# Classification of Energy Consumption in Buildings with Outlier Detection

Xiaoli Li, Chris P. Bowers, Thorsten Schnier, *Member, IEEE*

***Abstract*—In this paper, we propose an intelligent data analysis method for modelling and prediction of daily electricity consumption in buildings. The objective is to enable a building management system to be used for forecasting and detection of abnormal energy use. First, an outlier detection method is proposed to identify abnormally high or low energy use in building. Then a canonical variate analysis is employed to describe latent variables of daily electricity consumption profiles, which can be used to group the data sets into different clusters. Finally, a simple classifier is used to predict the daily electricity consumption profiles. A case study, based on a mixed use environment, was studied. The results demonstrate the method proposed in this paper can be used in conjunction with a building management system to identify abnormal utility consumption and notify building operators in real time.**

***Index Terms*—Energy management, Outlier detection, Electricity data, Canonical variate analysis, Modelling, Prediction**

## I. INTRODUCTION

ENERGY consumption of buildings (both residential and commercial) has steadily increased, reaching figures between 20% and 40% in developed countries [1]. The rise of energy demand in buildings will continue in the near future because of growth in population, long-term use of buildings and increasing demand for improved building comfort levels. Therefore, the energy efficiency of buildings is of prime concern for anyone wishing to identify energy savings. To this end, automated meter reading and smart metering systems have been employed to collect building energy data [1]. The aim of this data is to provide greater insight into how a building consumes energy and therefore what improvements are likely to be most effective in reducing consumption. It is estimated that improvements in building energy efficiency could contribute to the reduction of the current energy consumption of buildings by at least 20% in the EU, which is equivalent to saving 60 billion Euros annually [2]–[4]. Unfortunately metering data requires significant processing and analysis [5] which makes attaining these reductions a difficult and costly process.

Some intelligent methods have been proposed for the automated analysis of energy consumption data within a buildings energy management system. Electrical peak load forecasting plays a significant role in effective and economic operation of power utilities, Amin-Naseri and Sorroush [6] proposed a hybrid neural network model, by integrating a self-organizing map and a feed forward neural network, to predict daily electrical peak load. Developing an accurate and robust peak load forecasting methodology can lead to more accurate forecasting of electricity consumption, so we can significantly reduce the cost of operating power systems [7]–[10].

These methods are not computationally efficient and do not easily lend themselves to building energy management and control systems since the results are often difficult to interpret.

To ensure energy is being used most efficiently, a comparison between similar energy consumers can be used to examine consumption habits. Räsänen et al [11] used a self-organizing map approach to automatically create comparison groups based on individual building characteristics. The models, based on physical principles, can be used to predict the energy consumption in a building, but many parameters need to be accurately defined. A data-driven approach, based upon an artificial neural network, can replace a model based on physical principles [12]. Such a data-driven approach can forecast daily energy consumption, including heating, cooling and other electrical loads. This can also include the prediction of start-up and shut-down behaviour of an electrical system.

Although several intelligent methods have been proposed to predict energy demand, the fact that historical consumption data often contains abnormal building behaviour is often ignored. Successful approaches to energy data analysis must be reliable and effective yet robust in the analysis of captured building energy data. On the other hand building managers still require software and control systems to be easy to use and understand.

These intelligent methods can also be used to maintain occupant comfort levels within a building whilst minimising energy consumption [5], [13]. In particular, the use of intelligent decision support models for energy managers, energy auditors and the optimised operation of their units is of significance. Additionally, we could also make use of the recorded data to detect abnormal energy use. In [14], [15], an intelligent data analysis system was proposed which can automatically detect abnormal energy use in building, this information can be used to notify building mangers of issues with minimal delay, helping to reducing energy costs. For example, the detection of abnormal energy or other utility consumption may indicate wear or malfunctioning equipment, so it can also play a role in maintenance scheduling and early warning of equipment failure.

In this study, we propose an intelligent data analysis method for the automated classification of energy consumption profiles of a building. The classification model can be used to detect abnormal usage and forecasting energy consumption of a

Manuscript received 18 December, 2008, revised 24 April, 2009. This work is supported by the Engineering and Physical Sciences Research Council, U.K.



