# Data Analysis on Building Load Profiles: a Stepping Stone to Future Campus

Long Wang, Yong Ding, Till Riedel, Andrei Miclaus, Michael Beigl
*Karlsruhe Institute of Technology, TECO*
*Karlsruhe, Germany*
*Email:{wanglong,Ding,Riedel,Miclaus,Beigl}@teco.edu*

***Abstract*—For the sustainable development of smart cities across the globe, energy efficiency is becoming a major factor to the maintenance and planning of buildings. In order to demonstrate the potential of data driven approaches in understanding building energy usage, we conduct a data analysis study based on a 10-year data set from 361 buildings of a university campus that are equipped with 1951 smart meters. The preliminary results obtained from our analysis is presented. Results of both clustering analysis and prediction analysis offer a better understanding of common building energy usage as well as a better identification of anomalous behaviors in the usage patterns.**

## 1. Introduction

Probably the greatest challenges of our time lie in the area of sustainable development: reduction of CO2 emissions, increase of energy efficiency, safeguarding the world food supply and the integration of suitable measures into the human environment. Buildings are responsible for 40% of the energy consumption and 36% of the CO2 emissions in the EU [1]. By improving the energy efficiency of buildings, it is possible to reduce the total EU energy consumption by 5-6% and lower CO2 emissions by about 5% [1].

While reducing resource consumption and increasing efficiency is a safe investment in the long run, necessary investments have to be prioritized. Furthermore, measures need to reflect the actual usage of building infrastructure, that constantly changes based on the needs of its inhabitants. Metering data has the potential to reflect these dynamically changing usage very accurately. While energy companies use data for supply planning, currently, still insufficient (big data) tools are available for facility management to use metering data productively. Based on our experiences with analyzing data for industrial applications, we applied a process of potential analysis to identify potential data-driven innovations based on a data set provided to us by the facility management of a closed research campus comprising production facilities, office buildings, a computing center, a cafeteria, transport facilities, a small scale power plant and large scale research facilities. Buildings were built beginning in the 1950s until today based on different civil engineering standards, and have been reused and extended based on the changing demands of approximately 4000 people. Except for residential buildings the campus provides an interesting diverse mix of energy consumers in an area of two square kilometers. For most of those buildings we have nearly 10 years of energy consumption data available for analysis, which provides an excellent basis for further research.

In this paper reports our initial findings applying a data driven approach to the data set for identifying innovation potentials. While the process was explorative, we started with a number of initial questions:

- Is it possible to derive further understanding of the facilities beyond existing information?
- Can we derive hints for infrastructure planning from the data?
- Can we derive easy to understand classification and ranking schemes for buildings?
- What other data sources need to be combined with the data set for the sake of a better interpretation of the data?
- Does a fine-grained look into the data allow us to identify consumption patterns and to disaggregate different demands?

## 2. Related Work

In recent years, energy load forecasting has become one of the major areas of research in electrical engineering, especially short-term load forecasting has become increasingly important since the rise of competitive energy markets [2]. Load forecasting is, however, challenging, due to the inuence of many important exogenous variables. A wide variety of procedures has been tried for short-term load forecasting in the literature. These procedures can typically be classified into two categories of forecasting models [2]: time series (univariate) models, in which the load is modeled as a function of its past observed values, and causal models, in which the load is modeled as a function of some exogenous factors, particularly weather and social variables. More recently, machine learning techniques have been applied to the problem with a specific focus on probabilistic inference modeling [3], support vector machine or regression [4] and artificial neural networks [5]. Also random forest [6] and deep learning [7] have proven their worth for short-term load

forecasting. However, most existing researches of building load forecasting are based on only one or several buildings [8] [9] [10]. Researches based on large set of buildings with various metering data are still rare.

While improving the energy systems themselves has been a vastly studied field, much less research is published in assessing building stock based on available data sets. This is in part due to municipal separation of concerns and energy field market liberalization, which rarely makes it possible to gain access to all relevant data, particularly for research purposes. However there are still some interesting research work, like Mata et al. [11] provide an analysis on the current energy usage and associated carbon dioxide (CO2) emissions of the Swedish residential building stock, which includes single-family dwellings and multi-family dwellings. [12] introduces a bottom-up statistical methodology based on a Geographical Information System (GIS) to estimate the energy consumption of residential stocks across an entire city. Further, weighted robust regression and geographically weighted regression (GWR) models are applied to analyze the determinants and spatial patterns of water consumption in over 2300 multi-family buildings located in New York City. The results disclose the factors which have statistically significant effects on water use intensity [13]. The data set we have got possesses various resource consumption information of a university campus, which offer us potential to gain new hints for improving energy systems.

