

POWER QUALITY DATA ANALYSIS USING UNSUPERVISED DATA MINING

Ali Asheibi, David Stirling, Sarath Perera, Duane Robinson

Integral Energy Power Quality Centre
School of Electrical, Computer and Telecommunications Engineering
University of Wollongong

Abstract

The rapid increase in the size of databases required to store power quality monitoring data has demanded new techniques for analysing and understanding the data. One suggested technique to assist in analysis is data mining. Data mining is a process that uses a variety of data analysis tools to identify hidden patterns and relationships within large samples of data. This paper presents several data mining tools and techniques that are applicable to power quality data analysis to enable efficient reporting of disturbance indices and identify network problems through pattern recognition. This paper also presents results of data mining techniques applied to power quality data from an MV electrical distribution system to identify disturbance patterns.

1. INTRODUCTION

Power Quality (PQ) monitoring is an important issue for electricity utility customers due to the increasing penetration of equipment susceptible to power quality disturbances. With this type of equipment PQ disturbances can cause significant financial impact due to loss of production, damage to equipment, and disruption on related manufacturing processes [1]. For these reasons, large industrial and commercial customers are becoming proactive with regards to PQ monitoring. Pressure from electricity utility regulators, and the concept of electricity being sold as a product with measurable quality, requires that utilities also carry out extensive PQ monitoring programmes to ensure disturbance levels remain within predetermined limits [2]. For both utility and customer extensive PQ monitoring will eventually involve the storage and analysis of large amounts of data.

Routine preliminary analysis of PQ data typically involves the calculation of indices as specified in the relevant standards. This analysis typically includes reporting classical statistical parameters such as average, minimum and maximum [3]. In addition, histograms and cumulative frequency curves can provide useful descriptions of measured data. Moreover, cumulative probability values (95th or 99th percentiles) and probability distribution functions are recommended to represent continuous disturbances [2]. Although researchers have realised such large amounts of PQ data also hold much more information than that reported using classical statistical techniques for PQ monitoring [4], few have taken the opportunity to exploit this additional information. Such information could include

recognition of disturbance level pattern's prior to significant power quality events, relating plant or system events to disturbances, and identifying growth trends of disturbance levels. It is proposed data mining will provide an avenue to extract this information from PQ databases.

Data mining is a process that uses a variety of data analysis tools to identify hidden patterns and relationships within data. These tools are a mixture of machine learning, statistics and database utilities. Data mining has recently obtained popularity within many research fields over classical techniques for the purpose of analysing data due to (i) a vast increase in the size and number of databases, (ii) the decrease in storage device costs, (iii) an ability to handle data which contains distortion (noise, missing values, etc.), (iv) continuous progress in the implementation of automatic learning techniques, and (v) the rapid increase in computer technology [5]. The ultimate goal of data mining is to discover useful information from large amounts of data in many different ways using rules, patterns, and classification [6]. Data mining can be used to identify anomalies that occur as a result of network or load operation, which may not be acknowledged by standard reporting techniques.

2. DATA MINING TECHNIQUES

It is proposed that data mining can provide answers to the end-users about PQ problems by converting raw data into useful knowledge. This process can be completed using the following steps in an iterative manner (a) knowledge definition (by end user), (b) data selection, (c) data transformation, (e) data mining and extraction, (f) information assimilation,

and (g) report presentations [7, 8]. The data mining process differs from classical statistical methods in that solutions from statistical methods focus only on model estimation, while data mining techniques focus on both model formation and its performance. Another significant difference is that statistical methods fail to analyse data with missing values, or data that contains a mixture of numeric and qualitative forms. Data mining techniques, instead, can analyse and cope intelligently with records containing missing values, as well as a mixture of qualitative and quantitative data, without tedious manual manipulation [9, 10].

