Anomaly Detection and Visualization of School Electricity Consumption Data

Wenqiang Cui
Qunar.com Information Technology Co. Ltd.
100080 Beijing, China
e-mail: wenqiang903@gmail.com

Hao Wang
Norwegian University of Science & Technology
Postboks 1517, N-6025 Aalesund, Norway
e-mail: hawa@ntnu.no

***Abstract*—Anomaly detection has been widely used in a variety of research and application domains, such as network intrusion detection, insurance/credit card fraud detection, health-care informatics, industrial damage detection, image processing and novel topic detection in text mining. In this paper, we focus on remote facilities management that identifies anomalous events in buildings by detecting anomalies in building energy data. We have investigated five models to detect anomalies in the school electricity consumption data. Furthermore, we propose a hybrid model which combines polynomial regression and Gaussian distribution. Based on this model, we have developed a data detection and visualization system for a facilities management company to detect anomalous events in school electricity facilities. The system is tested and evaluated by the facilities managers of the company. According to the result of the evaluation, it reduces the effort required by facilities managers to identify anomalous events in school electricity facilities.**

Keywords-data analysis;anomaly detection; data visualization; facilities management; time-series; school electricity consumption data

I. INTRODUCTION

In recent years, with the ever growing shortage of natural resources, energy has been a major political, social and economic topic. It is now widely accepted that conserving energy and reducing energy consumption is of paramount importance. In the UK, building energy consumption has increased at a rate of 0.5% per annum, which is approximately 40% of total energy consumption [1]. However, buildings are widely reported to utilize energy inefficiently [2].

Although many techniques have been extensively investigated for building energy consumption modelling to design low-energy building, buildings often exceed the energy savings promised by their design. Anomalous events, such as faults in lighting equipment, can account for 2-11% of the total energy consumption for commercial buildings [3]. To identify anomalous events in buildings in time, increased attention has been given to remote facilities management.

Anomaly detection in building energy data is one of the most important methods to identify anomalous events in buildings. In different application domains, each anomaly detection problem has distinct features such as the nature of the data, availability of labelled data and type of anomalies to be detected. Most of the existing anomaly detection techniques solve a specific formulation of the problem [4]. Applying concepts from different disciplines such as statistics, machine learning, and information theory to a specific problem formulation is a challenge in anomaly detection.

In this paper, we present an innovative method and build a system to detect and visualize anomalous events in the school electricity consumption data for a facilities management company. This company provides facilities management service for over 40 schools in Scotland. Data on electricity consumption of these schools is recorded by their half hourly metering system. Every week facilities managers look over spreadsheet graphs of the data to identify anomalous events, particularly unusually high electricity consumption. Our system is used to reduce this tedious and time-consuming eyeballing of the data.

The goal of this paper is to build a data detection and visualization system to improve facilities managers' performance when they detect anomalous events in school electricity facilities. The paper is organized as follows. The methods of anomaly detection and data visualization are described in section II and III respectively. Then, section III shows the data detection and visualization system for school electricity consumption data. Finally, we present conclusions and discussion, and suggestions for future research in section IV and V.

II. ANOMALY DETECTION

Anomaly detection, also referred to as outlier detection, is the process of detecting patterns in a given data set that do not conform to an established normal behaviour [5]. Anomalies in data can correspond to some significant information in many diverse research areas and application domains. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination [6].

According to what kind of labels are available for a data set, there are three types of anomaly detection techniques:

- Supervised techniques build models for both anomalous data and normal data. An unseen data instance can be classified as normal or anomaly by comparing which model it belongs to.
- Semi-supervised techniques only build a model for normal data in the training data set. An unseen data instance can be classified as normal if it can fit the model sufficiently well. Otherwise, the data instance will be classified as anomalies.

- Unsupervised techniques do not need any training data. These approaches are based on the assumption that anomalies are much rarer than normal data in the data set.

The nonconforming patterns in data are commonly called anomalies or outliers. They can be classified into three following categories [4]:

- Point Anomalies: A point anomaly is a single independent data instance which does not conform to a well defined normal behaviour in a data set.
- Contextual Anomalies: A contextual anomaly is a data instance which is considered as anomalies in a specific context, but not otherwise.
- Collective Anomalies: A collective anomaly is a collection of related data instances which are anomalous with respect to an entire data set.

