

Feature Assessment in Data-driven Models for unlocking Building Energy Flexibility

Anjukan Kathirgamanathan^{1,2}, Mattia De Rosa^{1,2}, Eleni Mangina^{2,3}, Donal P. Finn^{1,2}

¹School of Mechanical and Materials Engineering, University College Dublin, Ireland

²UCD Energy Institute, O'Brien Centre for Science, University College Dublin, Ireland

³School of Computer Science, University College Dublin, Ireland

Abstract

Data-driven approaches are playing an increased role in building automation. This can, in part, be attributed to building operation and energy management system data becoming more readily accessible. A particular application is models to allow predictive control harnessing building energy flexibility, which is of interest to different stakeholders including; energy utilities, aggregators and end-users. Given the possibility of thousands of data features, feature selection becomes a critical part of the model development process. This paper considers various filter, wrapper and embedded methods applied in conjunction with three predictors in addressing the problem of constructing a suitable data-driven model to facilitate predictive control and provision of energy flexibility in a large commercial building. The feature selection algorithms are generally shown to significantly reduce model evaluation time and, in some cases, increase model accuracy. A random forest model with embedded feature selection was found to be the optimal solution in terms of model accuracy.

Introduction

Renewable energy sources such as wind and solar are intrinsically variable by nature and this creates a stability issue for the electrical grid with the fluctuating supply needing to be balanced with the demand (Lund et al. (2015)). With increasing penetration of such sources in most grids, balancing of national grids is becoming a more challenging problem. The flexibility to manage any mismatch can come from either the supply side (through the use of dedicated conventional power plants or storage) or from the demand side (Lund et al. (2015)). Hull (2012) categorises Demand Side Management (DSM) broadly as actions that influence the quantity, patterns of use or the primary source of energy consumed by end users. Buildings represent about 40% of the total primary energy consumption in Europe (Economidou et al. (2011)). This fact combined with their potential for thermal energy storage, makes buildings a very suitable candidate for the provision of energy flexibility. Most buildings use Heating, Ventilation and Air

Conditioning (HVAC) systems for space conditioning and this HVAC load can be shifted using the thermal mass of the building without compromising occupant comfort. The International Energy Agency (IEA) Energy in Buildings and Communities Program (EBC) Annex 67 (Jensen et al. (2017)) provides the current working definition of building energy flexibility as "the ability to manage building demand and generation according to local climate conditions, user needs, and energy network requirements".

To take advantage of thermal mass for demand shifting, a model capturing the thermal dynamics of the building and heating or cooling system is required in conjunction with a predictive control strategy. With buildings being complex, often non-linear in behaviour and no one building constructed or operated the same way, building control based on physics based models has often been limited to being rudimentary and non-predictive. Such approaches are hence unable to fully harness the energy flexibility buildings possess. Numerous studies have concluded that the biggest challenge in the mass adoption of intelligent building control is the cost and effort required to capture accurate dynamic models of buildings (Sturzenegger et al. (2016); Henze (2013)). On this premise, data-driven approaches show great potential for efficient and smart building control. The "Internet of Things" revolution has led to the rapid rise and use of sensors in building control and availability of building data including: HVAC system data, thermal comfort and internal air quality data, power consumption data, external weather data and occupancy data. This provides a significant data source to train data-driven models. Often there may be hundreds if not thousands of features at a modellers disposal. To develop an efficient and accurate data-driven model of a building, feature assessment is a critical element of the model development framework to avoid unnecessary model complexity or missing certain building dynamics. This involves primarily feature selection together with feature engineering. Feature selection is defined as the problem of selecting a subset of (m) features from a larger set (n) features or measurements

to optimise the value of a certain criteria over all subsets of size m (Narendra and Keinosuke (1977)). Feature engineering is domain specific by nature and can be a manual, difficult and time-consuming task, e.g., the addition of external weather data that may not be included with building BEM data. Guyon and Elisseeff (2003) give the primary aims of feature assessment as improving the prediction performance of the learning model, improving the efficiency and cost-effectiveness of the learning model and providing a better understanding of underlying phenomena and processes driving the response variable. Other benefits are aiding data visualisation and reducing data measurement and storage requirements.

This present paper considers the issue of feature assessment in data-driven models used for predictive control and provision of energy flexibility. A case study white-box model of a building is used to generate a significant database of features from which an optimal subset of features can be selected through the use of various feature selection algorithms. This research and the findings helps a move away from observational data to experimental data in the field of data-driven building energy modelling.

