Algorithmic Cleansing of Metered Building Performance Data

Stephan Hoerster, s.c.hoerster@umail.ucc.ie
Informatics Research Unit for Sustainable Engineering, University College Cork, Cork, Ireland

Thomas Willwacher, thomas.willwacher@math.uzh.ch
Institute of Mathematics, University of Zurich, Switzerland

Karsten Menzel, k.menzel@ucc.ie
Informatics Research Unit for Sustainable Engineering, University College Cork, Cork, Ireland

Abstract
Since the introduction of smart meters, a huge demand for the evaluation of energy consumption data can be experienced. Unfortunately, metered load curve data is often inconsistent and not free of faults. This results in increased cleansing efforts in order to evaluate and analyse the data. The interest in deeper analysis is often driven by monetary reasons or contractual agreements with external facility operators.
This research aims to eliminate the need for manual cleansing by introducing an algorithm which can be applied automatically on metered gas consumption data. Through a combination of gap detection and outlier identification, faulty data gets flagged. These occurrences get interpolated through a correlation between building heat consumption and outside temperature.
The proposed algorithm is evaluated with data sets from different buildings.
Its results reveal a huge increase in interpretability of the data.

Keywords: Data cleansing, load curve analysis, consumption forecast

1 Introduction

The recent and ongoing roll out of smart meters has increased the consumption data available from a building by a multitude. Load curves are data series captured by energy meters in defined time intervals. It is widely believed that the analysis of load curves has positive impact on daily operations, running systems and results in reduced energy costs (Li 2005).
At the present time, newly constructed buildings often get equipped with smart meters and depending on their size with a BMS (Building Management System). Older buildings are being retrofitted with smart meters to increase their detail of monitoring. In both cases, the captured data is usually recorded with data loggers. These devices usually speak a variety of field bus protocols. Their flexibility allows the communication with most kind of meters and sensors that can be found in buildings. Data loggers consolidate all monitored data from a building and transfer it to a database where it can be accessed for analysis purposes. Data acquisition techniques are further discussed in (Hoerster et al 2014).
With the introduction of high monitoring density, the likelihood of flawed data has increased accordingly. Reasons for bad or missing values are varied. These could be e.g. malfunctioning meters, bad wiring, errors while transmitting the data, or even power outages.

Nowadays, utility providers and Facility Management companies manage many thousand buildings. For them, reliable data is crucial for billing, analysis and optimization. For efficiency reasons, the manual correction of data needs to be minimized.

This research introduces a methodology to automatically cleanse gas consumption data. Since the gas consumption of a building strongly relates to its outside temperature, it can be used to interpolate missing or bad readings.

However, this equation is not fully applicable to cleanse electricity data since lighting and plug load is not related to the outside temperature. It may be used for electricity if a majority of the consumption relates to heating or cooling. However this is not the case for the buildings considered in this paper. Therefore, we will focus only on gas consumption data to demonstrate the cleansing methodology.

Weather data usually can be acquired from nearby weather stations. Depending on the size of buildings, it might even be feasible to install a local weather station. In this work we use weather data from both public and locally installed weather station.

In this research, a sophisticated interpolation algorithm combined with robust outlier detection is introduced. Its application validates and cleanses erroneous gas consumption. It is envisaged to implement this functionality in a monitoring system as automated data correction without any user interference. In order to prevent user confusion or doubt, any corrected data gets flagged as interpolated.

As the proposed methodology interpolates bad or missing data from a load curve through corresponding temperature readings, it may be exploited to forecast gas consumption for the same period where forecast weather data is available.

2 Related research

Our work emphasizes on cleansing of data. This discipline has been a topic in research for many years.

The identification of outliers and cleansing of data has been discussed widely in previous research (Abraham & Chuang 1989) (Ljung 1993). Here, regression analysis has been utilized to determine outliers in time series curves. However, their work considers time series as stationary, which renders their algorithms inappropriate for our work.

More recently, Chen et al. have introduced a smoothing method that corrects gaps and outliers in a load curve (Chen et al 2010). In comparison to our work, their algorithm is applicable to all kind of energy load curves, not just gas consumption. However, it requires a little amount of user interaction as their system needs the user to specify the ideal smoothing parameter. This renders it inapplicable to our approach as we require a system that cleanses data without user interaction.