2.1 Unsupervised Learning (USL)

For machine learning, the process of inductive learning essentially abstracts, generalizes or compresses observations into a model. There are two important learning strategies in machine learning and data mining techniques: Supervised Learning (SL) and Unsupervised Learning (USL). Supervised learning, or data classification, provides a mapping from attributes to specified classes or concept-groupings (i.e. classes are identified and pre-labelled in the data prior to learning). Unsupervised learning generally amounts to discovering a number of patterns, subsets, or segments within the data, without any prior knowledge of the target classes or concepts, that is, learning without any supervision. Since there are no predefined classes within the available PQ data, USL is used to identify statistically valid classes within the data itself. Generally one can describe the goal of USL as the discovery of structural patterns inside a set of, often multi-dimensional, data. Not all such discoveries are ultimately interesting or overly useful, thus it is the user's duty to evaluate the potential value of these, especially any new, anomalous or unexpected patterns.

Clustering is an important, if not core, technique in data mining, especially for USL [11]. In clustering, data is assembled subject to some measure of similarity into groups of like records. This similarity is mostly based on some geometric distance or a probability density model. Clustering as a data mining tool can also operate in two different ways. It can be used as an individual tool to identify, analyse and model the structures within a data set. It can also be used as a first step for further data mining processes, such as SL [12]. Typically, subsequent SL may facilitate additional insight and understanding, as well as an expression of certain patterns in terms of various specific attributes or features contained within the data. The primary focus of the remaining sections will be on USL data techniques in an attempt to identify patterns of

power quality disturbance behaviour in a typical MV radial distribution system.

2.2 Unsupervised Clustering with MML

Unsupervised clustering is based on a premise that there are several underlying classes that are hidden or embodied within a data set. The objective of these processes is to identify an optimal model representation of these intrinsic classes, by separating the data into multiple subgroups or clusters. The selection of data into candidate subgroups is typically subject to some form of objective function such as a probabilistic model distribution. For any arbitrary data set several possible models or segmentations might exist, each with a plausible assortment of formulated clusters. An evaluation procedure, such as Minimum Message Length (MML) encoding, is used on each in order to identify the best model.

This methodology is also well known as mixture modelling or intrinsic classification [13, 14]. Mixture models typically perform better than those based on a priori distance measures, such as a nearest neighbour algorithm, for example k-means [15]. The data mining software used thus far for automatic clustering of the PQ database is ACPro, an MML based algorithm similar to other intrinsic modelling tools such as Autoclass [16] and Snob [14]. Essentially it includes a comprehensive Bayesian clustering scheme that uses the finite mixture model, with prior distribution on all parameter values. Overall the software allows selection of the number of clusters and data precision used, and produces models structured as a collection of the means, variances and relative abundance of each of the constituent clusters. Besides forming cluster models on multivariate data, such tools can also generate predictions using the derived models [17, 18].

3. TEST 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 was also monitored at the source end of the MV feeders supplying the customers and the HV/MV substation transformer supplying the distribution network, 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 the 5th, 19th, and 49th harmonics every 10 minutes over a period of two weeks. Measurements were

taken from 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.

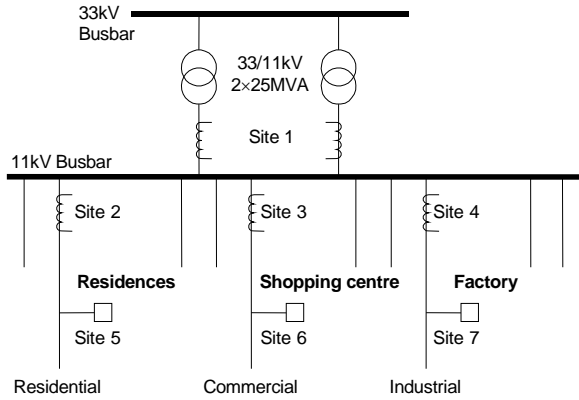


Figure 1: Schematic layout of test system

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 was pre-processed in text form before being entered into the ACPro software for analysis.

4. RESULTS AND OUTCOMES

As mentioned in Section 2, ACPro, a specialised data mining software package for the automatic segmentation of databases, was primarily used in this work. Segmentation (or clustering) using USL techniques was used to discover similar groups of records in the database, which in this case included clustering the harmonics 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. The outputs from the segmentation process were plotted using the graphical features of MS Excel. Figure 2 illustrates an example of the clusters discovered by the software for the voltage harmonics. Each cluster is data automatically grouped according to a learned pattern, and the abundance of each group is calculated over the full data range. In Figure 2 the clusters are labelled or allocated a type number that is ordered inversely proportional to the actual abundance, i.e. the most abundant cluster is seen as type 0 and those that are progressively rarer have a high value type numbers.

