ICESF
Nottingham | UK | 9–11 September 2019

The impact of data segmentation on modelling building energy usage

A. William Mounter [1], B. Nashwan Dawood [2], C. Huda Dawood[3]
[123]Teesside University, Stephenson St, Tees Valley, Middlesbrough TS1 3BA, United Kingdom
W.Mounter@tees.ac.uk

Energy is the lifeblood of modern civilisation, with buildings and building construction contributing to roughly 40% of the global energy usage and CO2 pollution. Predicting building energy consumption is essential for energy management and conservation; data driven models offer a practical approach to predicting building energy usage. The aim of this paper is to improve the data driven models available to aid facility managers in planning building energy consumption.

In this case study the 'Clarendon building' of Teesside University was selected for use in using it's BMS data (Building Management System) to predict the building's energy usage. With a particular focus on how data segmentation impacts a model's accuracy and computational time, in predicting temperature related building energy use. Specifically, the effect of segmenting data to accommodate seasonality. With each data segment to be used to train an ANN model (Artificial Neural Network), using ensemble models where data segmentation overlapped.

The potential of these models was compared on the grounds of accuracy and computational speed to each other, then discussed to identify the situational advantages and disadvantages of data segmentation. This study was performed as part of a larger study, in improving building energy use predictions during the operational period in the fields of incorporating user behaviour and accuracy over time.

Key Words: Buildings, Deep learning, Data segmentation, Energy, Prediction.

1. Introduction

The aim of this conference paper is to investigate the impact of data segmentation on the accuracy of building energy use predictions. Data segmentation being the process of dividing and grouping data based on chosen parameters, in this case timeframes, so that it can be used more effectively; (as opposed to data splitting, in which data is randomly split for cross validation usage).

To use an analogy, in cars, winter and summer tyres tend to perform better in their respective seasons than each other and all-season tyres, but poorer than each other and all-season tyres outside of their respective seasons. Would, in the case of machine learning, a model trained with only a season's recorded building data be more accurate at predicting said season's building energy use than a model trained with the variety of data from multiple seasons?

To investigate aim the Clarendon Building, part of Teesside University Campus, was selected for use in this study- due to the data rich environment it's BMS (Building Management system) provided. Previous studies into this building utilising square regression analysis typically had a baseline of "5% Mean Absolute Prediction Error (MAPE)" for the demands of each assets in one day ahead forecasts (Boisson et al.2019). However, the predictions lost accuracy as the rolling horizon increased.

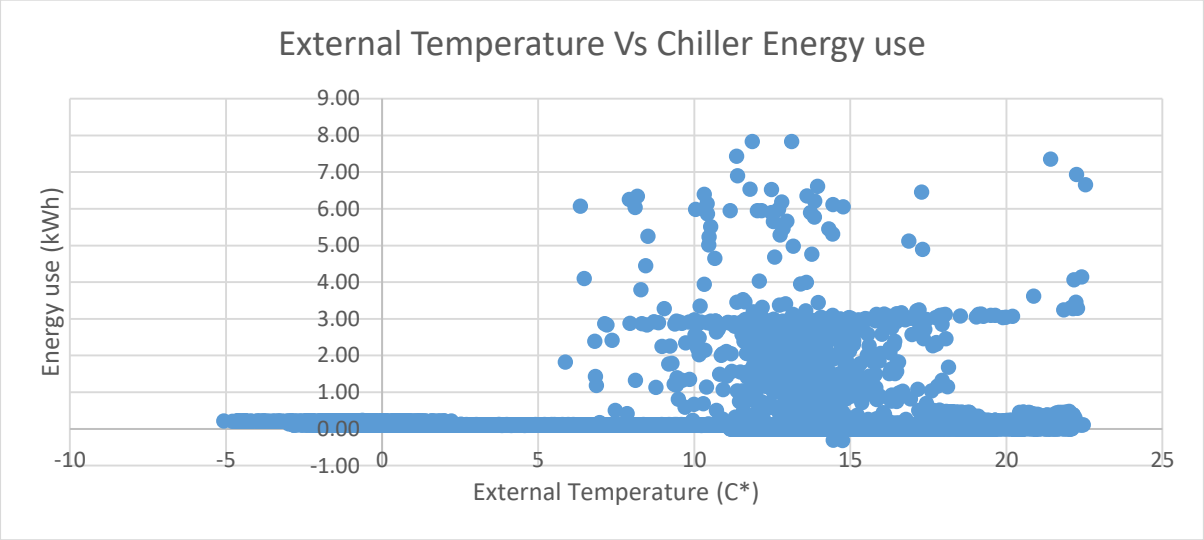
Figure 1: The Clarendon Building, Teesside university (Preston, 2019)