The authors are with the School of Computer Science at the University of Birmingham, B15 2TT, UK. (email: T.Schnier@cs.bham.ac.uk)

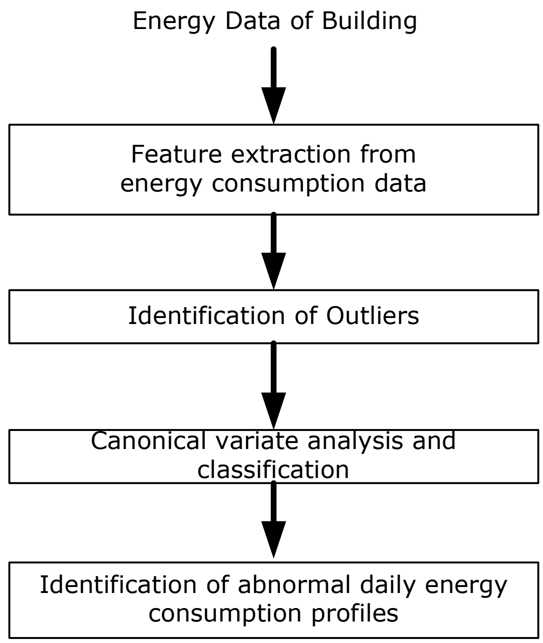



Fig. 1. The outline of the identification and model of the daily energy consumption profiles

building. In addition, such a robust classification model can provide a powerful tool for further automated analysis of energy consumption of buildings.

The next section describes the method, including the feature extraction of the daily energy consumption profile, the outlier detection of the daily energy consumption, the canonical variate analysis and a simple classifier. Finally, a case study is presented to demonstrate the performance of the method proposed in this paper.

## II. Methods

An outline of the method presented in this paper is shown in figure 1. The input to this method is in the form of time series data. In this study we use half-hourly energy consumption data, but other data sources such as gas, water consumption, heating loads, etc. are all viable.

The first step in this method is to extract features from the time series data, including the removal of seasonal variances in demand. The second step is to detect outliers in the existing data and remove these outliers. The remaining time series data is then used for further modelling. The final step is to group the daily energy consumption series in clusters. A clustering algorithm is used to classify the new data sets. The behaviour of the energy consumption can then be predicted. The details of each step in this method are now described.

### A. Feature extraction

Abnormal consumption of electricity should be detected and discarded with an outlier identification procedure before the construction of clusters for daily energy consumption. First, we need to extract the features from the energy time series data. Then, a statistical method is applied along with an outlier detector in order to identify and remove abnormal energy consumption.

The statistical features extracted from the daily energy consumption profile include the mean daily energy consumption, $m$, and the peak daily consumption, $pe$. To further explore the dynamical change (or dynamic characteristics) of the daily energy consumption data, an auto regression model (AR) is applied to analyse the daily energy consumption series.

Assuming that the value of the current sample, $y(t)$, in a data sequence $y(1), y(2), \ldots, y(N)$, can be predicted as a linearly weighted sum of the $p$ sample values, $y(t-1), y(t-2), \ldots, y(t-p)$, where $p$ is the model order and is generally chosen to be much smaller than the sequence length, $N$. The predicted values of $y(t)$ can be expressed as

$$y(t) = w + \sum_{i=1}^{p} a_i y(t-i) + e \tag{1}$$

where the weight $a_i$ denotes the $i$th coefficient of the $p$th-order model, $w$ is the intercept variable and $e$ is a noise parameter. Given the order $p$ of an AR($p$) model, the parameters $w$, $a_i$, and $e$ can be estimated by the stepwise least squares algorithm [16], [17]. Two algorithms, the Levinson-Durbin algorithm and Burg algorithm, can be used to estimate the coefficients of the AR model. A correct value of $p$ for a given data sequence is not known in advance, it is desirable to minimize the model's computational complexity by choosing the minimum value of $p$ that adequately represents the series being modelled. In this study, $p$ is given a value of 2 which provides sufficient fit to the data whilst simplifying further data analysis by resulting in only two coefficients. As a result, the new features include $a_1$ and $a_2$ from the daily energy consumption. Finally, we may extract four features from the daily energy consumption, the feature vector is $D = m, pe, a_1, a_2$.