In this paper, semi-supervised anomaly detection is adopted to detect anomalies in school electricity consumption data, and point and collective anomalies are the target of the detection. Because school electricity consumption data is in the form of a time series, two types of model: regression based model and kernel density estimation based model are investigated. Several existing methods, such as Yule-Walker, Akaike/Bayesian Information Criterion, Maximum Likelihood are adopted to estimate parameters for the models. This process is omitted in this paper.

A. Time Series

A time series is a sequence of data points that are ordered by a uniform time interval [7]. The school electricity consumption data used in this paper is in the form of a time series where the time interval is half an hour. It is recorded by a facilities management company. With half hourly frequency, the time series for one week contains 336 data instances. Fig. 1 shows one week's electricity consumption data of Kennoway Primary School from Monday to Sunday. In the Fig. 1, each data point represents the electricity consumption in the previously half hour.

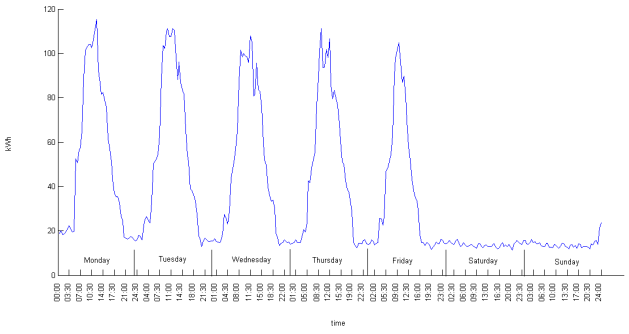

Figure 1. One week's electricity consumption data of Kennoway Primary School. The x axis is the electricity consumption (kWh). The y axis is the time.

With the school electricity consumption data, two types of anomalies are significant for facilities managers:

- One single high data point anomaly is often used to identify an anomalous meter. Because it is usually caused by a meter that records a wrong reading.
- A collection of continuous high readings is used to identity anomalous electricity facilities, such as heating being on at the wrong time.

The models used to detect anomalies in the data are investigated in the next section. Autoregressive (AR) model and Autoregressive-moving-average (ARMA) model are widely used in time series analysis [8]. However, they cannot capture the temporal structure of the school electricity consumption data because the data fluctuates greatly. Therefore, the investigation of these two models is omitted in this paper.

B. Polynomial Regression Model

In order to avoid the volatility and capture the temporal structure of the school electricity consumption data, polynomial regression is investigated to model the variation of the data for each time slot through a year, since the data of a time slot in a week should fluctuate slightly associated with the number of the week through one year.

Given a data set $\{(X_t, Y_t), t = 1, \cdots, T\}$where $X_t$is a scalar variable, the polynomial regression model is defined as:

$$Y_t = c + \sum_{i=1}^{p} \alpha_i X_t^p + \varepsilon_t$$

where c is a constant, ε_t is a random error conditioned on X_t, $\alpha_i$is the parameters of the model, p is the order of the model.

The order of the polynomial regression model used in this paper is 11. Furthermore, to improve the numerical properties of the polynomial regression model, X is centred and scaled by:

$$\hat{X} = \frac{X - \mu}{\sigma}$$

where μ is the mean of X, σ is the standard deviation of X.

This model is fitted to electricity consumption data of one time slot through a year. In one year, there are 52 data instances of each time slot. Fig. 2 shows the polynomial regression model which is fitted to one school's electricity consumption data for the first time slot of the year 2011.In the Fig. 2, the blue stars are the electricity consumption data and the red line is the polynomial regression model which is fitted to the data.

To predict one week's electricity consumption data, each polynomial regression model is used to predict one data instance of one time slot. Fig. 3 shows the predicted values of a week's electricity consumption data, in which the green line is the data of the year 2011 and the red line is the predicted values. In the Fig. 3, although the model does not predict the values of weekday sufficiently well, the prediction of Saturday and Sunday are accurate.

To detect anomalies in the data, the residual of each time slot is calculated. For one time slot, the residual is the difference between the actual value and the predicted value for this time slot. If the residual is greater than a threshold, the data instance will be detected as an anomaly. The threshold is computed for each school's model respectively based on two labelled datasets. One is used to train the model. Then the threshold is set and adjusted to guarantee that all the anomalies in the Saturday and Sunday's data can be

detected in another dataset based on the model. Fig. 4 shows the process of detecting anomalies in Saturday and Sunday's data by the polynomial regression model. In the Fig.4, the green line is the data of the year 2011 which is used to train the model. The blue line is the data of the year 2012. The red line is the predicted values. The black line is the residual between the data of the year 2012 and the predicted values. The yellow line is y=0, which is the threshold. Since the data of Saturday and Sunday in 2012 are much higher than that in 2011, the data of Saturday and Sunday in 2012 are anomalous. These anomalies can be detected by the residual. In Saturday and Sunday, if the residual of a data instance is greater than 0, the data instance will be detected as an anomaly. However, the residual of weekdays always fluctuates around 0. If we use the residual to detect the anomalies in a weekday, many normal data instances will be detected as anomalies. Therefore, this model cannot be used to detect the anomalies in a weekday.