This limitation has also been criticized in (Høverstad et al 2013), where a system is envisaged which automatically configures itself. For their work, data cleansing is performed in order to increase the accuracy of energy consumption forecasts for the next 24 hours. Their work involves several fairly complex load prediction models.

Statistical methods have also been studied to identify outliers (Davies & Gather 1993) (Ferguson 1961). Here, it is mostly assumed that the underlying distribution is known. However, load curves as real world application cannot fulfil this requirement per default.

Data mining techniques have also been utilized to identify outliers (Knox & Ng 1998) (Ramaswamy et al 2000). The downside of their approach is that these techniques are usually designed for structured data and might not work well on load curve data.

3 Methodology

In this chapter we propose an algorithm that can be applied to identify missing data and data of bad quality from gas consumption readings.

Gas is typically consumed for heating purposes. Therefore, its consumption can be related to the ambient temperature. This allows us to derive a building equation to replace erroneous data.

Faulty readings often result in large peaks in the consumption data, see figure 1 for an example. It seems impossible from this perspective to draw any meaningful conclusion with regards to the building's consumption behaviours. Note that the consumption peaks at almost 15,000 kWh.

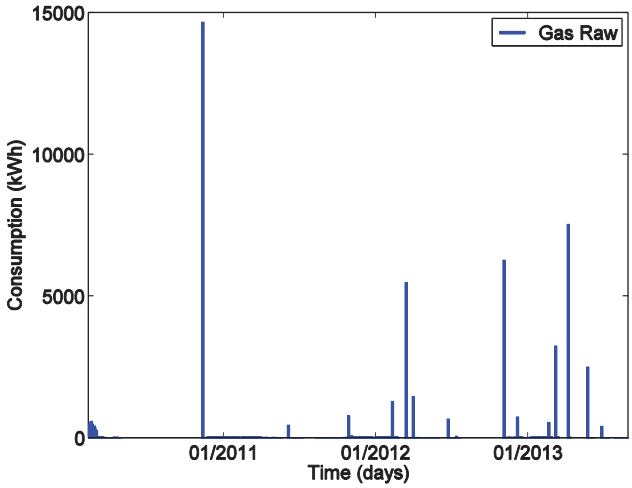

Figure 1 Raw gas consumption

While it seems that most values are 0, the largely magnified graph seen in figure 2 reveals more typical consumption patterns. The areas with zero consumption have decreased significantly, yet there are still gaps with apparently no data available. In this representation, it seems more likely that the consumption peaks less than 100 kWh. This emphasizes the impact of bad readings on graphs readability.

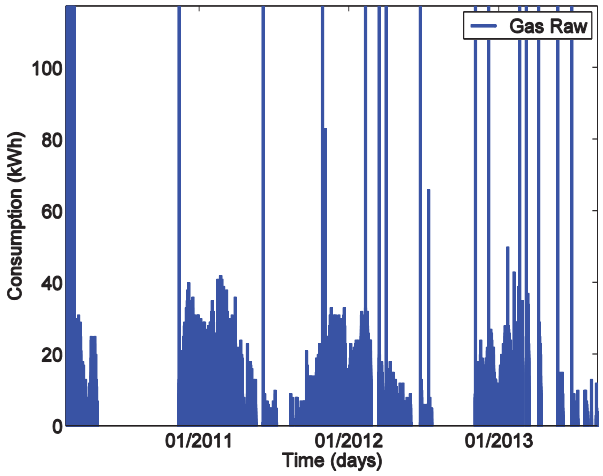

Figure 2 Magnified gas consumption

Our methodology aims to flag all occurrences of either missing data or bad data and interpolate these with replacement values. The algorithm has four input variables:

(i) gas consumption
(ii) timestamp corresponding to gas consumption
(iii) average daily temperatures
(iv) timestamp corresponding to average daily temperatures

As a first step, a regression analysis is performed on the data, by fitting a curve of the form as in equation 1.

$$W(T) = W_0 + \max(w(T_0 - T), 0) \quad (1)$$

Here, T is the outside daily average temperature, W_0 the base load, T_0 the heating threshold and w the temperature coefficient.