### B. Outlier detection

Outliers are defined as data that appears to be inconsistent with the rest of the data set. Statistically, these values are numerically distant from the remaining data. There is no rigid mathematical definition or ubiquitous method to determine whether or not an observed data point is an outlier. Some methods have been proposed to detect outliers in univariate and multivariate data sets [18]. In this study, we will employ two outlier detection methods for different sample sizes.

The first method is the generalized extreme studentized deviate (GESD). This outlier detection method is based on an assumption that the measured data has a normal distribution. This is because the outlier detection is conducted by using the mean and standard deviation of the data observed. In this method, two parameters should be determined: an upper bound, $Nu$, on the number of potential outliers and the probability, $\alpha$, of incorrectly declaring one or more outliers when no outliers exist. $Nu \leq \frac{1}{2}(N-1)$, where $N$ is the number of samples. Typically this method works well with sample sizes of more than 10 with a normal data distribution.

For smaller samples sizes, a second outlier detection method is used, the Q-test. First, the set of samples under examination are arranged in ascending order. The Q-value, $Q$, is then calculated as the ratio of the gap, between a sample and its neighbour, and the range of the samples. A critical Q-value, $Qc$, is defined with a confidence level (e.g. 95%). Finally, if $Q > Qc$, then the suspect value can be characterized as an outlier. Because this method is based on ordered statistics, there is no need to assume a normal distribution. In addition, this method performs well with small sample sizes. In this study, when the sample size is less than 10, the Q-test method is the chosen method for outlier detection.

For each cluster, any identified outliers are removed from the data set. The GESD or Q-test detection method are used to identify outliers for each individual feature based on the sample size. Thus, if there is more than one feature, the outlier detection method is used multiple times to determine outliers for a particular cluster.

### C. Canonical variates analysis

Canonical variate analysis (CVA) is a multivariate supervised discriminant tool which projects the original data into new axes called canonical variables (CVs) [22]. CVA estimates the space vectors which maximize the differences between groups in the original data according to a discrimination criterion which is aimed at maximising the separation between classes whilst minimising the separation within classes. Thus, the CVA method is more suitable for classification purposes than principle components analysis (PCA) [21]. A full mathematical description of the CVA approach can be found in [19] and [20]. Here we give a brief outline.

Given a data set $X_{N*M} = \{x_{ij}\}$, where $N$ is the sample number and $M$ is the number of variables, this data set is divided into $K$ groups, each group contains $n_i(i = 1, \ldots, K)$ samples. The within-group covariance matrix and the between-group covariance matrix are defined as

$$C_W = \frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T \tag{2}$$

$$C_B = \frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \tag{3}$$

where $x_{ij}$ is the $j$th sample in the $i$th group, $\bar{x}_i$ is the mean vector in the $i$th group and $\bar{x}$ is the mean vector in the overall samples.

To obtain CVs, a direction **W** needs to be determined, which satisfies the condition

$$C_B\mathbf{W} = \lambda C_W\mathbf{W} \tag{4}$$

where $\lambda$ and **W** represent the eigenvalue and eigenvector respectively. Once the direction **W** is determined, the canonical variates (CVs) may be estimated by

$$CV = X_c\mathbf{W} \tag{5}$$

where $X_c$ is the mean-centred data matrix. The number of CVs is $K-1$ ($K$ is the number of the group). If the matrix $C_W$ is non-singular, equation 4 can be written as

$${C_W}^{-1}C_B\mathbf{W} = \lambda\mathbf{W} \tag{6}$$

where the direction **W** solution can be considered an eigenvector of the matrix $C_W^{-1}C_B$ . As a result, the CVs can be obtained by (5).

In practice if the matrix $C_W$ is singular for multi-collinear data, for example the data in figure 4(a), a new optimal method is needed to estimate the direction **W**. In Nrgaard et al [20], Eq. (4) is re-written as a regression equation, such that

$$\mathbf{Y} = C_B\mathbf{B} + \mathbf{R} \tag{7}$$

where the columns of **Y** are the difference between each group mean and the overall mean, the columns of **B** are the directions for each group in **W** and **R** is a residual matrix [22]. Eq. (7) may be solved using a partial least squares approach.