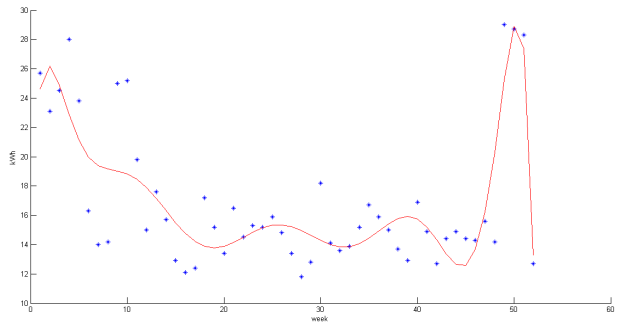

Figure 2.Polynomial regression model fitted to one school's electricity consumption data for the first time slot through the year 2011. The x axis is the week number from 1 to 52. The y axis is the electricity consumption data (kWh).

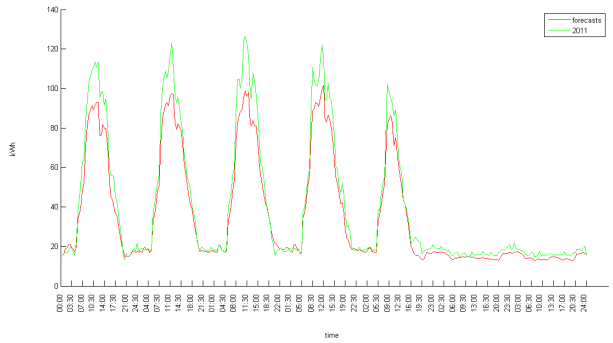

Figure 3.Predicted values a week.

C. *Kernel Density Estimation*

An alternate approach to model the data in the weekday is to fit a distribution to the data instances from one time slot for a year. For each time slot, the data instance with a very low probability can be considered as an anomaly. Based on the features of the data, kernel density estimation (KDE) is used to estimate the probability density function of the data in the weekday.

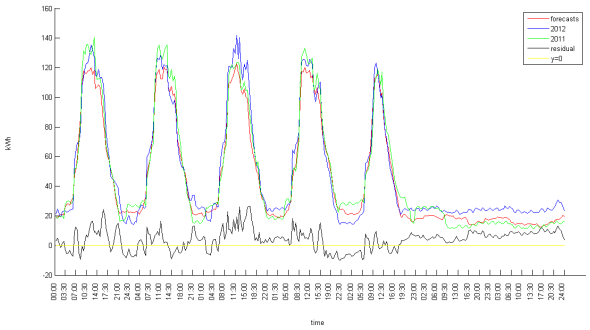

Figure 4. Detecting anomalies in Saturday and Sunday.

Given a data set $x_n = \{x_1, \cdots, x_n\}$ from an unknown continuous probability density function f, the Gaussian kernel density estimation is defined as:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-x_i}{h})^2}$$

where $K = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-x_i}{h})^2}$ is the Gaussian kernel function, h is the bandwidth, nh is the number of the data instances in the data set.

One school's electricity consumption data of the year 2011 is used to estimate the Gaussian kernel distribution for each time slot. Fig. 5 shows the Gaussian kernel distribution of the first time slot's data, in which the red '+' is the electricity consumption data and the blue line is the Gaussian kernel distribution of the data.

To detect anomalies in weekly data, the probability of each data instance is calculated by the Gaussian kernel distribution of the corresponding time slot. If the probability of a data instance is less than a threshold of the time slot, the data instance will be considered as an anomaly. With the Gaussian kernel distribution of one time slot, the threshold for the time slot is given by:

$$\text{threshold} = P_{min} + \alpha\times(P_{max} - P_{min})$$

where P_{min} is the minimum probability of the Gaussian kernel distribution, P_{max} is the maximum probability of the Gaussian kernel distribution, α is a coefficient. The process of adjusting the coefficient α for different schools' data is also based on two labelled datasets. One is used to train the model. Another is used to calculate the coefficient which guarantees all anomalies are detected in the dataset.