For example, in the curve shown in figure 3 the heating threshold T_0 is approximately 16 °C, the base load W_0 is 20 kWh and the gas consumption at T = 0 °C would be around 350 kWh.

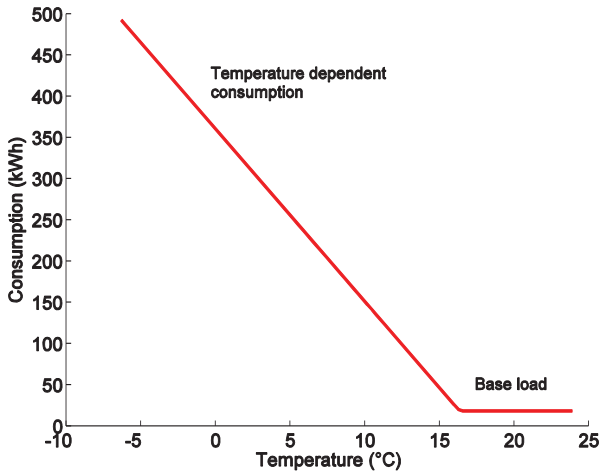

Figure 3 Ideal result of a regression analysis

Algorithmically, for fitting the above curve to the data and identifying outliers we employ a standard RANSAC scheme (Fischler & Bolles 1981). RANSAC has been proven to be very robust in a wide area of fields and applications, and works also well in our application.

In pseudo code the algorithm is as follows, where we assume that the data points (temperature and gas consumption) are $(T_j, W_j),\ j = 1, \dots, N$.

Initiate outliers: $o = \emptyset$
Do

Find W_0, T_0, w minimizing

$$\sum_{j=1}^{M} |W_j - W(T_j)| \quad (2)$$

where $j \in o$
Median Error:

$$err = median(\{|W_j - W(T_j)|\ j = 1, \dots, N\}) \quad (3)$$

Update outliers:

$$o = \{j\ |W_j - W(T_j)|\ > 8\ err\} \quad (4)$$

until convergence.

The nonlinear optimization problem in the first step of the loop is solved by the simplex algorithm (Lagarias et al).
Once the heat curve is fitted, we know (i) the parameters W_0, T_0 and w, and (ii) which data points are outliers. Next, the missing and/or faulty data points may be estimated to be $W(T)$, where T is the daily average temperature and $W(T)$ is as in equation 1.
This equation may then be used to also estimate future consumption values for the time interval provided. Here, the accuracy of the weather forecast plays a decisive role.

As an illustration, consider the data plotted in figure 4 together with the heat curve estimated by the above algorithm. The data consists of some serious outliers, resulting in a fitted heat curve that is not meaningful in its graphical representation. Here, the heat curve is represented as a straight line instead of the expected shape as depicted in figure 3.

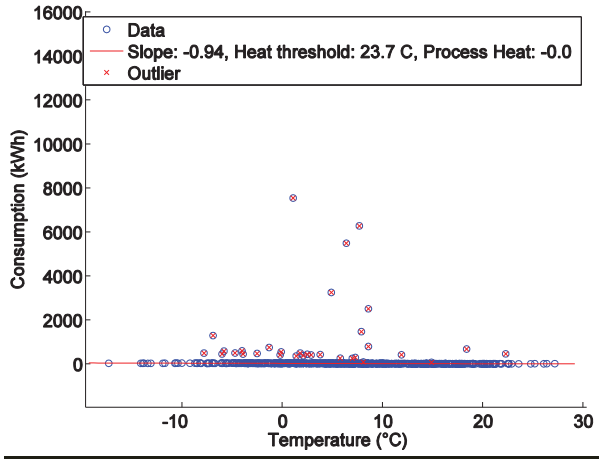

Figure 4 Scatter plot with daily consumption

In figure 5 the missing or bad data was estimated, resulting in values placed directly on the red artificial line. The output reassembles what has been illustrated in figure 5. After the cleansing process, the heat curve has again a shape as expected.