### D. Classifier

After the CVA, the CVs are obtained, then a linear discriminate analysis (LDA) classifies the samples into groups [22]. The discriminate function is defined as

$$\begin{aligned} F_i(CV) = \log(\pi_i) \\ & -1/2(CV - \overline{CV_i})^T C_{W,CV}^{-1}(CV - \overline{CV_i}) \\ & + \log|C_{W,CV}| \end{aligned} \tag{8}$$

where $i$ is the group number $(i = 1, \ldots, K)$, $\pi_i$ is the prior (and is equal to $\frac{1}{K}$), $CV$ is the CVs of the sample to be classified, $\overline{CV_i}$ is the average of CVs for group $i$ and $C_{W,CV}$ is the covariance matrix of the CVs. The group classification is chosen to be the $i^{th}$ group with the maximum value of $F_i$.

## III. Results

A case study is presented using real world data provided by a company based in Birmingham, UK. The data captures overall building electrical consumption, measured at 30 minute intervals, for an entire year (2006) on the site.

Figure 2 displays a time series plot of a months (January) electrical consumption. Periodical changes over a week can be seen. Typically, at the weekend, the site is closed, so the consumption of the electrical systems will be minimal. In contrast, during the working week (Monday-Friday), the electrical consumption look more sinusoidal in form. For each week there exists some variance due to production factors and climate effects (temperature, heating or cooling), however the global trends of the workdays are similar.

Figure 3 is a daily energy consumption series for Sundays only during the period of January through June. Based on observations on this time series data, the energy consumption on some Sundays is abnormal. To reduce seasonal effects, the consumption of each day is normalised by simply subtracting the minimal value of consumption on that day. The extracted features include the mean, peak and other model parameters. The outlier analysis method, utilising these daily profile features, identifies three abnormal cases, or Sundays,

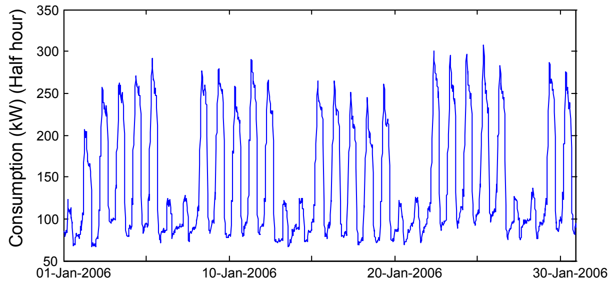



Fig. 2. Daily energy consumption profiles for January 2006

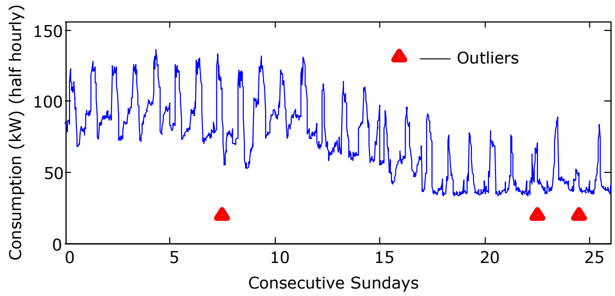
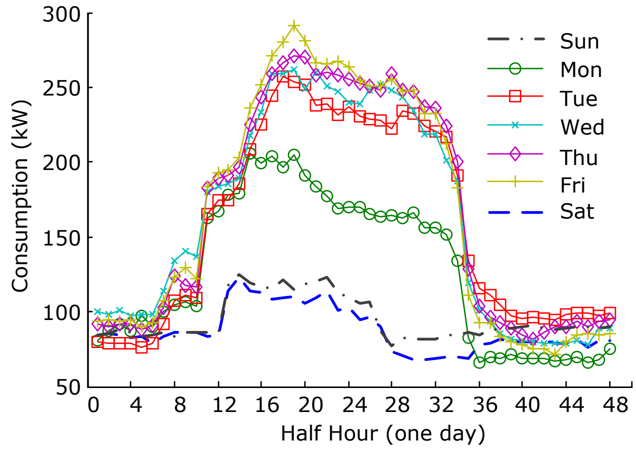



Fig. 3. The outlier detection of daily energy consumption on Sundays (January-June)