Fig. 6 shows the detected anomalies in weekly data by the Gaussian kernel distribution model. In the Fig. 6, the green line is the data of the year 2011 which is used to estimate the distribution. The blue line is the data of the year 2012. The red '*' is the detected anomalies in the data of the year 2012. Comparing the data of the year 2011 and 2012, the model has detected all the anomalies in the Monday, Friday, Saturday and Sunday's data. However, the Gaussian kernel distribution model also comes with the over-fitting problem which is shown in the Fig. 7. In the Fig.7, the green

line is the data of the year 2011. The blue line is the data of the year 2012. The red '*' is the detected anomalies.

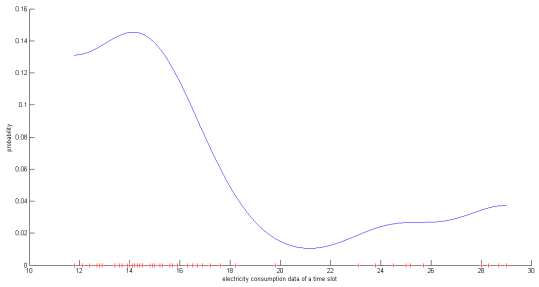

Figure 5.Gaussian kernel distribution of the first time slot's electricity consumption data. The x axis is the electricity consumption data. The y axis is the probability.

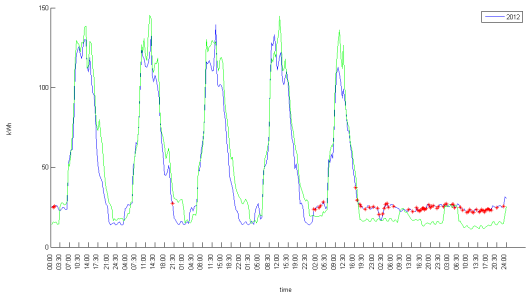

Figure 6.Anomaly detection in weekly data by the Gaussian kernel distribution based model.

Comparing the data of the year 2011 and 2012, the data of the year 2012 do not contain any anomalies. However, many normal data instances have been detected as anomalies.

D. Gaussian Distribution

To avoid the over-fitting problem of the Gaussian kernel distribution, Gaussian distribution is adopted. With Gaussian distribution, for each time slot, if a data instance is greater than an upper bound, it will be considered as an anomaly. The probability density function of the Gaussian distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the mean of the data,$\sigma^2$is the variance of the data.

Fig. 8 shows the estimated Gaussian distribution of one school's electricity consumption data for the first time slot of the year 2011. In the Fig.8, the red '+' is the electricity consumption data. The blue line is the Gaussian distribution of the data. To detect anomalies, the upper bound of one time slot is defined as:

$$\text{upperBound} = \mu + \alpha \times \sigma$$

where μ is the mean of the Gaussian distribution, σ is the standard deviation of the Gaussian distribution, and α is a coefficient. The process of adjusting the coefficient α is the same as previously described in section C.

Fig. 9 shows the process of detecting anomalies in the data by Gaussian distribution based model. In the Fig. 9, the green line is the data of the year 2011 which is used to fit the distribution. The blue line is the data of the year 2012. Comparing the data of the year 2011 and 2012, the model has detected the anomalies in Monday and Thursday of the year 2012. The thick part in the blue line is the detected anomalies. However, it does not detect the anomalies in Saturday and Sunday sufficiently well.

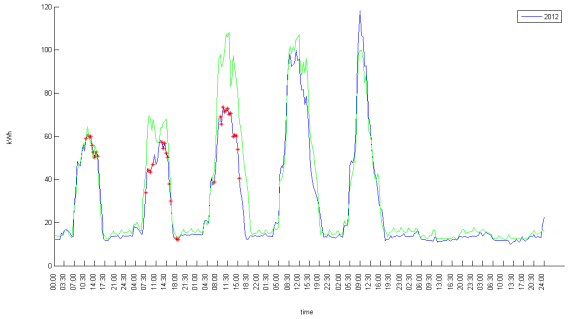

Figure 7.Over-fitting problem of the Gaussian kernel distribution based model.