## 3. Data Overview and Preprocessing

In this section, we present a short overview of our data set and talk about how we preprocessed the data. The data set is made available by the university facility management department, particularly for research dedicated to the improvement of the given energy infrastructure. The data set consists of smart meter data from the 1st of January 2006 to the 6th of May 2015. The total size is approximately 32GB. The meters have been read every 15 minutes, the value and time of cumulative meter readings have been recorded in the data set accordingly. The resource consumption over a time interval of 15 minutes is referred to as quarter consumption. Additional meter information regarding location, measuring medium, identifier etc. is also recorded. The data set comprises in total 1951 meters distributed among 361 buildings. According to the measuring, they are categorized into 11 types as seen in table 1.

Although additional data for a diversity of meters is available, in our first analysis we look at load data as one of the factors for improvement, due to the particularly high needs for cooling and lighting in our facilities (heating is the next obvious thing to look at). However, it is clear on the first glance that not all buildings have the same amount of data available, with some meters having only a few recent measurements. The second issue found is that for some buildings the assignment of meters is not clear. Therefore, buildings with no clearly assigned power meters have to be ignored. After removal of buildings with no power meter or too few readings, we obtained load data for 185 buildings.

| Medium | Quantity |
|---|---|
| Chemical wastewater | 223 |
| Compressed Air | 78 |
| Water Leakage | 8 |
| Gas | 4 |
| Cafeteria Waste | 1 |
| Cooling Water | 14 |
| Rain Water | 310 |
| Electricity | 780 |
| Drinking Water | 203 |
| Purified Water | 52 |
| Heat | 278 |

TABLE 1. MONITORING METER TYPES AND THEIR QUANTITY DISTRIBUTED AMONGST THE 361 BUILDINGS OF THE CAMPUS.

Having a further look at the data, we found that for some meters there are abnormally large quarterly consumption rates. One possible explanation is the historical exchange or reset of meters. Such anomalies (probable measurement failures) were replaced by last-known values. Beyond obvious failures, we had to clean some outliers whose values are considerably larger than the mean value. These outliers are filtered by applying a reasonably safe empirical 3-sigma-thresholds and replacing them with last-known values.

## 4. Load Profile Clustering

Building load profile clustering benefits the campus management department in two aspects: 1) It can help people enhance the comprehension on power consumption of each building, and improve supply planning. 2) It could help to design the Time of Use (ToU) tariff and contribute to formulation of a more efficient and balanced campus grid.

In this section, we explore the following questions: Are there any load cycle patterns in time domain of different buildings? Can we classify buildings into different categories based on these patterns? By investigating the load profile, it is easy to notice that there is a weekly pattern. Inspired by this observation, we restrict ourselves on weekly aggregated load data for clustering. We first upsample the data from 15 min/sample to 3 hours/sample, then we have 8 reading samples for each day and in total 56 samples for a week. K-means algorithm is employed to perform clustering, and two feature vectors are selected for the algorithm:

- **Absolute aggregated load**. We aggregate the absolute load data of all the meters in the same building week by week from 1st Jan 2006 to 6th May 2015.
- **Scaled load fluctuation**. The absolute aggregated load is scaled by subtracting mean value of itself to show the fluctuation of the load.

By trail and error, we set K as 8 for absolute load clustering to have enough load levels. Fig 1 shows that, 7 clusters distinguish each other by the absolute load value (there is one cluster excluded, which contains the aggregated load of all the other meters). The load tends to decrease during the weekend for some buildings but there do exist some buildings whose load does not change too much between weekdays and weekends. This phenomenon implies the

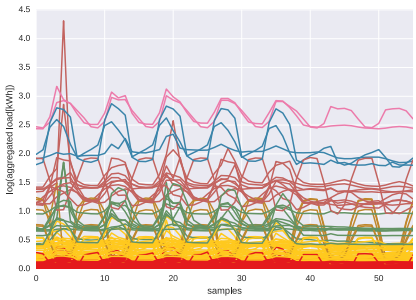


Figure 1. Clustering results of absolute building load profile (weekly aggregated)

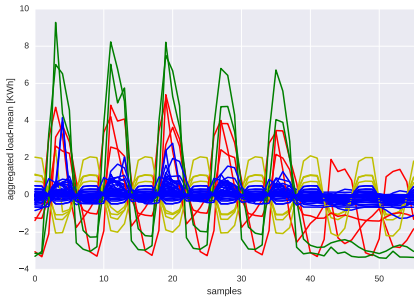


Figure 2. clustering results of scaled building load fluctuation (weekly aggregated)

| Absolute load | Load fluctuation | | | |
|---|---|---|---|---|
| | - | - | 1 | 1 |
| | 1 | - | 1 | 1 |
| | 6 | 1 | 3 | - |
| | 8 | - | - | - |
| | 2 | 3 | - | - |
| | 30 | - | - | - |
| | 127 | - | - | - |

Figure 3. Clustering results matrix. Rows of the matrix stand for clusters based on absolute load profile, columns stand for clusters based on load fluctuation. Numbers in the matrix represent amount of buildings in corresponding cluster.

effects of human activity and facility operation. It could be inferred that building loads with a clear weekday/weekend difference are dominated by human activity, while others with no clear weekday/weekend difference are dominated by facility operation.

