Utilising the Cross Industry Standard Process for Data Mining to Reduce Uncertainty in the Measurement and Verification of Energy Savings

Colm V. Gallagher, Ken Bruton, and D.T.J. O'Sullivan

Intelligent Efficiency Research Group, University College Cork, Cork, Ireland
c.v.gallagher@umail.ucc.ie

Abstract. This paper investigates the application of Data Mining (DM) to predict baseline energy consumption for the improvement of energy savings estimation accuracy in Measurement and Verification (M&V). M&V is a requirement of a certified energy management system (EnMS). A critical stage of the M&V process is the normalisation of data post Energy Conservation Measure (ECM) to pre-ECM conditions. Traditional M&V approaches utilise simplistic modelling techniques, which dilute the power of the available data. DM enables the true power of the available energy data to be harnessed with complex modelling techniques. The methodology proposed incorporates DM into the M&V process to improve prediction accuracy. The application of multi-variate regression and artificial neural networks to predict compressed air energy consumption in a manufacturing facility is presented. Predictions made using DM were consistently more accurate than those found using traditional approaches when the training period was greater than two months.

Keywords: Measurement and Verification, Data Mining, Energy Efficiency, Baseline Energy Modelling

1 Introduction

The European Union has issued the Energy Efficiency Directive (2012/27/EU) to ensure member states shift to a more energy efficient economy [1]. The effective implementation of the Directive relies heavily on the cumulative effect of energy savings across a number of projects. An example of this is the Energy Efficiency Obligation Scheme (EEOS) in Ireland, which is being implemented pursuant to Article 7 of the Directive [2]. Ireland has chosen to use the EEOS as a mechanism to ensure targets set out in the Directive are achieved. The scheme obligates energy distributors and retail energy sales companies to achieve energy efficiency improvement targets based on their market share. This structure requires the aggregated savings of multiple individual projects to reach these targets. Over or under estimation of savings in individual cases can lead to national targets not being met, therefore failing to achieve the overall objective of the Directive.

In the energy efficiency sector, Measurement and Verification (M&V) is the process of quantifying energy savings delivered by an Energy Conservation Measure (ECM). M&V is a requirement of a certified Energy Management System (EnMS). The Efficiency Valuation Organization publish standardised M&V guidelines entitled the International Performance Measurement and Verification Protocol (IPMVP). IPMVP is a framework of definitions and broad approaches for M&V. In addition to this, ASHRAE publish Guideline 14P, which provides detail on implementing M&V plans. Both guidance documents require metering to estimate energy savings.

There are two periods of analysis in the M&V of energy savings: the pre-ECM period and the post-ECM period. In most cases, real data is available for the pre-ECM period and post-ECM period (measured energy). To quantify the savings achieved, the energy consumption post-ECM must be compared to what the consumption would be had the ECM not been implemented. This is known as the adjusted baseline. This requires the baseline energy consumption in the pre-ECM period to be modelled and used to quantify the adjusted baseline in the post-ECM period by normalising post-ECM consumption to pre-ECM conditions. Figure 1 contains a graphical representation of this calculation process. Hence, M&V is not an exact science as there is always a margin of error in predicting energy consumption [3].

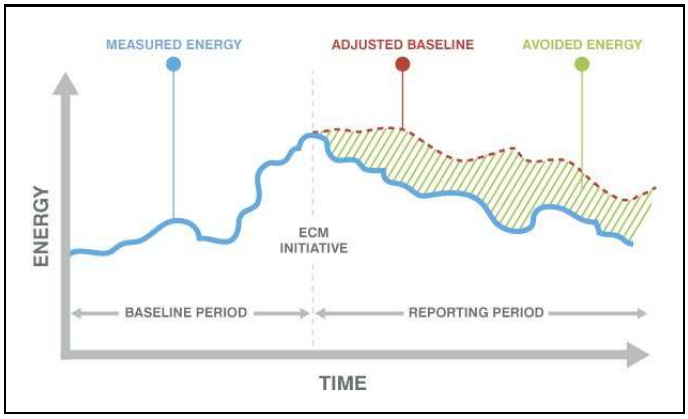

Fig. 1. Overview of measurement and verification savings calculation [4]