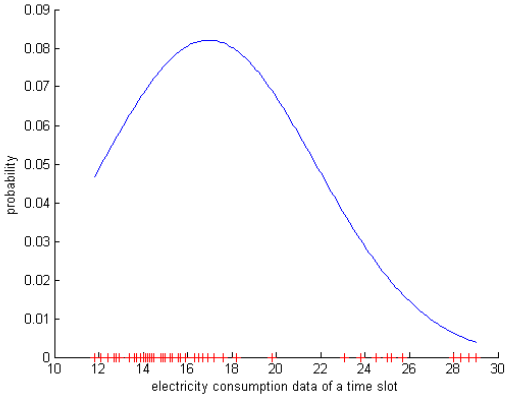

Figure 8.Gaussian distribution of the first time slot's electricity consumption data.

The Gaussian distribution based model can avoid over-fitting problem that appeared in Gaussian Kernel distribution based model. Figure 10 shows using Gaussian distribution based model to detect anomalies in the weekly data which only contains normal data. In the figure 10, the green line is the data of the year 2011. The blue line is the data of the year 2012. The red line is the upper bound. It shows that all normal data of the year 2012 are less than the upper bound.

E. Model Selection

The polynomial regression model detects the anomalies in the data of Saturday and Sunday. But it is not suitable for the data of weekdays. The Gaussian kernel distribution based

model detects all anomalies in the data. However, it results in a high false positive which is caused by the over-fitting problem. The Gaussian distribution based model detects the anomalies in the data of weekdays without the over-fitting problem. Therefore, in order to detect anomalies in weekly data $x_t(t = 1, \cdots, 336)$ with a low false negative and a high precision, a hybrid model which combines polynomial regression and Gaussian distribution is proposed:

$$\begin{cases} f(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, upperBound = \mu + \alpha \times \sigma \\ \qquad if\ 1 \le t \le 240 \\ x_t = \sum_{i=1}^{11} \beta_i t^i + \epsilon,\ \ threshold = c \\ \qquad if\ 241 \le t \le 336 \end{cases}$$

where the coefficient α and threshold c are adjusted according to different schools' electricity consumption data as mentioned in section B and D.

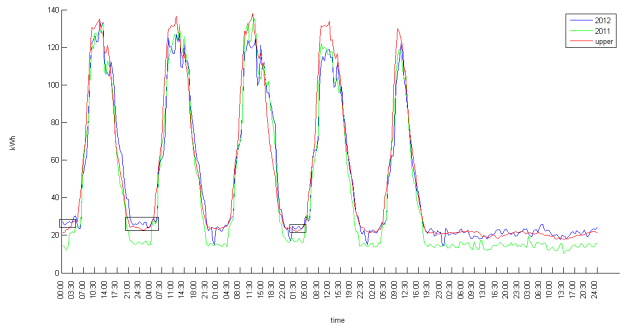

Figure 9.Anomaly detection based on Gaussian distribution model. The x axis is the time. The y axis is the electricity consumption data.

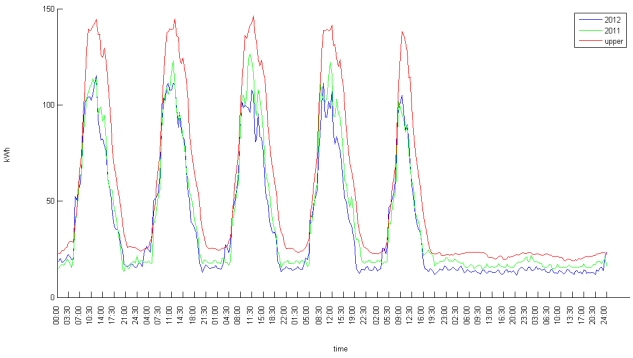

Figure 10.Anomaly detection on normal weekly data based on Gaussian distribution based model.

III. Visualization

A cluster heat map is a graphical representation of data where the individual values contained in a matrix are represented as colours taken from a colour scale [9]. To visualize the school electricity consumption data in a cluster heat map, each row is daily electricity consumption data and the colour of each cell is changed according to the data in which green represents low consumption, yellow represents medium consumption and red represents high consumption. An example of the heat map is shown in the Fig.12 in the next section.

F. Data Detection and Visualisation System

Based on the model proposed in the section E, we developed a data detection and visualization system for school electricity consumption data. In the system, the weekly data is visualized in a line chart and a heat map, in which anomalies are marked out. Fig. 11 shows the line chart of one week's electricity consumption data of the Musselburgh Grammar school. In the Fig. 11, the blue line is the data of the year 2012. The green line is the data of the year 2011. The red points are the detected anomalies in the data of the year 2012. Figure 12 shows the heat map of one week's electricity consumption data of the Musselburgh Grammar school.

