Conference on ELectricity Distribution, Lyon, 2015.

[2] R. J. Broderick and J. R. Williams, "Clustering methodology for classifying distribution feeders," in *Photovoltaic Specialists Conference (PVSC), 2013 IEEE 39th*, 2013, pp. 1706-1710.

[3] M. Sforna, "Data mining in a power company customer database," Electric Power Systems Research, vol. 55, pp. 201-209, 2000.

[4] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition: Elsevier Science, 2005.

[5] K. Samarakoon, J. Ekanayake, and N. Jenkins, "Reporting Available Demand Response," Smart Grid, IEEE Transactions on, vol. 4, pp. 1842-1851, 2013.

[6] Z. Dong and P. Zhang, Emerging techniques in power system analysis: Springer, 2010.

[7] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques: Elsevier, 2011.

[8] G. K. Smyth, "Nonlinear regression," Encyclopedia of environmetrics, 2002.

[9] T. Guo, P. Papadopoulos, P. Mohammed, and J. V. Milanovic, "Comparison of ensemble decision tree methods for on-line identification of power system dynamic signature considering availability of PMU measurements," in PowerTech, 2015 IEEE Eindhoven, 2015, pp. 1-6.

[10] H. Zhao, "Decision tree technology in data classification," in Applied Mechanics and Materials vol. 268, ed, 2013, pp. 1752-1757.

[11] "Matlab I., The MathWorks, Statistics and Machine Learning Toolbox User's Guide R2015b " 2015.

[12] A. J. Urquhart and M. Thomson, "Impacts of Demand Data Time Resolution on Estimates of Distribution System Energy Losses," Power Systems, IEEE Transactions on, vol. 30, pp. 1483-1491, 2015.

[13] D. C. Park, M. A. El-Sharkawi, R. J. Marks, II, L. E. Atlas, and M. J. Damborg, "Electric load forecasting using an artificial neural network," Power Systems, IEEE Transactions on, vol. 6, pp. 442-449, 1991.

[14] I. P. Panapakidis, G. K. Papagiannis, and G. C. Christoforidis, "Bus load forecasting via a combination of machine learning algorithms," in Power Engineering Conference (UPEC), 2014 49th International Universities, 2014, pp. 1-6.

[15] S. Kamalasadan, "Application of Artificial Intelligence Techniques in Power Systems," Asian Institute of Technology, Bangkok1998.

[16] T. Wang, G. Zhang, J. Zhao, Z. He, J. Wang, and M. J. Pérez-Jiménez, "Fault Diagnosis of Electric Power Systems Based on Fuzzy Reasoning Spiking Neural P Systems," 2014.

[17] "Quality of Supply and Market Regulation; Survey within Europe," KEMA Consulting2006.

[18] "VirginiaTech Research Data, [Online].Available: http://www.ari.vt.edu/research-data/."

[19] "The Value of Lost Load (VoLL) for Electricity in Great Britain, final report for OFGEM and DECC," London Economics2013.