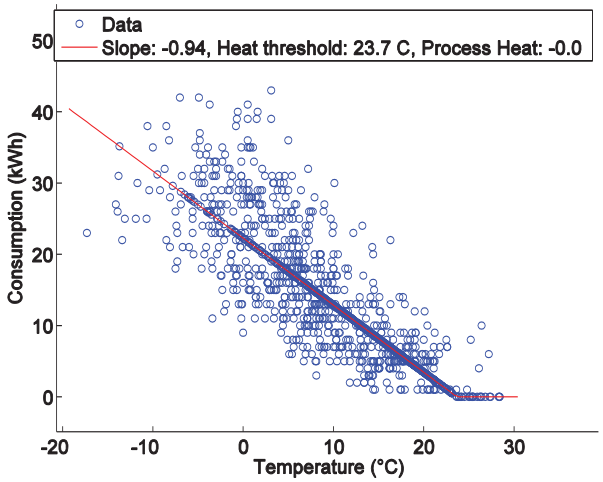

Figure 5 Scatter plot with bad readings replaced by calculated readings

The fitted heat curve can now be used to calculate W_T for each flagged data point. Afterwards, a direct comparison between raw and cleansed data can be examined. Figures

6 and 7 illustrate magnified examples of gas readings after applying the outlined methodology.

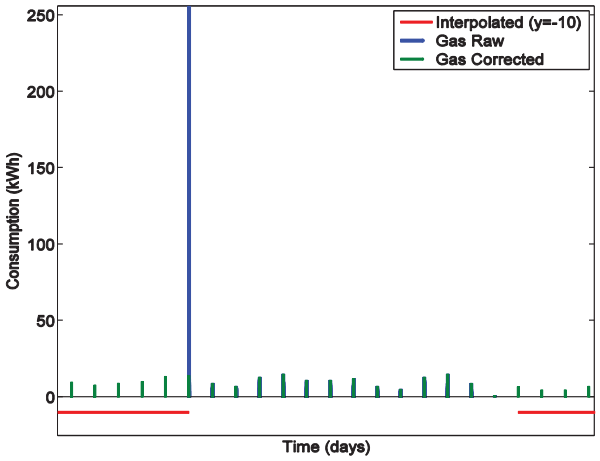

Figure 6 Magnified comparisons between raw and cleansed gas consumption

Both figures have occurrences where either missing data or outliers have been replaced with cleansed data derived from the applied methodology.
Note: In both representations, interpolated ranges are highlighted in red, placed below the x-axis at -10 for a better readability.

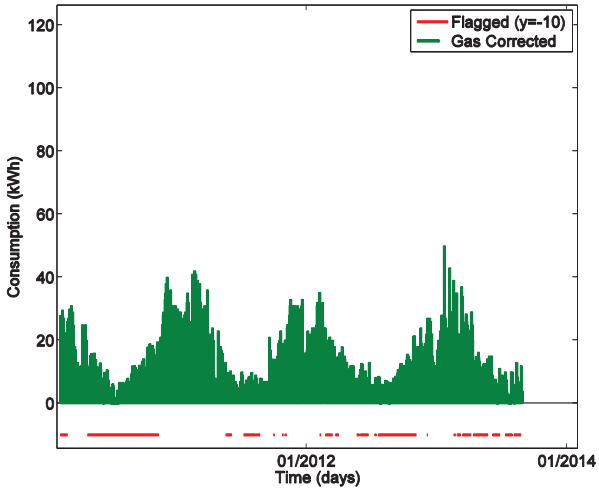

Figure 7 Cleansed data output after applying the outlined methodology

A data summary for the example used to discuss this methodology can be seen in table 1. It is interesting to note that only few outliers were responsible for the high spikes in a graphical presentation.

Table 1 Data quality summary

Criteria	Days
Period	1295
Good data	702
Missing data	560
Outliers	33
Total faulty	593

4 Results

The introduced methodology has been applied to metered readings from the following buildings:

- A multipurpose arena that hosts major sport events as well as concerts. Occasionally, business meetings take place in some of its VIP lodges.
- A public school that is only occupied from Monday to Friday.

Table 2 provides an overview about the data quality of the gas readings for the demo sites. Most historical readings are available for the school. With only 0.77% missing data, the completeness is higher than 99% which is a desirable target.
The historical data from the stadium covers a whole year. Missing data from the interval accumulates to 11.83 % which is a lot compared to the school building.