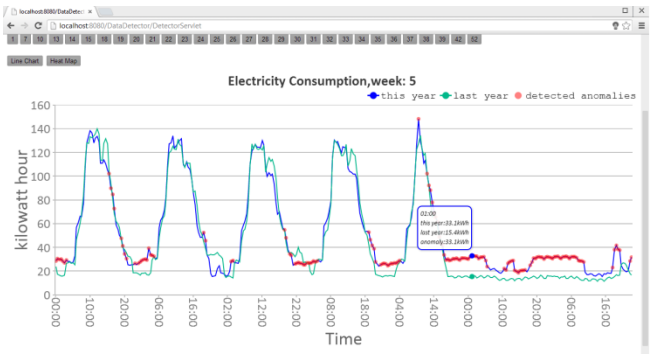

Figure 11.The line chart of one week's electricity consumption data of the Musselburgh Grammar school.

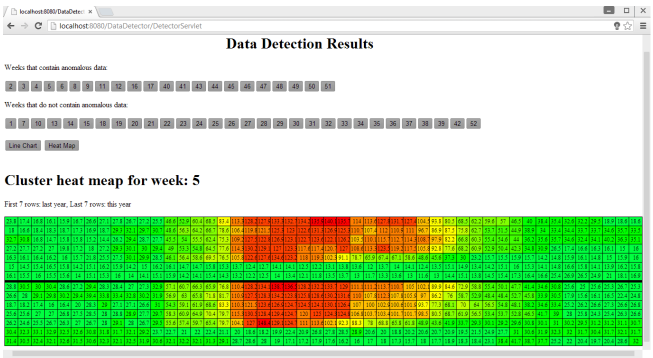

Figure 12.The heat map of one week's electricity consumption data of the Musselburgh Grammar school.

IV. Conclusion

In this paper, we have examined techniques to detect anomalies in a time series. The first, polynomial regression model is used for model the variations through the year. This model detects anomalies only in the data of Saturday and Sunday. We also examined two methods that estimate a distribution for the data of each time slot. The Gaussian kernel distribution based model detects all anomalies in weekly data. But it results in a high false positive which is caused by over-fitting problem. The Gaussian distribution

based model detects the anomalies in the data of weekdays without over-fitting problem.

To detect anomalies in the entire weekly data, we proposed a hybrid model which combines polynomial regression and Gaussian distribution. Moreover, based on this hybrid model, we created a data detection and visualization system for school electricity consumption data.

V. Future Research

Future research will focus on the data detection and the improvement of the system, especially, the automation of model fitting for different schools' electricity consumption data. Currently, the model used in the system need to be adjusted manually before detecting anomalies. Furthermore, we are working on the model optimisation based on more data (e.g., data of 10 years), such as adding season effect to the model. Finally, effort is put in the investigation of visualization techniques and the design of the user interface of the system.

Acknowledgments

The authors wish to thank prof. Nigel Goddard of the University of Edinburgh for the discussion on modelling, and the managers of the facilities management company for the discussion on the system.

References

[1] L. Perez-Lombard, J. Ortiz, and C. Pout, “A review on buildingsenergyconsumption information,”Energy and buildings, vol. 40, no. 3, pp.394–398, 2008.

[2] M. M. Ardehali, T. F. Smith, J. M. House, and C. J. Klaassen, “4641building energy use and control problems: An assessment of case studies,”ASHRAE Transactions-American Society of Heating RefrigeratingAirconditioning Engin, vol. 109, no. 2, pp. 111–121, 2003.

[3] Y. Heo, R. Choudhary, and G. Augenbroe, “Calibration of buildingenergy models for retrofit analysis under uncertainty,”Energy andBuildings, vol. 47, pp. 550–560, 2012.

[4] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: Asurvey,”ACM Computing Surveys (CSUR), vol. 41, no. 3, p. 15, 2009.

[5] B. Abraham and A. Chuang, “Outlier detection and time series modeling,”Technometrics, vol. 31, no. 2, pp. 241–248, 1989.

[6] V. Kumar, “Parallel and distributed computing for cybersecurity,”Distributed Systems Online, IEEE, vol. 6, no. 10, 2005.

[7] R. A. Yaffee and M. McGee,An Introduction to Time Series Analysisand Forecasting: With Applications of SAS®and SPSS®.AccessOnline via Elsevier, 2000.

[8] C. Chatfield,The analysis of time series: an introduction.CRC press,2003.

[9] L. Wilkinson and M. Friendly, “The history of the cluster heat map,”The American Statistician, vol. 63, no. 2, 2009.