The paper is structured as follows: the 'Background' section introduces the different types of feature selection algorithms that are in use and presents a literature review of the seminal works in feature selection. This is followed by a summary of the application of feature selection methods in the realm of building energy modelling and finally the identified research gap and aim of this study are presented. Next, the methods used in this case study are presented along with descriptions of the data and the case study building. Finally, in the results section, the various feature selection algorithms are compared with respect to their performance and identified features for building energy flexibility analysis. Note that the terms 'features' and 'variables' are often used interchangeably

Background

The main approaches to feature selection can be categorised as either filters, wrappers or embedded methods (Guyon and Elisseeff (2003); Molina et al. (2002)). Filter methods are independent of the chosen machine learning model and are used as a pre-processing step with features ranked on the basis of correlation or mutual information criteria. The wrapper method is specific to the machine learning model chosen and uses the model to evaluate and search through subsets of variables. A wide variety of search techniques can be used, including forward search, backward search, Recursive Feature Elimination (RFE), branch-and-bound and Genetic Algorithms (GA), to list a few (Kohavi and John (2011)). Although wrapper methods are generally more com-

putationally intensive than filter methods, embedded methods incorporate feature selection as part of the training process of the model itself and may promise to be more efficient than wrapper methods (Guyon and Elisseeff (2003)), i.e., decision/regression trees and L1 (Lasso) Regularisation.

Current seminal works in feature selection such as Guyon and Elisseeff (2003) present a key message that a unifying theoretical framework to feature selection is lacking due to the diverse approaches available. They highlight the importance of having baseline performance values to compare any approach selected, to understand and quantify the benefits of feature selection. Kohavi and John (2011) present a summary of wrapper methods with the core components that are required being a search space, operators, a search engine and an evaluation function. They emphasise that in general, one should look for optimal features with respect to the specific learning algorithm or predictor and training set at hand. On a similar thread, Molina et al. (2002) showed that different feature selection algorithms behave differently to different data particularities.

The development of data-driven models for building control and harnessing energy flexibility is a nascent field. Even when considering data-driven models used for energy consumption forecasting, few researchers have developed a systematic approach to feature assessment in the development stage of building a predictive model. Numerous studies have selected features purely based on domain knowledge or the features that were actually available. Given that no one building is the same, individual building characteristics can be missed. Zhao and Magoulès (2012) claimed their study to be the first attempt to discuss how to select a subset of features for statistical models applied to the prediction of building energy consumption, through two filter approaches (correlation coefficient and gradient guided feature selection). Kapetanakis et al. (2017) considered the issue of input feature selection looking at linear and monotonic correlations between the features, but was restricted to assessing only thermal loads of commercial buildings. A few studies have investigated wrapper methods in the problem of feature selection. Fan et al. (2014) used RFE together with eight widely used predictive algorithms. Zhang and Wen (2019) developed a methodology tested on both real and synthetic data combining a filter and wrapper approach. The model which was developed with systematic feature selection results showed better accuracy and generalisation in the application of short term building energy forecasting. Finally, there have also been some studies in building energy consumption prediction utilising embedded methods, e.g., Jain et al. (2014) used a L1 Lasso in forecasting the energy consumption of multi-family residential buildings with accurate predictions possible without data from external

sensors such as temperature and occupancy.

With data driven models being increasingly used for building control and for the assessment and provision of energy flexibility, and given very limited existing research in data driven models used in this application, this paper focuses on the performance of various feature selection algorithms and analysis of the data features that are most relevant to this application. Given the potential benefits of feature selection such as increased accuracy and simplicity, this work has the potential to lead to faster and more cost-effective data-driven models enabling greater energy flexibility to be unlocked from a larger range of buildings.

Methods

For flexibility assessment, two response variables are of interest: Zone air temperature (to ensure thermal comfort limits are not violated and to model the thermal dynamics of the building) and the power consumption of the building (to estimate the change in power consumption that can be achieved through demand response measures).

This study considers three predictors: linear regression, Support Vector Regression (SVR) and random forests. The first two models are chosen as they are linear (although SVR can map non-linear functions through the use of kernel functions as explained later) and hence can easily be integrated with a predictive control optimisation framework. Random forests are chosen because they are effective at capturing non-linear and complex behaviour and the work of Jain et al. (2016) has shown that random forests are capable of being integrated with receding horizon control using the technique of separation of control and disturbance variables.

SVR is a form of Support Vector Machines (SVM) that finds a decision function or model to represent the relationship between the features and the target. Where a linear function is not enough to map the relationship, the problem can be mapped to a higher dimensional feature space through the use of kernel functions. See Smola and Scholkopf (2003) for further details on the formulation.

One of the biggest drawbacks of a classical decision tree is its tendency to overfit to training data and random forests are an ensemble method that was developed to combat this. Many parallel learners exploiting independence between the learners are averaged to reduce the error of the ensemble predictor. Each tree in the ensemble is built from a sample drawn with replacement (i.e. a bootstrap sample) from the training set. In addition, instead of using all the features, a random subset of features is selected, further randomizing the tree. See Breiman (2001) for the seminal work and more detail on random forests.