The normalisation of the adjusted baseline to an acceptable degree of certainty is a critical step in M&V. The methods most commonly used to predict the adjusted baseline are reviewed in Sect. 2.3. This study proposes an alternative M&V methodology that utilises data mining (DM) to harness the power of the energy data available. DM is being utilised as a mechanism to progress M&V of energy savings towards more accurate and reliable results. This is possible as DM enables efficient processing of the data and prediction of the adjusted baseline, hence maximising the potential power of the available data.

2 Related Work

2.1 Data Mining in Energy Engineering Applications

Continual improvement and development is a vital component in the success of an EnMS. This requirement is being satisfied through the use of DM to support and implement these systems. DM has successfully been used to define, develop and implement an EnMS. Velázquez et al. utilised a DM approach to identify key performance indicators and subsequently energy consumption models [5].

Energy consumption has also been predicted through the use of DM with a view to allowing for more informed decision making. This was achieved by extracting information from unstructured data sources, processing the data and presenting it in a manner that maximises its use to the decision maker [6]. If patterns in energy consumption are not obvious, DM has been shown to be the most effective mean of capturing consumption trends in the case of efficiently maintaining buildings [7]. Also, in a study which applied DM techniques to optimise building heating, ventilation and air-conditioning performance, DM has been shown to accurately model the operation of buildings, when supplied with sufficient data [8].

2.2 Modelling Energy Consumption

In engineering, the ability to predict electrical loads in an accurate manner is a valuable tool in demand side management. A DM approach using unsupervised learning has been taken in M&V to estimate baseline load for demand response in a smart grid [9]. The self-organizing map and K-means clustering were the modelling techniques applied. The results of which were compared to the accuracy of day matching methods, which is a simplistic approach to predicting energy consumption. Root mean square error was reduced by 15-22% on average through the use of DM techniques. Hence, the suitability of more complex modelling algorithms was vindicated [9].

The use of soft computing models to improve the accuracy of electrical load forecasting has also been investigated. Neuro-fuzzy systems have been proven to perform better than artificial neural networks and statistical forecasting based on Box-Jenkins ARIMA model [10]. Electrical load forecasting for smart grids analyse data across a larger project boundary than a typical M&V case. M&V of energy savings generally focuses on a smaller project boundary which in some cases is as large as an entire facility, while in other cases it can be as small as only covering a single piece of equipment. Hence, the application of DM techniques, such as those mentioned, to M&V of smaller scale electrical loads should be investigated.

2.3 Baseline Modelling in M&V At Present

Modelling techniques that are commonly used in industry include linear regression, day matching and change-point regression models. Walter et al. stated

that a limitation of methods used for predicting baseline energy is the inadequate quantification of uncertainty in baseline energy consumption predictions. The importance of uncertainty estimation is highlighted as being essential for weighing the risks of investing in ECM [11]. Reviews of the possible modelling techniques have been carried out. One such study compares five models which include change point models, monthly degree-day models, and hourly regression models. This presented a general statistical methodology to evaluate baseline model performance. The study showed that results generated using 6-months pre-ECM data, i.e. training data, may be just as accurate as those that use a 12-month training period [12]. Crowe et al. investigated using baseline regression models for individual homes to move towards an automated M&V approach using interval data. The results were found to offer a promising first step in the process, while recommendations were made to develop more a robust M&V methodology [13].

This paper assesses the suitability of using DM to provide this robustness in M&V. The need to progress the methods used to quantify the adjusted baseline in M&V projects has been highlighted. As the methods used at present are rigid in nature, they tend to generalise the variables affecting energy consumption. DM is proposed to progress this aspect of M&V through the use of complex modelling techniques that are capable of capturing the trends of energy consumption. The case study presented reviews the application of data mining for energy consumption prediction in a biomedical manufacturing facility. Each stage of the data mining process is detailed within the context of the case study.