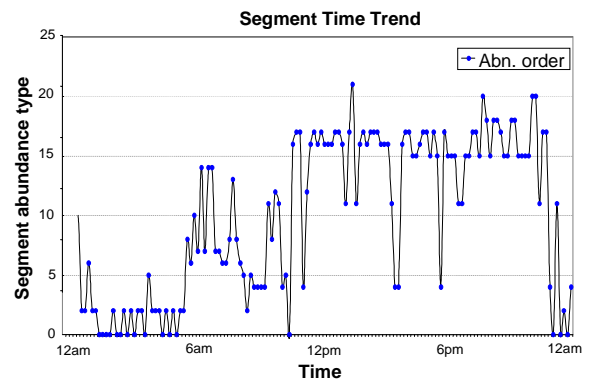


Figure 2: Abundance of clusters obtained through USL for harmonic voltage data over one day

Using data mining important patterns can be discovered when sorting the clusters by order of their mean value and relating these ordered clusters to their abundances. An example of this is shown in Figure 3 where it can be seen that the values of 5th harmonic currents and voltages from Site 7 are normal at high abundance level (to the left hand side) and have abnormal values at low abundance level (to the right hand side). The magnitudes of the harmonic voltages and currents of Figure 3 have been normalised to their respective mean values (expressed in per unit) and are shown to be below

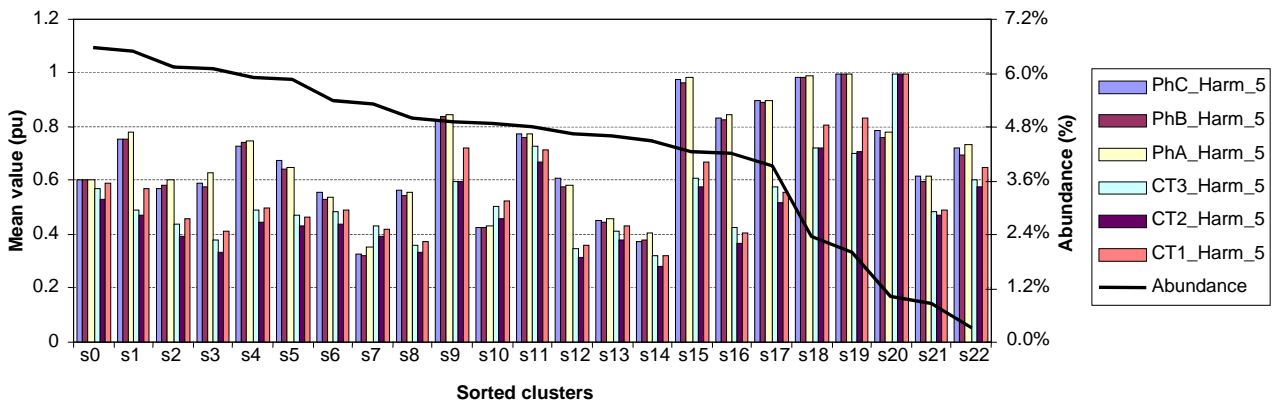


Figure 3: Abundance of clusters of 5th harmonic current and voltages over each phase of monitoring results where PhA, PhB, PhC are the phase voltages, and CT1, CT2, CT3 are the phase currents

average for a majority of clusters. Clusters that are of significant abundance and contain high magnitude PQ disturbances may lead to equipment damage or disoperation, and thus would indicate further investigation may be required.

It is proposed that the clusters obtained through USL may provide important information as to equipment operation, e.g. clusters containing high 5th harmonic currents and voltages in Figure 3 may indicate occurrences of significant customer harmonic contributions. The utility and/or customer may then use this information to carryout further investigations, coordinate harmonic emissions with other customers, or mitigate harmonics disturbance more effectively.

Several other interesting patterns may be extracted from the graphs of the obtained clusters. These patterns are easy to recognise and the new information that these patterns hold can be used to solve some simple power quality problems. The extracted information is in the form of relationships among harmonic data attributes that are explained in the following respective figures.