Table 2 Demo site data quality

Evaluation criteria	Arena	School
Total period (in days)	365	1425
Readings (in minutes)	60	60
Expected readings (per day)	24	24
Expected readings (in total)	8784	34272
Actual readings	7745	34007
Missed readings (in %)	11.83	0.77

The type of data defects is summarized in table 3. Here, especially for the school the percentage of poor data is much higher than what was expected when comparing with table 2. This is explained by the number of outliers, which were detected during the flagging process of the cleansing algorithm.

Table 3 Categorized data defects

Evaluation criteria	Arena	School
Total days	365	1425
Days with missing data	39	117
Outliers	0	119
Good (in %)	89.04	83.44
Bad (in %)	10.96	16.57

The data sets have been evaluated with the proposed methodology. A comparison of the data before and after correction is depicted in figures 8 and 9.

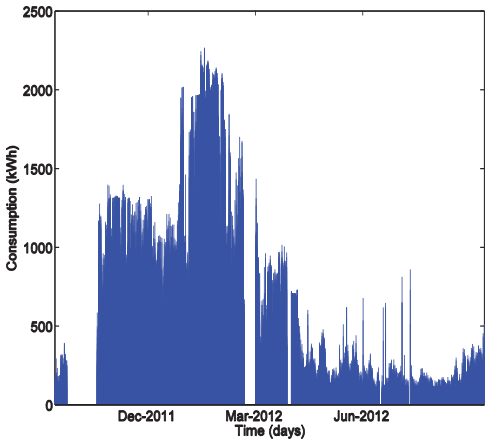

Figure 8a Arena: Raw gas consumption

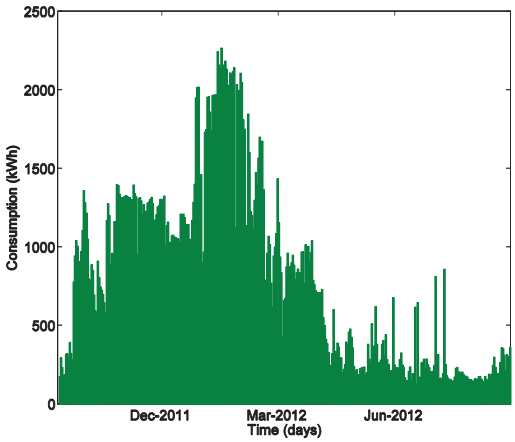

Figure 8b Arena: Corrected gas consumption

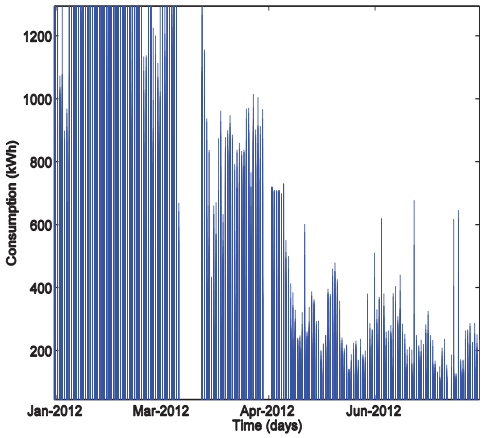

Figure 8c Arena: Detailed raw consumption

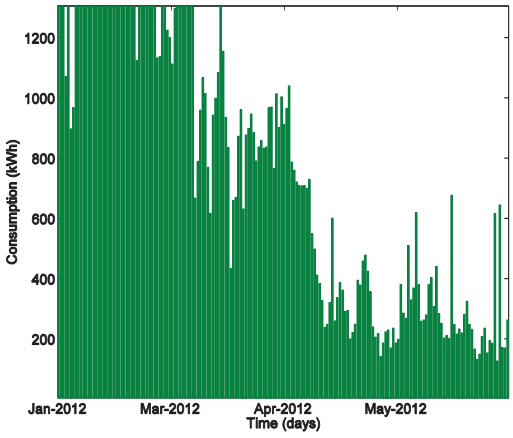

Figure 8d Arena: Detailed corrected consumption

The arena data does not suffer from any outliers. Instead, there are several gaps where data is missing. The total recorded gas consumption including the measurement gaps can be seen in figure 8a. In figure 8b, these gaps were interpolated by applying of our methodology. A more detailed interpolation of missing data can be seen in figures 8c and 8d. Here it was zoomed into a selected part of the data to increase visibility.