2. Research Method

From the Clarendon building, two main datasets were available: October 2017 to May 2018 and October 2018 to May 2019. These datasets contained 15-minute averages of building elements energy usage, as well as sensory data of the internal and external environmental temperatures, containing approximately 23,000 data events each. Of these building elements, the building chiller system was selected for use in modelling due to the impact seasonality would have on the overall usage of the chillers.

Whilst square regression analysis was used in the previous study into the Clarendon's energy usage patterns, ANN (Artificial Neural Networks) were selected for use for modelling within this study. This was due to ANNs ability to interpret non-liner data (as shown in figure 2) compared to other machine learning methods such as multiple linear regression (Which interprets non-linear data poorly) (Zeyu, W & Ravi, S. 2015). Or in the case of Support vector regression, which is also capable of interpreting non-linear data in irregular energy usage environments, due to the size of the datasets available. SVR possessing greater accuracy in smaller datasets than ANN, but being out performed by ANNs in larger datasets (Grolinger, K, Et al, 2016).

Figure 2



ANNs are based upon the concept of establishing a relationship between independent and dependent variables, though the use of training algorithms (Abbas et al, 2019). In this case study, establishing a relationship between the independent variables of the external and average internal temperature to the dependent variable of the chiller's energy usage in a feedforward neural network. The impact of data segmentation on this process would be observed through changing the size of the training data between: yearly, seasonally, monthly, weekly and daily. Creating an ANN for each dataset and comparing them on the grounds of the percentage error and computation relative to the size of the timeframes being predicted and the time between the training data and predicted events.

2.1. Method limitations

As the external temperature training data used the temperature at the time of each event, opposed to what the external temperature was predicted to be before the event, this would represent an absolute ideal situation. Where in predicting true future events, the difference between the accuracy of the predicted temperature would impact the overall prediction of the building's energy usage and predicting one year into the future in this manner would be significantly inaccurate.

3. Results and Discussion

3.1. Computational time

Whilst choosing the optimum number of hidden layers for the ANN, too few and the ANN would be too linear to predict the outputs, too many and the ANN would overfit the model. It was observed that the computation time did not exceed 1 second until the number of hidden layers approximately exceeded 1000 regardless of size of the dataset used. As changing the size of the dataset did not visibly affect the computational time of the process up until 1000 hidden layers, it can be assumed that the number of times the data is processed has a more significant impact on the computational time than the size of the dataset itself. For all following data 10 hidden layers were used, due to the comparably less percent error observed.