3 Application of CRISP-DM for Purposes of M&V: A Case Study in a Manufacturing Facility

The CRoss Industry Standard Process for Data Mining (CRISP-DM) was identified as a method to further standardise the M&V methodology and enable more accurate estimation of energy savings. A case study was carried out to assess the viability of the methodology proposed in this paper. Figure 2 outlines the procedure applied. Rapidminer Studio (v7.0.001) was the software used to implement the CRISP-DM [14].

3.1 Business Understanding

A biomedical manufacturing facility was chosen as a case study to assess the feasibility of the application of DM to aid M&V. A quality understanding of the business under analysis was essential to interpret the results at the modelling and evaluation stages of the process. This was achieved by carrying out a process walk-through, studying process flow diagrams, and piping and instrumentation diagrams. A knowledge of the systems within the boundary of analysis was acquired from this process and any additional issues were discussed with the facility's engineering team. The boundary of the analysis was the electrical energy consumption across the entire manufacturing facility.

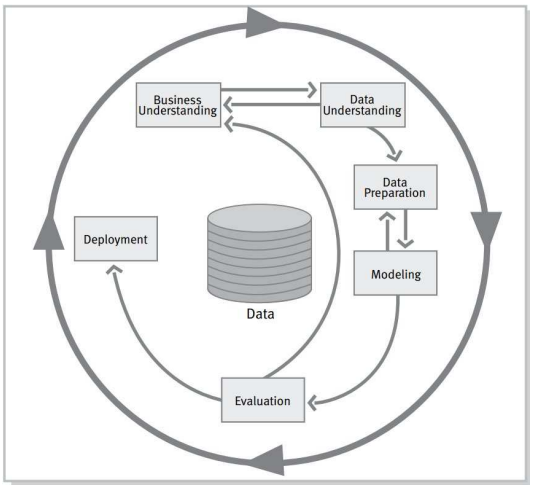

Fig. 2. Phases of the CRISP-DM reference model [15]

3.2 Data Understanding

The data understanding phase of the CRISP-DM reference model was completed through investigation into the information technology infrastructure at the facility. An understanding of the flow of energy consumption data and the databases in which it was stored was gained. This enabled the data preparation stage detailed in Sect. 3.3 to be carried out in an efficient manner with all pre-processing completed using as little resources as possible. The objective of streamlining the M&V process was considered at each stage of the study.

3.3 Data Preparation

Energy consumption data is often difficult to compute due to the nature of the metering. Cumulative meters are generally used for electrical energy and as a result, pre-processing must be completed on the outputted data. In the case under investigation, this was completed prior to being output to the user. However, despite this pre-cleansing of the data, outliers remained in the data set as the pre-cleansing process did not remove all anomalies. Therefore, the data preparation stage was utilised to remove any remaining outliers in the data set delivered to the user. Figure 3 illustrates the steps undertaken to prepare the data for the modelling stage.

Two data sources were used to gather the data required for a complete analysis of the electrical energy consumers on-site: energy management software and wind turbine management software. The electrical energy consumed on-site is measured by cumulative kilowatt-hour (kWh) meters. Pre-processing of this data involved detecting outliers caused by meter errors and converting the data from kWh to average electrical loads in kilowatts (kW). The second step was

