2006

Identification of Load Power Quality Characteristics using Data Mining

Ali Asheibi
University of Wollongong, ali_asheibi@uow.edu.au

David A. Stirling
University of Wollongong, stirling@uow.edu.au

Duane Robinson
University of Wollongong

Publication Details

This paper was originally published as: Asheibi, A, Stirling, D & Robinson, D, Identification of Load Power Quality Characteristics using Data Mining, Canadian Conference on Electrical and Computer Engineering, Ottawa, Canada, 7-10 May 2006, 157-162. Copyright IEEE 2006.

Identification of Load Power Quality Characteristics using Data Mining

Abstract

The rapid increase in computer technology and the availability of large scale power quality monitoring data should now motivate distribution network service providers to attempt to extract information that may otherwise remain hidden within the recorded data. Such information may be critical for identification and diagnoses of power quality disturbance problems, prediction of system abnormalities or failure, and alarming of critical system situations. Data mining tools are an obvious candidate for assisting in such analysis of large scale power quality monitoring data. This paper describes a method of applying unsupervised and supervised learning strategies of data mining in power quality data analysis. Firstly underlying classes in harmonic data from medium and low voltage (MV/LV) distribution systems were identified using clustering. Secondly the link analysis is used to merge the obtained clusters into supergroups. The characteristics of these super-groups are discovered using various algorithms for classification techniques. Finally the a priori algorithm of association rules is used to find the correlation between the harmonic currents and voltages at different sites (substation, residential, commercial and industrial) for the interconnected supergroups.

Disciplines

Physical Sciences and Mathematics

Publication Details

This paper was originally published as: Asheibi, A, Stirling, D & Robinson, D, Identification of Load Power Quality Characteristics using Data Mining, Canadian Conference on Electrical and Computer Engineering, Ottawa, Canada, 7-10 May 2006, 157-162. Copyright IEEE 2006.

IDENTIFICATION OF LOAD POWER QUALITY CHARACTERISTICS USING DATA MINING

Ali Asheibi, David Stirling, Duane Robinson
Integral Energy Power Quality and Reliability Centre
School of Electrical, Computer and Telecommunications Engineering
University of Wollongong
email: atma64@uow.edu.au

Abstract

The rapid increase in computer technology and the availability of large scale power quality monitoring data should now motivate distribution network service providers to attempt to extract information that may otherwise remain hidden within the recorded data. Such information may be critical for identification and diagnoses of power quality disturbance problems, prediction of system abnormalities or failure, and alarming of critical system situations. Data mining tools are an obvious candidate for assisting in such analysis of large scale power quality monitoring data.

This paper describes a method of applying unsupervised and supervised learning strategies of data mining in power quality data analysis. Firstly underlying classes in harmonic data from medium and low voltage (MV/LV) distribution systems were identified using clustering. Secondly the link analysis is used to merge the obtained clusters into super-groups. The characteristics of these super-groups are discovered using various algorithms for classification techniques. Finally the a priori algorithm of association rules is used to find the correlation between the harmonic currents and voltages at different sites (substation, residential, commercial and industrial) for the interconnected super-groups.

Keywords: *power quality, harmonics, data mining.*

1. Introduction

Due to the high level of uncertainty with regards to power system networks, along with multidimensional parameters such as voltage, current and impedance, utility engineers are now beginning to rely on the classification tools of data mining techniques to support decisions of assessing the security of operation of power systems [1]. Data mining methods may be used in load forecasting to build predicting models and to discover relationships between input and output variables such as weather parameters, seasonality, and load profiles [2]. Decision trees, the elementary tools for defining rules for pattern recognition in data mining, can be used to discover new unseen rules for short and long term load forecasting and the probable demand surplus/deficit arising from unusual weather can be predicted from these rules [3].

Essential in applying data mining tools to power quality data is the ability to identify the various underlying classes associated with the sites monitored and power quality disturbances of interest. Two important learning strategies exist in data mining and machine learning techniques for such analysis: supervised learning (SL) and unsupervised learning (USL).