In an attempt to distinguish the most significant harmonic, the 5th harmonic current and total harmonic current distortion (ITHD) were analysed. The results of this analysis are illustrated in Figure 4. By inspecting the patterns of Figure 4, a strong relationship between 5th harmonic current and (ITHD) is verified, i.e. ITHD rises and falls with the value of 5th harmonic. This indicates that for the test system the 5th harmonic current is a significant contributor to the overall distortion level, as previously reported using traditional analysis in [19].

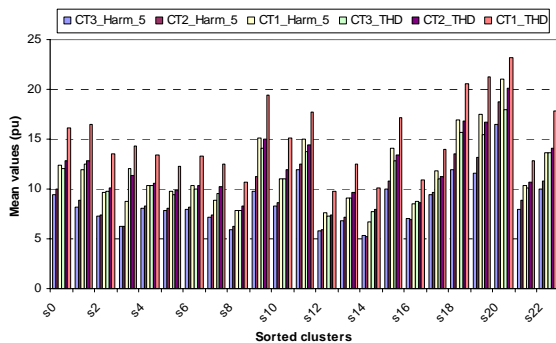


Figure 4: Cluster of 5th harmonic current and ITHD over all three phases from Site 7

USL cluster analysis was repeated for the 5th harmonic voltage and total harmonic voltage distortion (VTHD) of Site 6. The analysis was limited to 10 clusters, and the resulting clusters are illustrated in Figure 5. As shown in Figure 5, the strong relationship between ITHD and 5th harmonic

current in Figure 4 is repeated for VTHD and 5th harmonic voltage. This was also found to be true for all other sites monitored.

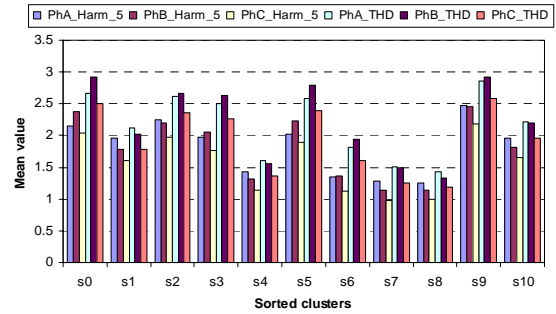


Figure 5: Clusters of 5th harmonic voltage and VTHD over all three phases from Site 6

By comparing clusters of 5th harmonic voltage and current, as shown in Figure 6, it can also be concluded that the 5th harmonic voltages arising are due mainly to the customer rather than contributions from upstream for Site 6. This is expected, as a significant component of the impedance that the LV customer sees will come from the MV/LV transformers. Also the harmonic contributions from the commercial load are expected to be high with significant power electronics driven loads (computers, air conditioners, etc.).

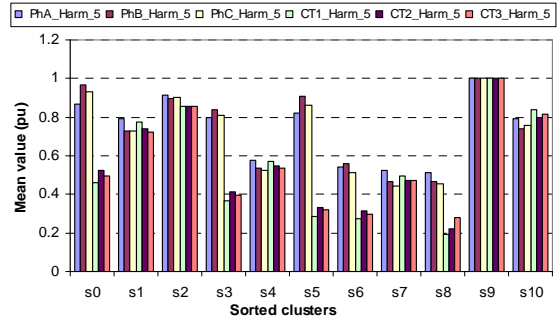


Figure 6: Clusters of 5th harmonic voltages and currents over all phase three phases from Site 6

The above analysis was repeated for the 19th and 49th harmonics (data for other harmonics not available). The resulting clusters of Figures 7 and 8 illustrate less distinct relationships for the 19th and 49th harmonics. However, it is unclear at this stage as to whether these results were affected by the limitations in readings of the instruments, as there relative magnitudes were quite small.

In general the least abundance clusters are the most interesting ones, as they are unexpected to occur. In power quality specifically these uncommon patterns are usually investigated to identify anomalies that occur as a result of network or load operation. As can be seen from Figure 9, the least abundant cluster

(s5) has the highest level of 5th harmonic current over all other clusters. This concept may also be used to identify the most significant distorting loads.