Figure 9a depicts all historic readings from the school. The huge amount of outliers intrigues the viewer to believe that the consumption is much higher than it actually is. What cannot be seen in this figure is that some outliers were as high as 10.000 kWh. Due to readability reasons we applied a vertical zoom on the raw data. Otherwise the load curve would appear flat. The corrected consumption retrieved through the cleansing process can be seen in figure 9b. No peaks higher than 60 kWh remain in the cleansed data. Figures 9c and 9d illustrate a magnified version of the schools consumption. In here, the cleansing of outliers can be easily seen.

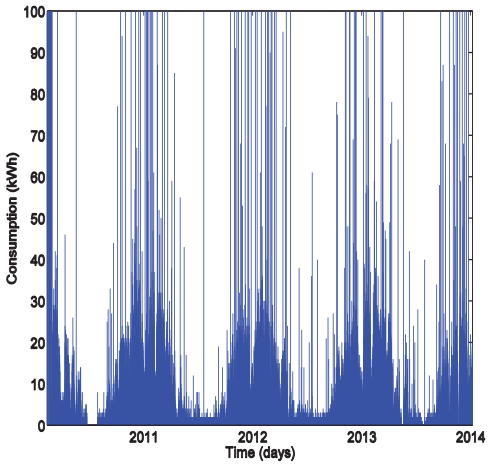

Figure 9a School: Raw gas consumption

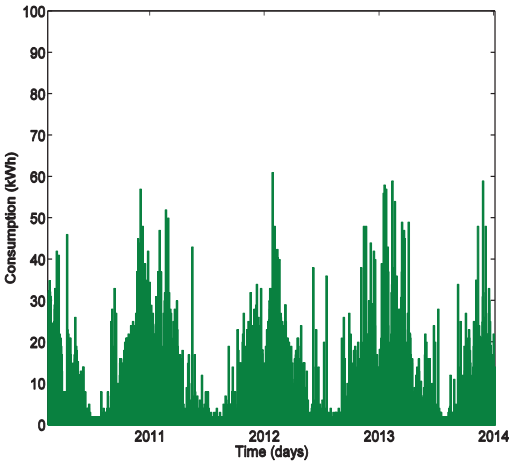

Figure 9b School: Corrected gas consumption

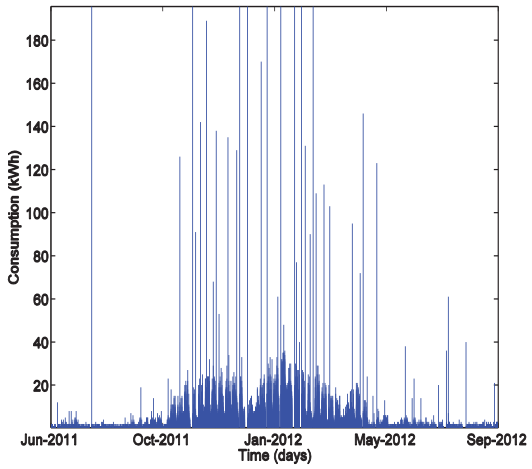

Figure 9c School: Detailed raw gas consumption

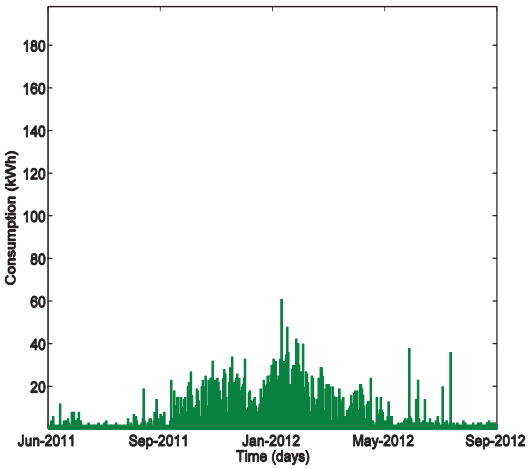

Figure 9d School: Detailed corrected gas consumption

5 Conclusion

This paper outlines a methodology for an automated processing and correction of metered gas consumption data. Its application is aimed at an increased comprehensibility of smart meter readings. It is envisaged that the methodology is suitable for the handling of large building pools without the emphasis on manual data cleansing. The method was tested against readings acquired from operational buildings which featured wide differences with regards to gapless and accurate data.