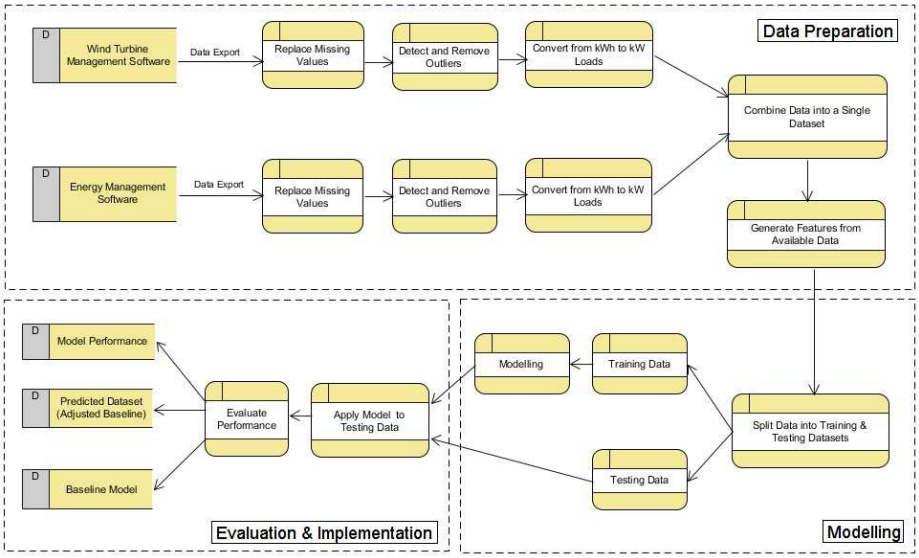

Fig. 3. Application of CRISP-DM in case study

required in order to analyse all data in the same format and units. The individual datasets were then joined together and forwarded to the modelling stage described in Sect. 3.4.

The use of data mining methods to pre-process the data reduced the resources required to clean the dataset into a useful form. Without the use of these DM techniques, the detection of anomalies can be a manual, time-intensive process. The process developed in the software environment can be applied to any data export from the systems within the analysis boundary. Hence, future projects can utilise this resource, further streamlining the M&V process.

3.4 Modelling

The dataset output from the data preparation stage was in a clean and functional format as a result of the data cleansing performed. For the purposes of this case study, the compressed air load was the chosen quantity to be modelled, as it was the most appropriate variable to highlight the power of the available energy data. When the load was analysed at a high-level, there was no clear and obvious correlation to other significant energy users on-site. The other significant energy users were more predictable due to scheduling of equipment and the presence of standard operating procedures. An electrical meter measured the total electrical energy consumed by the compressed air generation system in 15-minute intervals.

The modelling techniques applied to the dataset were multivariate linear regression and feed-forward neural networks. Following a review of the techniques that could be applied to the dataset, these were found to be the most appropriate

for this analysis. A total of 11 variables were available to be used to construct a model for compressed air consumption. These variables accounted for 44% of electricity consumption across the facility. The modelling process was developed within the software environment and the variables utilised to build each model were chosen automatically by the software based on significance to the attribute being predicted.

A traditional M&V approach was also considered for the purposes of evaluating the effectiveness of the techniques proposed. This approach consisted of single variable linear regression modelling. For M&V purposes, this is an approach that most practitioners are familiar with through the use of Microsoft Excel [16]. A correlation matrix was generated to assess which variable in the dataset had the greatest influence on the compressed air load, the quantity to be predicted. The electricity consumed by production related equipment had the highest correlation coefficient with a value of 0.902, hence this was chosen to be used as the input variable for developing the single variable regression model for compressed air consumption. Figure 3 illustrates the modelling process carried out.

Following initial modelling using the techniques described above, the results were evaluated. As per the CRISP-DM methodology, the data understanding stage was revisited within the context of the modelling process. The process of constructing the models was then refined to ensure the knowledge contained in the dataset was accurately represented and the power of this knowledge was maximised. This was achieved by choosing modelling parameters, such as tolerance levels and learning rates, in a heuristic manner. For the feed-forward neural network, this process identified a learning rate of 0.3 and a total of three layers as the most appropriate for the data under analysis. Similarly, the minimum tolerance used to construct the linear regression models was chosen as 0.05.