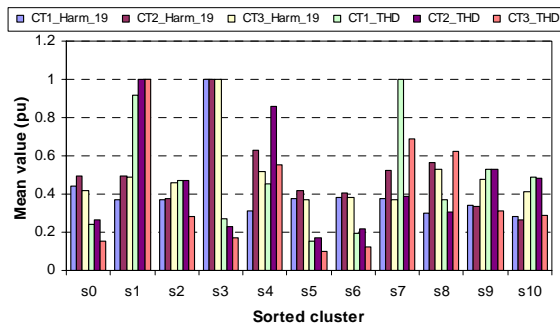


Figure 7: Clusters of fundamental and 19th harmonic current over all three phases from Site 6

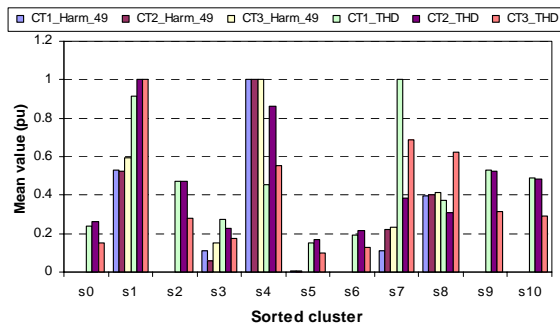


Figure 8: Cluster of fundamental and 49th harmonic current over all three phases from Site 6

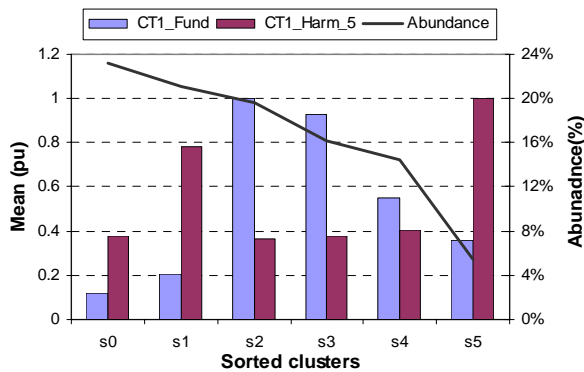


Figure 9: Fundamentals and 5th harmonic currents clusters in a single phase

Figure 10 illustrates the pattern of the clusters over the period of one week. In Figure 10 each cluster is represented with a colour in greyscale in proportion to the abundance of that cluster, i.e. the least abundant cluster will appear as black and the most abundant cluster will be the lightest shade of grey. Noticeable characteristics from Figure 10 include the two distinctive darker patterns towards the left hand side of the MV substation data (Site 1). This indicates the least abundant occurrences, appearing during the mornings of the weekend days. Also the

commercial site, Site 6, exhibits a recurring pattern of harmonics over each day, noting that the shopping centre is in operation seven days a week.

The residential and industrial customer clusters are somewhat more random than the other sites, suggesting that harmonic emission levels follow no well defined characteristics.

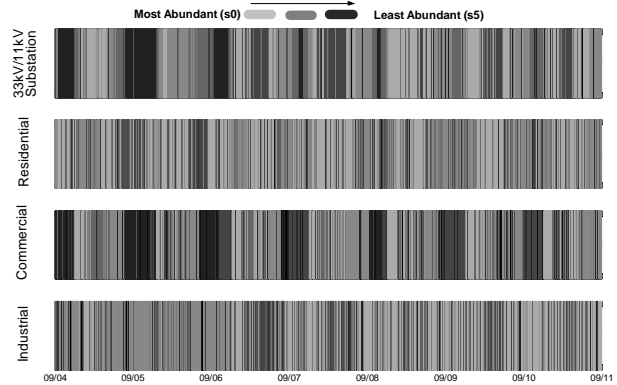


Figure 10: Clusters of harmonic emissions from the different customer loads and system overall for a one week period

Using techniques similar to the above mentioned it is perceived that data mining will become a useful tool for identifying additional information from PQ monitoring data, beyond that which is obtained from standard reporting techniques. This may include finding PQ disturbance “foot prints” for particular harmonic sources, as has been completed in Figure 10.