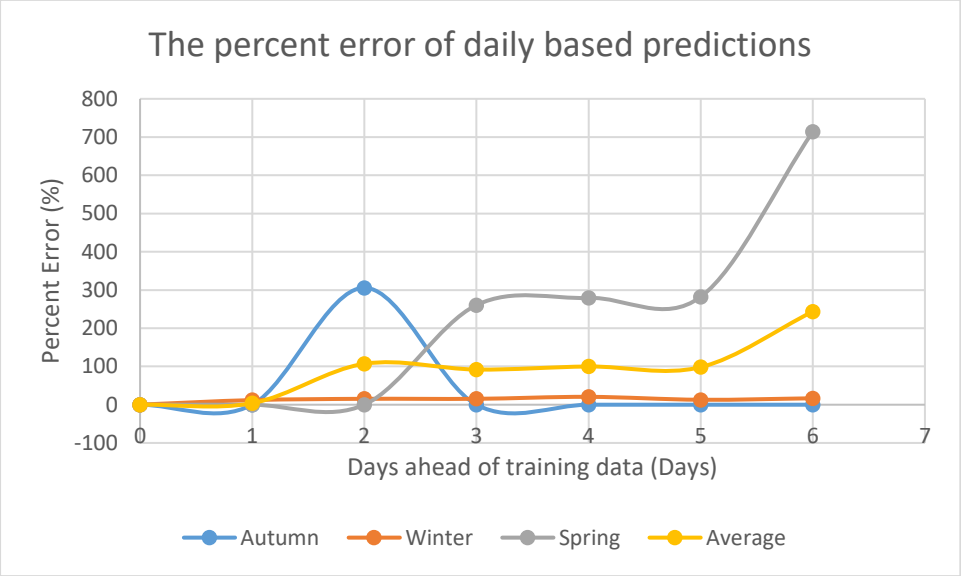
Figure 3

Number of hidden layers	10^0	10^1	10^2	10^3	10^4
Percent Error	4.46E-01	4.09E-01	4.22E-01	4.93E-01	2.87E+01
Computation time (Hour, Min, sec)	0.00.01	0.00.01	0.00.01	0.03.20	00:33:38

3.2. Prediction accuracy

The following is a selection of the percentage errors observed:

Figure 4

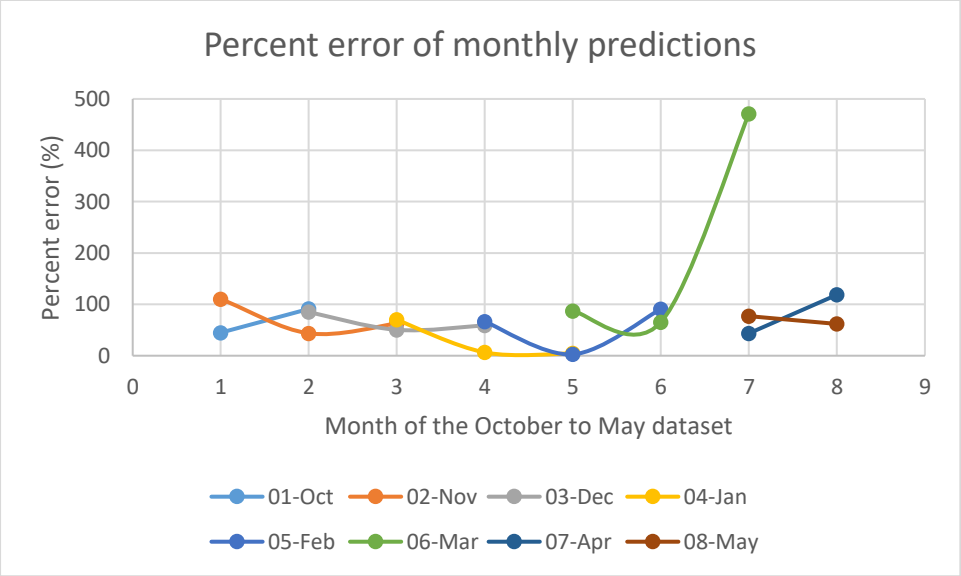


The above is a graphical representation of three days randomly selected from each season and used to model the following week. It was observed that the smaller the training dataset, the more accurate on average it's predictions would be, assuming the areas being predicted did not exceed the parameters of the original training data. Varying from as low as 0.005% error to 715% in a single week of predictions.