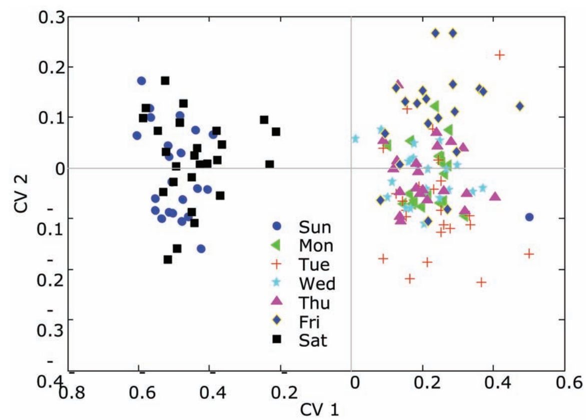
which are indicated by solid triangles as shown in figure 3. The results are in agreement with human observations. These are examples of cases that would demand further investigation by the site energy manager. Once these outliers are detected, they are discarded, and the remaining data is used for further modelling.

To highlight the variability in daily consumption profiles, the energy consumption profiles of a week are plotted in figure 4(a). Since we have no prior knowledge of the composition of energy consumption, we will divide the data set into 7 groups ($K = 7$), corresponding to each day of the week. All public holidays are categorised as Sundays. In this study, the first half of the data corresponding to the months January to May inclusive, is analysed using the CVA method, the first two canonical variates are plotted in figure 4(b). Two clear clusters are observed, one corresponds to non-workdays, including weekends and holidays, the other to workdays. This method contains six canonical directions. Linear Discriminate Analysis (LDA) is applied to the six canonical variates to form a classifier for energy consumption profiles. The remaining data from June to December are used to test the model. The weekend days and workdays can be exactly identified. However, the differences amongst workdays can not be identified, (see figure 4(b)).

In terms of the above analysis, we may divide the daily energy consumption profiles into two classes, workdays and non-workdays. First, we analyse the two groups (workday and non-workdays, K=2) based on the time series data upon which the model was built upon, from January to May inclusively. The number of canonical directions is one-dimensional (K-1). Figure 5 shows the canonical weight vector and the canonical variates of the 151 samples. As can be seen in figure 5(a), the largest canonical weights are concentrated in the ranges 8-




(a) Typical energy consumption profile for 7 days



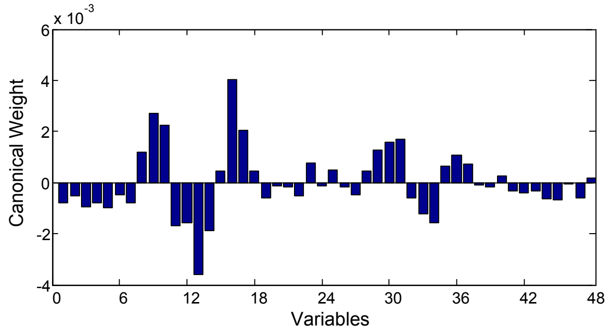

(b) 1st and 2nd Canonical Variates of day of week

Fig. 4. Canonical Variate Analysis of seven days of the week (January-May)

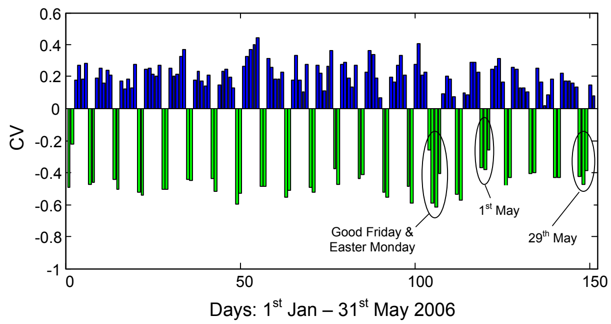
18 and 26-36 in the variable space. This change matches the difference of samples (see figure 4(a)). Figure 5(b) indicates this method can obtain a promising discrimination. A negative bar means that this day is a non-working day. Typically these are in pairs corresponding to weekends. However, in a few highlighted examples, the effect of additional UK public or bank holiday can be seen. These highlighted examples correspond to the Easter period (Good Friday/Easter Monday) and the UK public holidays held on the first and last Monday in the month of May.