In supervised learning, target or output classes are identified and relabelled in the data prior to the application of a learning algorithm. Unsupervised learning on the other hand has no output classes. USL techniques are typically successful in situations where there are many variables or attributes describing a large volume of data. Power quality data is an example of such a domain in that it has many electrical parameters (voltage, current, impedance, etc.) in three phases and includes many power quality disturbances (harmonics, unbalance, sags, etc.). To report these power quality disturbances, a huge amount of data (10 GB/week/site) is required to be attained from many distributed sites. The USL techniques of data mining (e.g. clustering and link analysis) can be applied in power quality as a first step to find classes in the data and merge the similar classes into super-groups. The SL techniques (e.g. classification and association) are then used to extract rules behind the formed super-groups [4].

This paper first introduces the data mining tools utilised for the analysis, then discusses the data set based on harmonics data from a power quality survey of a study MV/LV distribution system, and finally presents the results from the applied data mining analysis.

2. Unsupervised clustering with Gaussian mixture model

Clustering is an important technique in data mining, machine learning and communication systems. It is a powerful approach that can discover underlying and meaningful groups of data. In part, clustering can be considered as a learning process. It is also a useful tool for analysis of complex data sets, such as for lossy image compression in communication systems [5]. Clustering also divides or segments an initial collection of data into a certain number of groups or clusters. As a result the data residing in each cluster are similar, whereas data across different clusters are dissimilar. Most of the criteria that are used to measure similarities or dissimilarities are based on geometric distance (for example, K-means algorithms) or probability density function estimation, such as Gaussian mixture model (GMM) [6]. Any random variable in practice can be represented by Gaussian or normal distributions as long as this random variable is unaffected or affected by independent variables almost identically [7].

The data mining software used thus far to accomplish automatic clustering of the power quality (PQ) data for this study is ACPro, which assumes that the distribution of data under study was generated from a sum of simpler distributions [8]. As an extension to this analysis finite mixture models using Bayesian clustering are used to produce a range of models. Subsequently a minimum message length (MML) encoding algorithm is used to identify the best model [9]. This MML based algorithm is similar in nature to other intrinsic modelling tools such as Autoclass and Snob [10].

3. Similarities and dissimilarities between discovered clusters

Each cluster developed by the model in the analysis was further investigated with respect to all other clusters in terms of similarities or dissimilarities. The number of probability density functions in each cluster equals the number of variables or features of the cluster. The features for every pair of clusters were compared by calculating the Kullback-Lieber (KL) distance between each pair [11]. The smaller the distance between each pair, the more similar these clusters are. This means that this pair has a similar amount of information so that they can be linked together to form one group. This link analysis process is continued until each single cluster is linked to one of the formed super-groups based on the KL distance. The link analysis techniques used to form the network from KL distances are explained in the next section.

4. Super-group formation using MDS

As mentioned above, similar clusters were merged into super-groups. A multidimensional scaling algorithm (MDS) [12], which is a dimension reduction technique that can reduce the high dimensional data into smaller dimensions down to one dimension, was used to form a network from KL distances. Using Knowledge Network Organising Tools (KNOT), which is essentially an MDS algorithm, the distances between clusters that are found by calculating KL distances are confirmed and the super-group abstractions are formed by removing the links exceeding a dissimilarity threshold between two different super-groups.

5. Study distribution system

To illustrate the use of the data mining analysis tools PQ monitoring results from three MV electricity utility customers on a typical MV distribution system were obtained. Data from the source end of the MV feeders supplying the customers and the HV/MV substation transformer supplying the distribution network was also obtained, as shown in Figure 1. Although not selected specifically for the application of data mining the test system involved capturing PQ data using standard parameters and monitoring intervals and thus it was perceived the true applicability of data mining to PQ data would be illustrated. The monitored data included voltage and current readings of the fundamental, THD, and 3rd, 5th, and 7th harmonics every 10 minutes over a period of two weeks. Measurements were taken from the HV/MV zone substation transformer voltage transformers (VT) and current transformers (CT) at the MV feeder CTs and the LV side of each customer's 11kV/430V distribution transformer. The selected customers represented different load types, i.e. primarily residential, commercial or industrial sites. The locations of PQ monitoring devices at Sites 1-7 are illustrated in Figure 1.

The residential site consists primarily of residential homes in an inner suburban location. The commercial site is a large shopping centre operating seven days a week. The industrial site is a medium sized factory manufacturing paper products such as paper towelling. Data from the sites monitored were pre-processed in text form before being entered into the ACPro software for analysis.