Figure 5

Percent error	Day 1	Day 1 vs 2	Day 1 vs 3	Day 1 vs 4	Day 1 vs 5	Day 1 vs 6	Day 1 vs 7
Autumn	0.005	0.007	305.738	0.016	0.017	0.016	0.011
Winter	0.223	12.077	15.416	15.417	20.594	12.883	16.810
Spring	0.003	0.003	0.003	260.504	279.331	282.110	714.127
Average	0.077	4.029	107.052	91.979	99.981	98.336	243.649

Figure 6



Conversely, it was observed that the larger the training dataset, the higher the mean percent error, but the less negative impact anomalies and the time from the training data would have on the predictions. Accuracy would decrease the further the predictions from the training data's relative point in the year, though accuracy would increase as the predictions approached the training data's relative point in the following year.

Figure 7

Percent error	2017/2018	2018/2019	1-year predictions	Increase in Error
Autumn	71.41	60.23	73.13	12.90
Winter	23.95	54.37	75.41	21.04
Spring	62.35	61.30	67.83	6.52
Year	64.38	55.17	71.99	16.82

In figure 7 the accuracy of models developed in the 2017/2018 period are compared with models developed in 2018/2019 as well as used to predict the 2018/2018 based upon its independent variables. Of which it can be observed that in predicting in the short term, greater data segmentation produced greater accuracy, but the further the prediction is away from the greater the error caused by data segmentation.

4. Conclusion

In conclusion, data segmentation can have both a negative and positive impact upon the accuracy of predicted building energy usage dependent upon the duration of the predicted period and the time between the training data and the predicted events. The smaller the predicted period and the time between the predicted event and training data, the more positive the effect of segmenting the data. The greater the size of the predicted period and time between the predicted event and training data the more negative the impact of data segmentation will have on prediction accuracy. This is likely due to the to larger the period and further away the prediction, the increasing likely there will be anomies outside of the range of the training data.

Under the ideal conditions of predicting one day into the future, using a one-day segment to train the ANN, with completely accurate temperature data, an average mean percent error of 4% could be achieved. It can be expected that this error would increase, in the case of predicting future energy usage based upon predicted weather data for the external temperatures and the building temperate comfort zone for the internal.

Based upon these results, four main areas of future work were identified:

- Investigating alternative forms of data segmentation, such as building active and dormancy periods.
- Investigating the accuracy of smaller data segments such as hours in predicting shorter periods into the future
- Using predicted weather data to investigate its impact on prediction accuracy.
- Investigating the accuracy of other machine learning techniques, such as SVR for use in smaller data segments, or other types of ANN and training algorithms.

5. References

preston.(2019). *Tessideuniversity.* Available:http://prestonvsteesside.e-monsite.com/album/teesside-university/. Last accessed 2019.

Pierre BOISSON, Simon THEBAULT, Sergio RODRIGUEZ, Sylvia BREUKERS, Richard CHARLESWORTH, Sarah BULL, Igor PEREVOZCHIKOV, Mario SISINNI, Federico NORIS, Mihai-Tiberiu TARCO, Andrei CECLAN, Tom NEWHOLM,. (2017). *DR Bob D5.1.*Available: https://www.dr-bob.eu/wpcontent/uploads/2018/10/DRBOB_D5.1_CSTB_Update_2018-10-19.pdf. Last accessed 2019.

Husain Abbas, Yousef A. Al-Salloum, Hussein M. Elsanadedy, Tarek H. Almusallam,

ANN models for prediction of residual strength of HSC after exposure to elevated temperature, Fire Safety Journal, Volume 106, 2019, Pages 13-28

Wang, Zeyu & Srinivasan, Ravi. (2015). A review of artificial intelligence based building energy prediction with a focus on ensemble prediction models.

Mauro Ribeiro, Katarina Grolinger, Hany F. El Yamany, Wilson A. Higashino, Miriam A.M. Capretz, Transfer learning with seasonal and trend adjustment for cross-building energy forecasting, Energy and Buildings, Volume 165, 2018, Pages 352-363