Fig. 2 shows clustering results of scaled load fluctuation with K set as 4. Mean values are subtracted from the load profiles of each building to scale them to the same level. The load fluctuation provides us with a new angle on viewing and distinguishing building load profiles. It is interesting to note that the yellow cluster works in a reverse phase style compared with buildings in other clusters. The peaks of its load always appear at night. The reason might be that the equipment in these buildings needs to operate during the night and shut down during the day. We see that there is a potential for load balancing, it is interesting to compare the cluster results with building types and usage types to check the potential information gain.

We combine and reorganize the results above to obtain the clustering results matrix shown in Fig. 3. The numbers in the matrix stand for the amount of buildings with different absolute load and load fluctuation. There are 7 rows in the matrix which corresponds to 7 absolute load clusters with increasing load, and 4 columns representing 4 load fluctuation clusters with increasing fluctuation amplitude. It shows that a large part (127/185) of the buildings could be characterized as relatively low absolute load and low load fluctuation. And there is a decreasing trend for the amount of buildings when the absolute load increases. The first column could be described as nearly constant low load fluctuation, the second column is with slightly higher fluctuation but working with an opposite load phase, the third column is with high fluctuation, the fourth column is with highest fluctuation and stable valley value at weekend.

## 5. Load Profile Prediction

Based on the knowledge acquired from clustering in last section, we further investigate the problem of load profile prediction. Due to European Data Protection Regulation and privacy concerns, we do not have detailed information about the users. Intuitively most campus buildings are planned based on functionality. We assume users sharing the same power meter belong to the same faculty so that they are supposed to behave similarly. Without loss of generality, we selected building A with two power meters and performed load forecasting for both individual meters and the building itself. Support Vector Machine (SVM) is used in our load profile prediction.Considering the clear weekday/weekends and day/night patterns in the building load profiles, the following features are included in the feature vector for training:

- Normalized historical time series load data with 7 * 24 * 4 bits.
- 24-hour information with 24 bits.
- Weekday information with 7 bits.
- Holiday information with 1 bit.

We train the SVM model with a one-year data set (from 1st Jan 2006 to 31st Dec 2006) and test the model with the rest of the data. The criteria we used for measuring the forecasting error between the actual and predicted load is Mean Squared Error (MSE).

As we can see in Fig. 4 and Fig. 5, the trained model could provide load prediction which matches the real data quite well for both individual meter and building cases. Still there are some anomalies in the load profile which go beyond the model's prediction capabilities. Additionally, the trained model tends to generate larger prediction error for weekends compared to weekdays. The MSE performance on prediction for the individual meter and building (see Fig. 6) are quite similar (so we omit the figure for individual meter), which tells that the model works well for both cases. Note that the MSE values increase considerably in the 9th year and drop in the 10th year, which implies there might be some sudden changes in building usage. These anomalies should be excluded for load prediction.

## 6. Discussion and Future Work

Although we have derived some preliminary results on clustering and prediction, it is believed that the work we have done on this data set discloses only a small part of its value. There are still quite a lot of unmined golden nuggets. For instance, taking the data of other media into account could provide further insight on building clustering

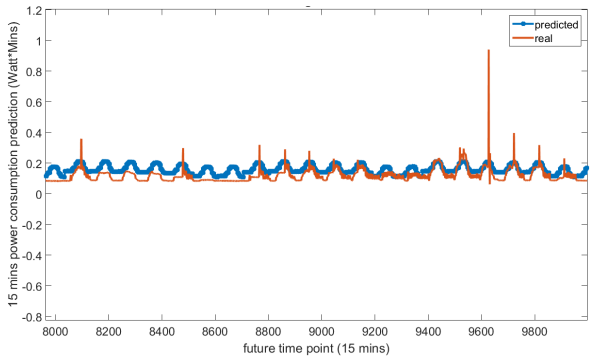


Figure 4. A close look into load forecasting for power meter with ID1

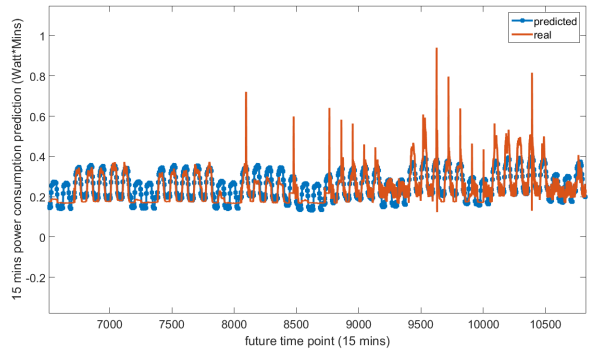


Figure 5. A close look into load forecasting for Building A

results. Viewing the data from a different domain through transformation could also provide new possibilities for clustering. Further, it is interesting to explore if there is any relevance between load profile and building location. For load forecasting, it helps to further improve the prediction performance if the effects of season, weather and outside temperature are taken into account. Fortunately, we have acquired high-resolution environment data for the same campus shortly before the submission of this manuscript. Correlation analysis between the environmental factors and building load profiles is scheduled for future work.