6. Daily harmonics variations

Harmonics are a continuous distortion to the voltage or current waveforms. Non-linear loads, either single phase or three phase, are the main source of harmonics which may lead to a number of problems at both utility and customer sites. Such problems include power factor correction capacitor failure, overheating of neutral conductors and false tripping of electrical distribution equipment [13]. Losses due to harmonics in industrial systems can increase operational costs and decrease the useful life of system equipment.

System response of distribution networks can enhance or attenuate the effects of harmonics on the power system and is governed by such parameters as system impedance, the presence of capacitor banks, and the amount of passive load connected to the system. Cyclic variation of the non-linear and

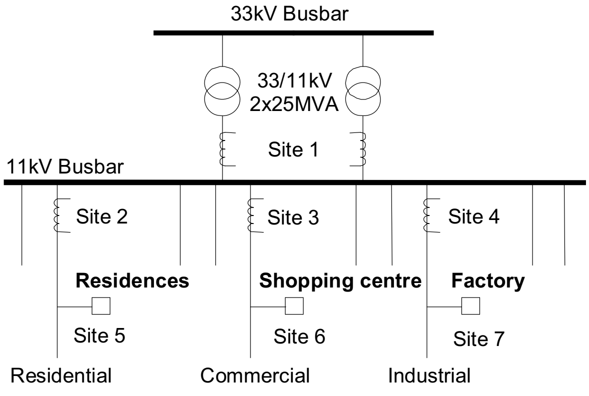

Figure 1. Schematic layout of test system.

passive load and capacitor switching are the main reasons for fluctuating individual harmonic voltage levels over each day.

There are many individual harmonic components that can exist in power systems, the most prominent is the 5th harmonic based on recordings over the previous decade, as its daily pattern contributes the most to Total Harmonic Distortion (THD) [14]. The next major individual harmonics are the 3rd and 7th. Each individual harmonic has its own daily pattern which is typically repeated over the week (weekdays & weekend), however the values of harmonic components are not necessarily the same during specific times of different weeks. Furthermore, different site types (i.e. residential, commercial, and industrial) produce different patterns for the same individual harmonic. Nevertheless, individual harmonic daily behaviour can be classified into different classes, e.g. the 5th harmonic voltage shown in Figure 2 using data from [15] increases during the early evening, stays at the highest level, and then late in the evening decreases rapidly and remains at a predominantly low level for the rest of the day, thus suggesting three nominal classes.

7. Classification and association rules in harmonics

The decision tree (C5.0 algorithm) for classification techniques is used to define the rules behind each super-group that describe the level of harmonics and consecutively the a priori algorithm of association rules is used to explore which site (residential, commercial or industrial) is causing what level of harmonics elsewhere for each rule. Data mining software called Clementine is used in this section for both classification and association algorithms.

As mentioned in Section 3, the automated segmentation generalized by ACPro was primarily used in this work. A segmentation (or clustering) using USL techniques was used to discover similar groups of records in the database, which in this case included clustering the harmonic data from the test system. The number of clusters obtained was automatically determined based on the significance and confidence placed in the measurements, which can be estimated using the entire set of measured data. Data from different sites (site 1, 5, 6, 7) were used as input data to the software and 11 clusters (s0, s1, s2, .., s10) were produced with different abundance, mean and standard deviation as shown in Figure 3.