3.5 Evaluation

The models developed in Sect. 3.4 were applied to a new dataset containing energy consumption knowledge for a time period outside that with which the models were constructed. This independent dataset was used for cross validation of the models. Relative error was the metric used to evaluate the performance of each model. The modelling technique that resulted in the lowest relative error was deemed as the most appropriate for the given data set. The equation for N data points is as follows:

$$Mean\,Relative\,Error = \frac{1}{N}\sum_{i=1}^{N}\frac{pred(i)\ - real(i)}{real(i)}\ . \tag{1}$$

Table 1 contains the relative error of each model in predicting the compressed air load during the testing period. The training period duration was varied to assess the affect that this had on model performance. It was found that the single variable regression model is the most appropriate modelling technique in cases where the training data available is less than 62 days, as this technique

minimised the relative error. For greater training periods, the multivariate linear regression model and feed-forward neural network performance improved greatly and surpassed that of the single variable regression model.

Table 1. Relative error in predictions from each model.

Training Period	Testing Period	Single Variable Linear Regression	Multivariate Linear Regression	Feed-forward Neural Network
31 Days	28 Days	40.4%	64.32%	46.81%
62 Days	28 Days	37.18%	38.18%	58.26%
92 Days	28 Days	29.92%	22.09%	23.09%
111 Days	28 Days	28.67%	21.49%	28.16%
139 Days	28 Days	26.84%	19.87%	21.6%

Multi-variate linear regression performed with the lowest relative error for the longest training period. As energy data is largely affected by the outdoor air temperature, historical data over a 12-month period is usually used for analysis. This ensures a wide range of operating conditions are contained within the data. A full cycle of a facility's operation is also required to capture all levels of energy consumption. The manufacturing facility in this study operates continuous processes so a full cycle of operation is not as defined as that of a batch process. Hence, a default value of 12-months of data would suffice. However, approximately 6-months of data that was fit for use was available at the time of this analysis.

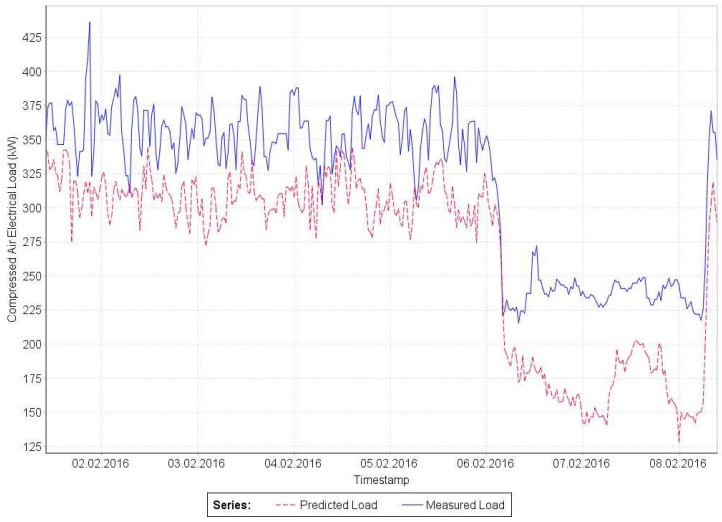

Fig. 4. Sample of prediction performance of multi-variate linear regression model

Taking the prediction accuracy of the models developed across the range of training periods, multi-variate linear regression was the most appropriate technique for modelling the baseline compressed air energy consumption in this case. Figure 4 illustrates a sample of the prediction accuracy of this leading performing model, developed using 139 days of training data. The decrease in compressed air load on February 6 is as a result of reduced plant operations on weekends. The performance gap between the traditional modelling technique and multi-variate linear regression was approximately constant at 7% for training periods greater than 92 days. A longer period of analysis must be considered to investigate this relationship further.

3.6 Deployment

Deployment of the models generated would consist of predicting the adjusted baseline in the post-ECM period. This would then be compared with metered data to identify the savings made. For an M&V project, the largest training dataset would be used in order to capture the widest range of production levels possible. Hence, given that the models constructed using a DM approach were more accurate than those constructed using a traditional modelling approach, deployment in an M&V application would improve the performance of the energy savings estimation.