## 7. Conclusion

A data-driven potential analysis on building load profiles was conducted in this paper. Load data of 185 buildings, out of 361 buildings equipped with 1951 meters, were selected and cleaned for further process. A better understanding of building usage pattern is achieved through building load profile clustering. Additionally, an SVM based load profile forecasting model was trained and tested for both building and individual meter cases. Results suggested good prediction accuracy and a high remaining potential for insights out of the data set.

## Acknowledgments

This research was supported in part by..... In particular, the authors thank KIT Campus North for their kind support.

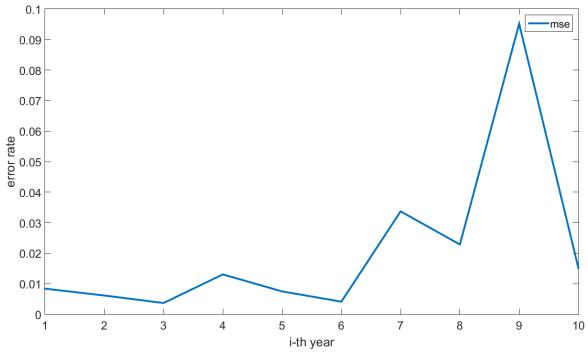


Figure 6. MSE of load forecasting for Building A

## References

[1] "European commission: Building energy efficiency," https://ec.europa.eu/energy/en/topics/energy-efficiency/buildings, accessed: 2017-06-13.

[2] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Transactions on power systems*, vol. 16, no. 1, pp. 44–55, 2001.

[3] G. Wang, A. Kowli, M. Negrete-Pincetic, E. Shafieepoorfard, and S. Meyn, "A control theorist's perspective on dynamic competitive equilibria in electricity markets," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 4933–4938, 2011.

[4] D. Niu, Y. Wang, and D. D. Wu, "Power load forecasting using support vector machine and ant colony optimization," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2531–2539, 2010.

[5] C. Xia, J. Wang, and K. McMenemy, "Short, medium and long term load forecasting model and virtual load forecaster based on radial basis function neural networks," *International Journal of Electrical Power & Energy Systems*, vol. 32, no. 7, pp. 743–750, 2010.

[6] Y.-Y. Cheng, P. P. Chan, and Z.-W. Qiu, "Random forest based ensemble system for short term load forecasting," in *Machine Learning and Cybernetics (ICMLC), 2012 International Conference on*, vol. 1. IEEE, 2012, pp. 52–56.

[7] E. Busseti, I. Osband, and S. Wong, "Deep learning for time series modeling," *Technical report, Stanford University*, 2012.

[8] J. Massana, C. Pous, L. Burgas, J. Melendez, and J. Colomer, "Short-term load forecasting in a non-residential building contrasting models and attributes," *Energy and Buildings*, vol. 92, pp. 322–330, 2015.

[9] Y. Chen, P. Xu, Y. Chu, W. Li, Y. Wu, L. Ni, Y. Bao, and K. Wang, "Short-term electrical load forecasting using the support vector regression (svr) model to calculate the demand response baseline for office buildings," *Applied Energy*, vol. 195, pp. 659–670, 2017.

[10] H. Chitsaz, H. Shaker, H. Zareipour, D. Wood, and N. Amjady, "Short-term electricity load forecasting of buildings in microgrids," *Energy and Buildings*, vol. 99, pp. 50–60, 2015.

[11] É. Mata, A. S. Kalagasidis, and F. Johnsson, "Energy usage and technical potential for energy saving measures in the swedish residential building stock," *Energy Policy*, vol. 55, pp. 404–414, 2013.

[12] A. Mastrucci, O. Baume, F. Stazi, and U. Leopold, "Estimating energy savings for the residential building stock of an entire city: A gis-based statistical downscaling approach applied to rotterdam," *Energy and Buildings*, vol. 75, pp. 358–367, 2014.

[13] C. E. Kontokosta and R. K. Jain, "Modeling the determinants of large-scale building water use: Implications for data-driven urban sustainability policy," *Sustainable Cities and Society*, vol. 18, pp. 44–55, 2015.