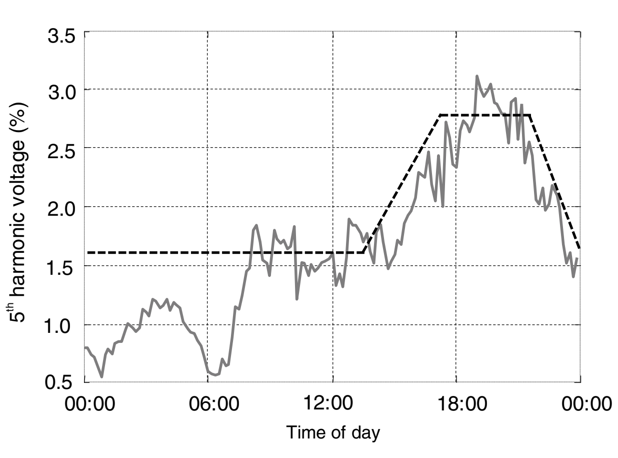

Figure 2. 5th harmonic voltage at commercial site

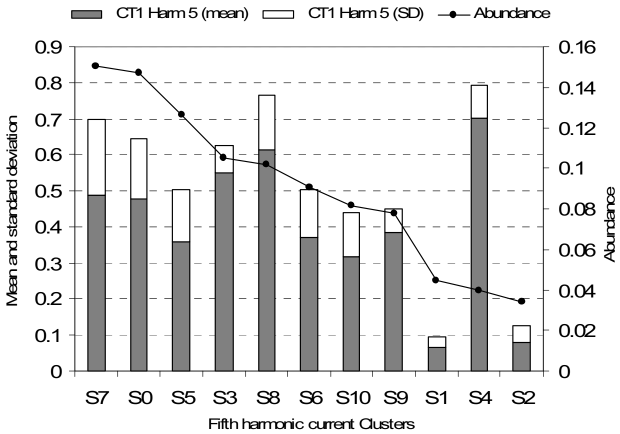

Figure 3. Abundance, mean and standard deviation for each cluster of 5th harmonic current per phase

The KL distance between the clusters was measured by the software and the lower triangular 10x10 matrix of KL distances from the 11 clusters as shown in Table 1. These distances were sorted to find the most similar clusters, such as clusters (s5, s10) with KL=34 and (s7, s10) with KL=36, and the most different cluster such as (s1, s3) with KL=3186 and (s1, s0) with KL=2674.

The links between these clusters were visualized using KNOT by reducing the 11 dimensional model into a two dimensional graph and then forming the super-group when removing the links that exceeded the dissimilarities threshold. The obtained super-groups (A, B, C, D and E) are shown in Figure 4. Most of the super-group abstractions are formed based on the site type (industrial, commercial, and substation) excluding super-group B, which is formed from clusters containing data from all sites. This can be attributed to the substantial levels of harmonics, as discussed at the end of this section. The residential site does not seem to have a particular

Table 1. Kullback-Lieber distances between clusters

S0										
S1	**2674**									
S2	832	232								
S3	62	**3186**	**1157**							
S4	181	2486	941	178						
S5	59	1077	358	185	127					
S6	51	1277	361	173	169	**37**				
S7	51	2518	871	107	155	58	142			
S8	102	2773	1003	113	169	145	201	39		
S9	450	1486	612	519	649	194	234	471	365	
S10	115	867	332	233	153	**34**	107	**36**	70	116
	S0	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**

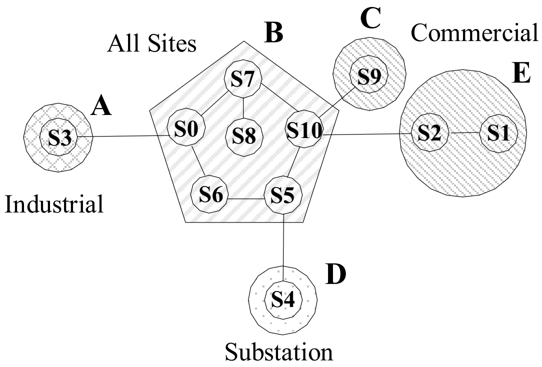

Figure 4. Super-group abstractions by m MDS.

super-group, which means that the influence of harmonic emissions (or participation) from this site is very low. These super-groups are plotted over one week and overlaid by the sites (substation, residential, commercial and industrial), as in Figure 5, to show the synchronisation of more than one super-group at different sites.

As can be seen from Figure 5, super-group A at the industrial site is synchronised with both super-group D at the substation site and super-group E at the commercial site early in the morning, each day of the week. This means that there is a harmonic link between these sites at that time. To discover the characteristics of these super-groups, the decision tree (C5.0 algorithm) of classification techniques in Clementine is used and the rules are generated. It can be stated that the level of harmonics at each site provides the second criterion for forming these super-groups, after the site type in Figure 4.

The average level of harmonic voltages is normalised to 1.0 per unit, which is also the maximum value for harmonic currents.

Figure 6 illustrates the high level of harmonic voltage (Rule B1) occurs at the sites simultaneously most of the time. This means that there is an interaction between these sites, i.e. one or more sites causing high 5th harmonic voltage to occur at the other sites. The suggested one is the commercial site where Rule B-1 is taking place more frequently. This may be due to the existence of a large number of fluorescent lamps. The industrial site also is included in B1 but tends to occur as a reduced proportion. Likewise average to high levels of 3rd harmonic voltage (Rule B4) can be attributed to the industrial site. This is most likely due to small AC motors that are used at the industrial site. Another observation is the coincidence of Rule A1 at the industrial site with D1 at the substation site late at night and early in the morning, which means that the possibility of high harmonic levels still exists even when load current is low.