Results have shown that faulty consumption readings from gas meters can be cleansed in order to obtain replacement values of adequate quality. Our algorithm delivered robust results even with large quantities of corrupt data. It derives missing values and replacement values for outliers from the buildings heat curve. As this heat curve is dependent on the outside temperature, the algorithm will only produce meaningful values for readings collected from main meters. Any sub meter has more dependencies and unless these are considered as well, the application of the methodology will produce deficient data. Further research is required to adequately consider and cleanse individual sub meters.

It should not be neglected that through the introduction of this methodology, the forecast of gas consumption becomes possible. When temperature data from a weather forecast is fed into the algorithm, it will output the estimated gas consumption for the corresponding period. This becomes particularly interesting for a demand based procurement of energy or load balancing algorithms. Additionally, it may support building operators in maintaining sustainable systems operation.

Future work could further increase the accuracy of the cleansed data through differentiation between multiple building usages. Therefore, a higher accuracy may be achieved once buildings consumption is categorized e.g. by data collected during work hours and data collected off-peak. For each category, an individual heat curve would need to be calculated.

We limited the application of the data cleansing algorithm to gas consumption data only. The reason for this is that the algorithm heavily relies on the outside temperature. One may only see electricity consumption depending on the outside temperature if a majority of the consumed electricity is used either for heating or for cooling purposes. This is not the case for the buildings we evaluated.

We have shown that through the introduction of the demonstrated methodology, manual cleansing efforts of monitored gas consumption can be significantly reduced. We believe that this is a great benefit not only for Facility Management companies and utility providers, but also for energy analysts and lastly building owners.

Acknowledgements

This work is funded by the Irish HEA PRTLI-5 programme and Bilfinger HSG FM. The authors would like to thank Bilfinger HSG FM for providing the consumption data used to evaluate the presented work.

References

Abraham, B. and Chuang, A. (1989). *Outlier detection and time series modelling*. Technometrics, vol. 31, no. 2, pp. 241–248

Chen, J., Li, W., Lau, A., Cao, J. and Wang, K. (2010). *Automated load curve data cleansing in power systems*. Smart Grid. IEEE Transactions on, 1(2):213–221.

Davies, L. and Gather, U. (1993). *The identification of multiple outliers*. J. Amer. Statist. Assoc., vol. 88, no. 423, pp. 782–792.

Ferguson, T.S. (1961). *On the rejection of outliers*. Proc. 4th Berkeley Symp. Math. Statist. Probab, vol. 1, pp. 253–287

Fischler, M. A., and Bolles, R. C. (1981). *Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography*. Communications of the ACM, vol. 24, no. 6, pp. 381-395

Hoerster, S., Katzemich F., Menzel, K. (2014). *A Methodology for Data Logging and Retrieval from Remote Sites*. European Conference on Product & Process Modelling, Vienna, Austria

Høverstad, A., Tidemann, A. and Langseth, H. (2013). *Effects of Data Cleansing on Load Prediction Algorithms*, CIASG, IEEE

Knox, E.M. and Ng, R. T. (1998). *Algorithms for mining distance-based outliers in large datasets*. Proc. Int. Conf. Very Large Data Bases

Lagarias, J.C., J. A. Reeds, M. H. Wright, and P. E. Wright. (1998). *Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions*. SIAM Journal of Optimization, vol. 9 no. 1, pp. 112-147

Li, W. (2005). *Risk Assessment of Power Systems: Models, Methods, and Applications*. IEEE Press—Wiley. New York.

Ljung, G.M. (1993). *On outlier detection in time series*. J. R. Statist. Soc. Ser. B (Methodol.), pp. 559–567

Ramaswamy, S., Rastogi, R., Shim, K. (2000). *Efficient algorithms formining outliers from large data sets*. ACM SIGMOD Rec., vol. 29, no. 2, pp. 427–438