5. CONCLUSION

Power quality (PQ) data from an MV distribution system containing residential, commercial, and industrial customers has been analysed using data mining techniques. Data mining, in particular cluster analysis arising from unsupervised learning (USL) techniques, has been shown to be able to identify useful patterns within the data set. These patterns may help to identify the need for further investigation, coordinate harmonic emissions amongst customers, and assist in mitigation of PQ disturbances.

Significant results obtained from the cluster analysis include:

- (i) Verification of a strong relationship between 5th harmonic and THD levels for both current and voltage at customer sites.
- (ii) Significantly high harmonic disturbance levels usually only occur for short periods of time (low abundance).
- (iii) Identification of “footprints” of overall system harmonic distortion, and residential,

commercial and industrial customer harmonic emissions.

Future work will include more in depth analysis of clustering techniques, and identification of other useful patterns for reporting PQ data. Supervised learning (SL) will also be used to establish trends between plant equipment, load cycles, and PQ disturbances.

6. REFERENCES

- [1] R. Lamedica, G. Esposito, E. Tironi, D. Zaninelli, A. Prudenzi, "A survey on power quality cost in industrial customers", IEEE PES Winter Meeting, 2001.
- [2] R.C. Dugan, M.F. McGranahan, S. Santoso, H.W. Beaty, "Electrical Power Systems Quality", 2nd Edition, McGraw-Hill, 2002.
- [3] V. Gosbell, D. Robinson, R. Barr, V. Smith, "How should power quality be reported", EESA Electricity 2002 Conf., Canberra, 2002.
- [4] W.W. Dabbs, D.D. Sabin, T.E. Grebe, H. Mehta, "Probing power quality data", IEEE Computer Applications in Power, Vol.7, No.2, 1994, pp.8-14.
- [5] R. Groth, "Data Mining: Building Competitive Advantage", Prentice Hall, USA, 2000.
- [6] H. Mannila, "Data mining: machine learning, statistics, and databases", Proc. 8th Inter. Conf. on Scientific and Statistical Database Systems, 1996.
- [7] C. Olaru, L. Wehenkel, "Data mining", IEEE Computer Applications in Power, Vol.12, 1999.
- [8] S. Santoso, J.D. Lamoree, "Power quality data analysis: from raw data to knowledge using knowledge discovery approach", IEEE PES Summer Meeting, 2000.
- [9] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, "Advances in knowledge discovery and data mining", AAAI Press, Menlo Park, California, 1996.
- [10] C. Westphal, T. Balxton, "Data Mining Solutions: Method and tools for solving real-world problems", Wiley, USA, 1998.
- [11] J. Han, M. Kamber, "Data mining: concepts and techniques", Morgan Kaufmann Publishers, San Francisco, 2001.
- [12] O. Maimon, M. Last, "Knowledge Discovery and data mining: The info-Fuzzy Network (IFN) Methodology", 1st Edition, Kluwer Academic Publisher, 2001.
- [13] C. Wallace, "Intrinsic Classification of Spatially Correlated Data", The Computer Journal, Vol. 41, No. 8, 1998
- [14] C. Wallace, D. Dowe, "Intrinsic classification by MML – the Snob program", Proc. 7th Australian Joint Conf. on Artificial Intelligence, World Scientific Publishing Co., Armidale, Australia. 1994.
- [15] R. Munro, "Seneschal: classification and analysis in supervised mixture-modelling", Proc. 3rd Inter. Conf. on Hybrid Intelligent Systems, 2003.
- [16] P. Cheeseman, J. Stutz, "Bayesian Classification (AutoClass): Theory and Results", in "Advances in Knowledge Discovery and Data Mining", AAAI Press/MIT Press. 1996.
- [17] I.H. Witten, "Data mining: practical machine learning tools and techniques with Java implementations", Morgan Kaufmann, San Francisco, California, 2000.
- [18] J. Oliver, T. Roush, P. Gazis, W. Buntine, R. Baxter, "Analysis Rock Samples for the Mars Lander", American Association for Artificial Intelligence, 1998.
- [19] V.J. Gosbell, D. Mannix, D.A. Robinson, B.S.P. Perera, "Harmonic survey of an MV distribution system", AUPEC, Brisbane, 2000.