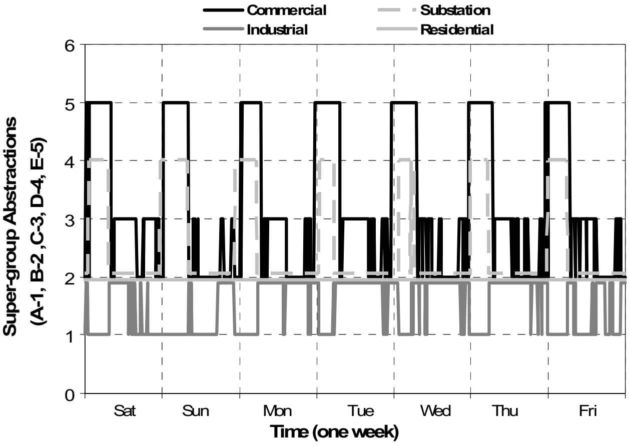

Figure 5. Super-groups in all sites over one week.

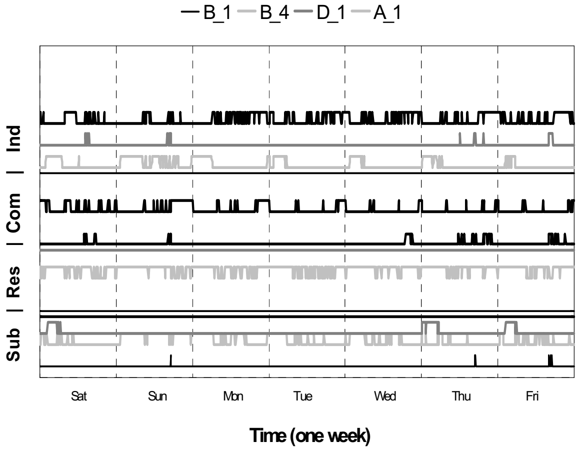

Figure 6. Rules of high level harmonics at all sites over one week.

The next step involved using an a priori algorithm of association rules in Clementine to categorize the correlation between the variables at different sites for the interrelated super-groups. The association rules that were uncovered are many, of which one example is selected, which indicates that the 5th harmonic current at the industrial site is responsible for the 5th harmonic voltage existing at all sites. To further analyse the harmonic current and voltage relationship identified by the super-groups, correlation coefficients were determined for the data set within the super-group against the 5th harmonic current emissions from the industrial site. The resulting correlation coefficients were classified as follows; weak (0-0.33), medium (0.33-0.66) and strong (0.66-1.0). Figure 7 shows the correlation between 5th harmonic voltage at the substation site and the 5th harmonic current at the industrial site in phase A (86.3%). However, this correlation is lower in phase C (43%) as shown in Figure 8, which indicates a significant level of unbalance between the phases.

Although the example provided in this study identifying the source of the most significant harmonic emissions could also be completed through manual analysis of data, the method utilising the data mining tools suggested allows many such patterns to be identified simultaneously. In this manner the outlying events of the power quality data set, e.g. the more

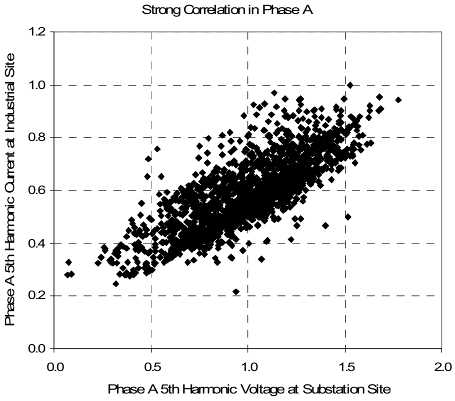

Figure 7. Strong correlation in Phase A between industrial site and substation site.

significant harmonic voltages and currents, can easily be identified for further analysis, minimising the time required for analysis by the distribution engineer.