Given a model built using data from January to May, we can now perform validation by comparing model predictions against real data for dates ranging from June to December inclusively. The results show that there is an above 99% positive predictive value in this case. The predicted behaviour is shown in figure 6 with daily data labelled by day of year.

To explore the details of abnormalities in the predictions, three sections are identified and labelled in figure 6. These sections are selected because they contain working days which are classified as non-working days but do not correspond to weekend days. Label A is used to refer to data for days $237-241$, label B for days $237-241$ and label C for days $305-309$. These three highlighted abnormalities are analysed in detail and plotted in figures 7-9 respectively.




(a) Canonical weights




(b) Canonical variates

Fig. 5. The Canonical Analysis of workdays and weekends, including public holidays, from January to May inclusive

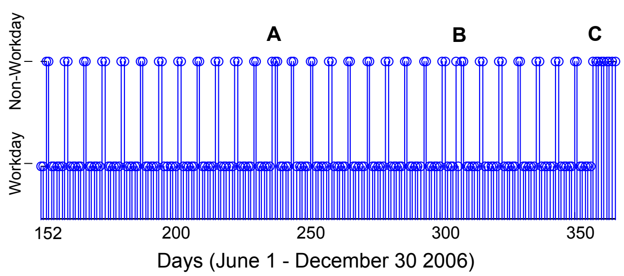
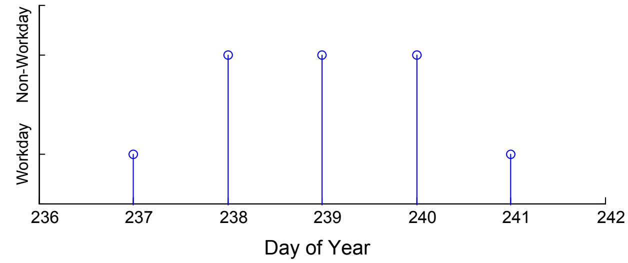



Fig. 6. The test result for the data from June 01 - December 30 2006

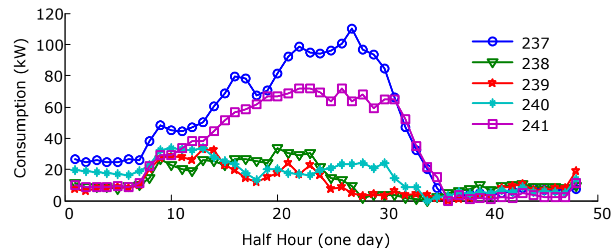
Figure 7 shows the classification results and corresponding daily energy consumption data for section A from figure 6. Classifications, given in figure 7(a), show that three consecutive days are classified as non-working days. This corresponds to the UK Bank Holiday held on August 28th (day 240). The daily energy consumption profile for this day is similar to that for a weekend day as shown in figure 7(b).

Figure 8 corresponds to section B in figure 6. The classification results are shown in figure 8(a). Day 306 is classified by the system as a non-working day but is in fact a working day (Tuesday) as is confirmed by the energy consumption series shown in figure 8(b). This represents a single mistaken classification by the system.

Figure 9 shows the half hourly consumption data corresponding to the time period denoted by label C in figure 6. This is the Christmas holiday period of 2006, for which the site was closed for days 357-360. The classifications and original daily energy computation are shown in figures 9(a) and 9(b) respectively.

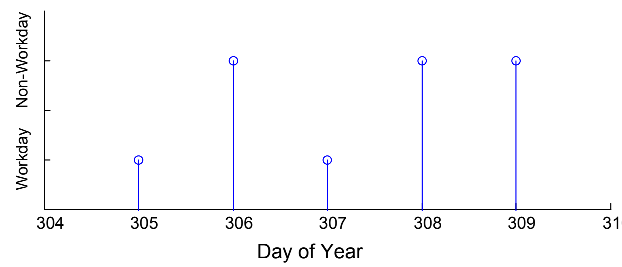



(a) Classifications

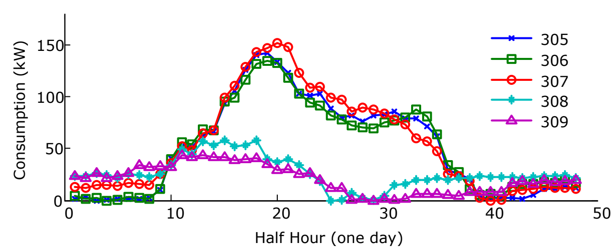



(b) Daily energy consumption

Fig. 7. The classification results for the data at section A.