4 Conclusions

The prediction of compressed air electricity consumption in a large biomedical manufacturing facility was chosen to apply and review the proposed methodology. Multi-variate linear regression models and feed-forward neural networks were constructed using a variety of training periods. Single variable linear regression was also performed as it is a common approach used in M&V. Cross-validation of all models found that the traditional approach was more appropriate when the training period was two months or less. Training periods longer than this resulted in the models developed using a DM approach performing with improvements in relative error of approximately 7%. The data mining process was found to improve the prediction accuracy, therefore reduce the uncertainty in energy savings estimation for M&V purposes. More detailed analysis using a calendar year of training data should be performed to further substantiate the findings of this case study. The use of DM to improve performance and reduce the resources required for M&V was presented with promising results for this case study. One can broaden the scope of this analysis across a variety of M&V applications to advance the research into the subject area.

5 Future Work

The research presented in this paper is part of a wider study which has a primary objective to move towards automating the process of M&V. The results of the

CRISP-DM included in this paper form an initial evaluation of the potential and effectiveness of DM in these circumstances. Future research will consist of broadening the scope of the analysis presented in this paper through the use of larger datasets, alternative case studies and more complex baseline modelling techniques.

Acknowledgements. The authors would like to acknowledge the financial support of the Science Foundation Ireland MaREI Centre and the NTR Foundation.

References

1. European Parliament: Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on energy efficiency (2012)
2. Sustainable Energy Authority of Ireland: Energy Efficiency Obligation Scheme. Technical Report (2014)
3. Drive, B., Afb, T.: Measurement & Verification Handbook. Technical Report (1998)
4. C^3 Resources, `http://www.c3resources.co.uk`
5. Velázquez, D., González-Falcón, R., Pérez-Lombard, L., Marina Gallego, L., Monedero, I., Biscarri, F.: Development of an energy management system for a naphtha reforming plant: A data mining approach. Energy Conversion and Management 67, 217–225 (2013)
6. Peral, J., Ferràndez, A., Tardío, R., Maté, A., de Gregorio, E.: Energy consumption prediction by using an integrated multidimensional modeling approach and data mining techniques with Big Data. In: Indulska, M., Purao, S. (eds.) ER 2014. LNCS, vol. 8823, pp. 45–54. Springer, Heidelberg (2014)
7. Gao, Y., Tumwesigye, E., Cahill, B., Menzel, K.: Using Data Mining in Optimisation of Building Energy Consumption and Thermal Comfort Management In: Kou, G., Peng, Y., Franz, I.S., Chen, Y., Tateyama, T. (eds.) SEDM 2010, pp. 434–439. IEEE Xplore (2010)
8. Ahmed, A., Korres, N.E., Ploennigs, J., Elhadi, H., Menzel, K.: Mining building performance data for energy-efficient operation. Advanced Engineering Informatics 25, 341–354 (2011)
9. Park, S., Ryu, S., Choi, Y., Kim, H.: A framework for baseline load estimation in demand response: Data mining approach. In: IEEE SmartGridComm 2014, pp. 638–643. IEEE Xplore (2014)
10. Abraham, A., Nath, B.: A neuro-fuzzy approach for modelling electricity demand in Victoria. Applied Soft Computing 1, 127–138 (2001)
11. Walter, T., Price, P.N., Sohn, M.D.: Uncertainty estimation improves energy measurement and verification procedures. Applied Energy 130, 230–236 (2014)
12. Granderson, J., Price, P.N.: Evaluation of the Predictive Accuracy of Five Whole-Building Baseline Models. Technical Report, Lawrence Berkeley National Laboratory (2012)
13. Crowe, E., Reed, A., Kramer, H., Kemper, E., Hinkle, M.: Baseline Energy Modeling Approach for Residential M & V Applications. Technical Report, Northwest Energy Efficiency Alliance (2015)
14. RapidMiner Studio 7, `https://rapidminer.com/products/studio`
15. The Modeling Agency, `https://the-modeling-agency.com/crisp-dm.pdf`
16. Bonneville Power Administration: Regression for M & V: Reference Guide. Technical Report May (2012)