8. Simulation Results

To study the harmonic interaction between the sites, an investigation was carried out by modelling and simulating the distribution system under study in time domain using the PSCAD^{TM}/EMTDC® electromagnetic transient software program. The available data was used to calculate the variables needed for the simulation. The passive load harmonics models were represented as a resistor (e.g. incandescent lights, ovens, heaters, etc.) in parallel with a series impedance of reactor and resistor (dynamic impedance of an induction motor), as shown in Figure 9. To differentiate between different load types (residential, commercial or industrial sites), different values of power factor (Pf) and an allocation factor (Af) were assigned for each load type depending on phase shift between voltage and current and the percentage of resistive loads to the inductive loads for each site respectively. The load factor (Lf) which is the percentage of full load was also considered for each load type. The values of Pf, Af and Lf selected for the simulation are shown in Table 2.

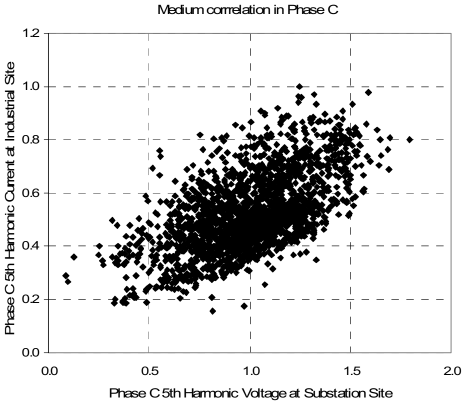

Figure 8. Medium correlation in Phase C between industrial site and substation site.

Table 2. Practical values of power factor (Pf), allocation factor (Af) and load factor (Lf)

Site	Residential	Commercial	Industrial
Pf	0.9	0.85	0.8
Af (%)	85	75	50
Lf (%)	35	40	85

Field data from the study system was used to determine the magnitude of harmonic current injections at sites 5, 6 and 7 (see Figure 1). To represent the harmonic contribution at medium voltage (MV) from other loads on the system, 5th harmonic currents were also injected at a magnitude determined by the monitored values at sites 1, 2, 3 and 4, combined using the second summation law from [16]. This attempts to account for all the other 11kV feeders that are connected to the same substation.

The aim of the simulation was to verify the contributions from each customer site to the overall harmonic voltage levels existing on the system identified using the data mining techniques. Figure 10 illustrates the harmonic voltage at the zone substation 11kV busbar (site 1) alongside the harmonic current contribution from the industrial customer (site 7). The significant contribution to harmonic voltage levels by the industrial site identified in Figures 5 and 7 using the data