(a) Classifications

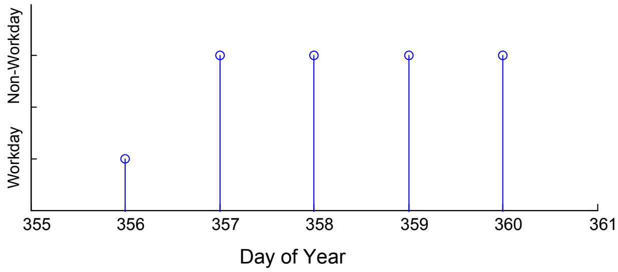



(b) Daily energy consumption

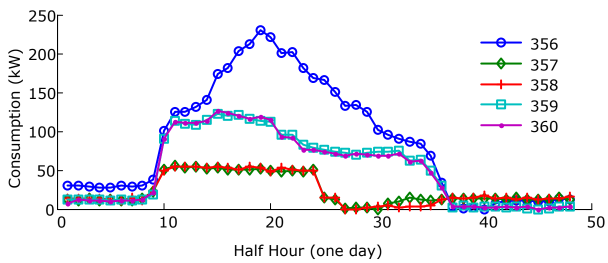
Fig. 8. The classification results for the data at section B.

## IV. Conclusion

In this paper, a new method for classification of building energy consumption data is proposed. First, the features of the daily energy consumption profiles are extracted by statistical methods and an autoregression model applied. Outlier detection methods are then used to isolate abnormal data points from further analysis. Canonical Variate Analysis combined with Linear Discriminate Analysis is then used to build a classification model based on historical data. As a result, abnormal energy consumption can be identified in real time, enabling building managers to investigate and correct problems as they occur. Of course, identifying and removing causes of abnormal energy use ensures a more efficient environment, and not just




(a) Classifications




(b) Daily energy consumption

Fig. 9. The classification results for the data at section C.

in terms of the building energy costs. In this system, the algorithms applied are computationally efficient and robust, so can feasibly be integrated into existing building energy management and warning systems. Further work will aim to build on this classification technique to provide additional tools for automated analysis of metered building energy data.

## References


[1] L. Prez-Lombard, J. Ortiz & C. Pout, A review on buildings energy consumption information, Energy and Buildings, 2008;40(3):394-398

[2] K. Blok, Improving energy efficiency by 5% and more per year. Journal of Industrial Ecology 2004;8(4):87-99.

[3] K. Blok, D.de Jager & C. Hendriks, Economic Evaluation of Sectoral Emission Reduction Objectives for Climate Change, ECOFYS Energy and Environment, AEA Technology, National Technical University of Athens, Utrecht, March 2001

[4] S. Lechtenboehmer & S. Thomas, The mid-term potential for demand-side energy efficiency in the EU. Wuppertal Institute for Climate, Environment and Energy, 2005.

[5] H. Doukas, K. Patlitzianas, K. Iatropoulos & J. Psarras, Intelligent building energy management system using rule sets. Building and Environment 2007;42(10):3562-9.

[6] M.R. Amin-Naseri, A.R. Soroush, Combined use of unsupervised and supervised learning for daily peak load forecasting, Energy Conversion and Management, 2008;49(6):1302-1308

[7] S. Tzafestas & E. Tzafestas, Computational Intelligence Techniques for Short-Term Electric Load Forecasting. J Intell Robotic Syst 2001;31(7):7-68.

[8] K. Metaxiotis, A. Kagiannas, D. Askounis & J. Psarras, Artificial intelligence in short term electric load forecasting: a state-of-the-art survey for the researcher. Energy Convers Manage 2003;44: 1525-34.

[9] H.S. Hippert, C.E. Pedreira & R.C. Souza, Neural networks for short-term load forecasting: a review and evaluation. IEEE Trans Power Syst 2001;16(1):44-55.

[10] N. Amjady, Short-term hourly load forecasting using time-series modeling with peak load estimation capability. IEEE Trans Power Syst 2001;16(4):798-805.

[11] T. Räsänen, J. Ruuskanena & M. Kolehmainen, Reducing energy consumption by using self-organizing maps to create more personalized electricity use information, Applied Energy, 2008;85(9):830-840.

[12] A.H. Neto, F.A.S. Fiorelli, Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption, Energy and Buildings, 2008;40(12):2169-2176.

[13] H. Doukas, C. Nychtis & J. Psarras, Assessing energy-saving measures in buildings through an intelligent decision support model, Building and Environment, 2009;44(2):290-298.

[14] J. E. Seem, Pattern recognition algorithm for determining days of the week with similar energy consumption profiles, Energy and Buildings, 2005;37(2):127-139.

[15] J. E. Seem, Using intelligent data analysis to detect abnormal energy consumption in buildings, Energy and Buildings, 2007;39(1):52-58.

[16] A. Neumaier & T. Schneider, 2001: Estimation of parameters and eigenmodes of multivariate autoregressive models. ACM Trans. Math. Softw., vol.27, pp.27-57.

[17] T. Schneider & A. Neumaier, 2001: Algorithm 808: ARfit - A Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. ACM Trans. Math. Softw., vol.27, pp.58-65.

[18] V. Barnett & T. Lewis, Outliers in Statistical Data, 3rd ed., John Wiley & Sons, New York, 1994.

[19] R.A. Johnson & D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, 1992.

[20] L. Nrgaard, R. Bro, F. Westad & S.B. Engelsen, A modification of canonical variates analysis to handle highly collinear multivariate data. J. Chemometr. 2006;20(8-10):425-435.

[21] T. Canals, J.R. Riba, R. Cantero, J. Cansino, D. Domingo & H. Iturriaga, Characterization of paper finishes by use of infrared spectroscopy in combination with canonical variate analysis, Talanta 2008;77(2):751-757.

[22] W. J. Krzanowski, Principles of Multivariate Analysis. A User's Perspective, Oxford University Press, 2000.


**Xiaoli Li** received B.S.E. and M.S.E. degrees from Kun-ming University of Science and Technology, and a Ph.D degree from Harbin Institute of Technology, China, in 1992, 1995, and 1998, respectively, all in mechanical engineering. From April 1998 to Oct. 2003, he was a Research Fellow of the Department of Manufacturing Engineering, City University of Hong Kong, of the Alexander von Humboldt Foundation at the Institute for Production Engineering and Machine Tools, Hannover University, Germany, a Post doc fellow at the Dept. of Automation & Computer Aided Engg, Chinese University of Hong Kong. In 2002, he was appointed professor at the Institute of Electrical Engineering, Yanshan University, China. Currently, he works in CERCIA, School of Computer Science, The University of Birmingham, UK. His main areas of research are computation intelligence, monitoring, signal processing and data analysis, manufacturing system.

**Chris P. Bowers** received a BSc degree in Computer Science and Physics from Keele University in 2001, MSc in Natural Computation in 2002 and PhD in Computer Science in 2006, both from the University of Birmingham. He is currently a Research Fellow with the Centre of Excellence in Computational Intelligence and Applications (CERCIA) at the School of Computer Science at the University of Birmingham. His main research interests involve applications of computational intelligence and robust optimisation particularly in the context of complex and dynamic environments.

**Thorsten Schnier** has been a Research Fellow with the University of Birmingham since 1999, and Centre of Excellence for Computational Intelligence and Applications (CERCIA) since its inception in 2003. Before coming to Birmingham, Dr Schnier read Electrical Engineering (BSc, MSc) at the University of Hannover and Computer Science (PhD) at the University of Sydney. His Ph.D research at the Key Centre for Design Computing involved work on Computational Creativity, Evolutionary Design, and particularly the role of representations. Following his PhD he worked as a lecturer at the Key Centre. He was Principle Investigator of the 'Nature Inspired Creative Design' network, an AHRC/EPSRC sponsored 'Designing for the 21st Century' initiative. He is Principle Investigator on a project under the EPSRC 'Feasibility Studies in Energy' program on "Advanced Analysis of Building Energy Performance", and supervisor on an UKERC Doctoral Training Grant on "Advanced Computational Techniques for Modeling Electricity Markets".