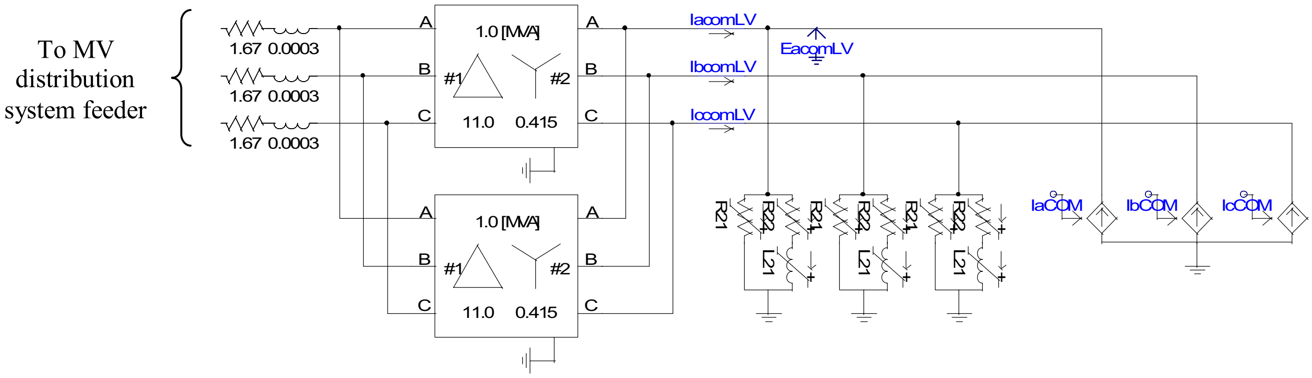

Figure 9. PSCAD Simulation Schematic of LV Customer Installation

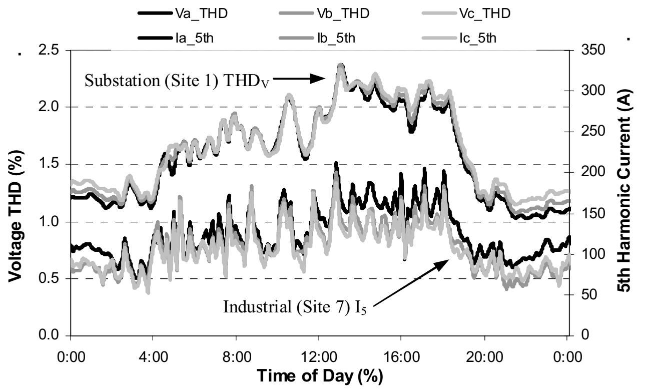

Figure 10. Simulated Output of Zone Substation 11kV busbar voltage THD and Industrial Site 5th harmonic current.

mining techniques is repeated in the PSCAD simulation output.

Although in preliminary stages only it is anticipated that the PSCAD simulation exercise will be further utilised to test the application of data mining techniques as the entire parameter set of the system is known, as opposed to the physical system where extensive monitoring is required to fully understand the system operation.

9. Conclusion

Data mining, both unsupervised learning and supervised learning, has been shown to be able to identify useful patterns within the power quality data set. Significant results obtained from cluster analysis, classification and association rules to illustrate the applicability of data mining in power quality data have been developed. Link analysis and visualisation techniques are also data mining tools that can assist in discovering useful patterns and relationships that are present in power quality data sets. The super-groups formation from different sites has been investigated and the effect of these sites on each other has been examined. The causality of unwanted distortion was also discovered using the classification rules.

Monitored data from a study MV/LV distribution system has been utilised in the development of the analysis tools. The collected data has also been reinforced with the development of an example system in simulation.

References

[1] L. A. Wehenkel, *AUTOMATIC LEARNING TECHNIQUES IN POWER SYSTEMS*. Boston: Kluwer Academic, 1998.

[2] S. Rahman and R. Bhatnagar, “An expert system based algorithm for short term load forecast”, *IEEE Trans. on Power Systems*, Vol. 3, No. 2, pp. 392-398, May 1988.

[3] B.D. Pitt, “Application of data mining techniques to electric load profiling”, *PhD Thesis*, Manchester Institute of Science and Technology, pp. 197, 2000.

[4] J. Han, *DATA MINING: CONCEPTS AND TECHNIQUES*. San Francisco: Morgan Kaufmann, 2001.

[5] D.J.C. MacKay, *INFORMATION THEORY, INFERENCE, AND LEARNING ALGORITHMS*. New York: Cambridge University Press, 2003.

[6] R.O. Duda, *PATTERN CLASSIFICATION*. New York: Wiley, 2001.

[7] A.D. Aczel, *COMPLETE BUSINESS STATISTICS*. Boston: McGraw-Hill/Irwin, 2002.

[8] A. Asheibi, D. Stirling, S. Perera, and D. Robinson, “Power quality data analysis using unsupervised data mining”, *AUPEC*, Brisbane, September 2004.

[9] R.A. Baxter and J. Oliver, “Finding overlapping components with MML”, *Statistics and Computing*, Vol. 10, pp. 5-16, 2000.

[10] C. Wallace and D. Dowe, “Intrinsic classification by MML - the snob program”, *Proc. 7th Aust. Joint conf. on Artificial Intelligence*, Armidale, Australia, (1994).

[11] K. Solomon, *INFORMATION THEORY AND STATISTICS*. New York: Dover Publications, Inc., 1997.

[12] W. Schvaneveldt, *PATHFINDER ASSOCIATIVE NETWORKS*. New Jersey: Alpex. 1990.

[13] M. H. Shwehdi, *et al.*, “Harmonic flow study and evaluation of a petrochemical plant in Saudi Arabia”, *LESCOPE*, pp. 165-172, June 2002.

[14] E. Duggan and R.E. Morrison, “Prediction of harmonic voltage distortion when a nonlinear load is connected to an already distorted supply”, *IEE Proc. Generation, Transmission and Distribution*, Vol. 140, No. 3, pp. 161-166, May 1993.

[15] D. Robinson, “Harmonic Management of MV Distribution Systems”, *PhD Thesis*, University of Wollongong, 2003.

[16] IEC 61000-3-6, “*Electromagnetic compatibility (EMC) - Part 3: Limits - Section 6: Assessment of emission limits for distorting loads in MV and HV power systems - Basic EMC publication*”, 1996.