The overall feature assessment procedure is described graphically in Figure 1. Further details on each step are provided below.

Step 1. Generate Synthetic Data

The US Department of Energy Large Office archetype white-box model (using EnergyPlus) has been taken and modified to be the testbed building providing the synthetic data for training the data-driven model initially (Deru et al. (2011)). This building is a 12 storey building with a floor space of 46,000 m^2 (Figure 1). The building 'core mid' zone is investigated as one of the response variables as this zone represents the majority of the zonal temperatures, being the largest zone per floor and representing 10 of the 12 floors through symmetry properties of the simulation. The building has a gas boiler for heating, two water-cooled chillers for cooling and a multizone variable air volume (MZ VAV) system for air distribution. A combined PV-battery system was added to the existing model. The building complies with the minimum requirements of ASHRAE Standard 90.1-2004 and the version for climate zone 4C was selected, the closest climate zone to that of Dublin, Ireland, for which the weather data is used. Weather data, extracted from the weather file used in the simulation, also forms a part of the synthetic data and features studied. The model uses a simulation time-step of 15 minutes and was simulated for an entire year to generate the training and test dataset. The data was split with 75% of the dataset being used for training (January to September) and 25% for testing (October to December). The advantages of using a white-box model to generate synthetic data is that a comprehensive database can be generated for training that can be used to compare the various feature selection methods without concern for data quality issues that plague most real datasets from buildings. The question of how the feature selection algorithms correspondingly perform on this real data is the subject of future research. However, note that approximately 13,000 features were available as outputs from the white-box model of the 'Large Office' Building with many of these variables not being practically measurable or observable. An initial manual filtering was done to remove such variables from the EnergyPlus output to simulate only data that would realistically be output from a Building Energy Management System (BEMS).

Steps 2-4. Feature Selection (Data Pre-processing)

Although the synthetic data does not contain any missing data, this step was added to the workflow so that when it is used with real data, any feature that has more than a certain threshold of missing values is removed from the dataset. Features with a single value contain no useful information for a predictive model and hence these are also removed from the dataset. Redundant features are those that are highly correlated with one another and hence one of them is considered redundant as it does not add any

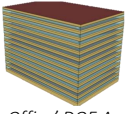
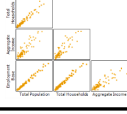

Step 1	Generate DR Synthetic Data  <ul style="list-style-type: none"> 1 year data, 15 min resolution Split into train/test Keep only observable features 'Large Office' DOE Archetype	~13,000 features ↓ 3,886 Features X																								
Step 2	FS: Remove features with missing data Any features with more than 40% missing values removed	3,884 features																								
Step 3	FS: Remove features with single values Any features with single values removed	3,648 features																								
Step 4	FS: Remove collinear features  Remove redundant features (Pearson correlation > 0.9 with another feature)	143 features																								
Step 5	FE: Add Proxy Variables  <ul style="list-style-type: none"> Hour of day, day of week Lag terms 	151 features																								
Step 6	FE: Normalise Data Scale feature data to [0,1] range	151 features																								
Step 7	FS: Apply & Evaluate FS Algorithms <table border="1" data-bbox="263 929 678 1176"> <thead> <tr> <th>FS Algorithm /Predictor</th> <th>Linear Regression</th> <th>SVR</th> <th>Random Forest</th> </tr> </thead> <tbody> <tr> <td>All Features</td> <td>ALL-LR</td> <td>ALL-SVR</td> <td>ALL-RF</td> </tr> <tr> <td>Filter - Pearson</td> <td>F-P-LR</td> <td>F-P-SVR</td> <td>F-P-RF</td> </tr> <tr> <td>Wrapper - RFE</td> <td>W-RFE-LR</td> <td>W-RFE-SVR</td> <td>W-RFE-RF</td> </tr> <tr> <td>Wrapper - GA</td> <td>W-GA-LR</td> <td>W-GA-SVR</td> <td>W-GA-RF</td> </tr> <tr> <td>Embedded</td> <td></td> <td></td> <td>E-RF</td> </tr> </tbody> </table>	FS Algorithm /Predictor	Linear Regression	SVR	Random Forest	All Features	ALL-LR	ALL-SVR	ALL-RF	Filter - Pearson	F-P-LR	F-P-SVR	F-P-RF	Wrapper - RFE	W-RFE-LR	W-RFE-SVR	W-RFE-RF	Wrapper - GA	W-GA-LR	W-GA-SVR	W-GA-RF	Embedded			E-RF	various Features X'
FS Algorithm /Predictor	Linear Regression	SVR	Random Forest																							
All Features	ALL-LR	ALL-SVR	ALL-RF																							
Filter - Pearson	F-P-LR	F-P-SVR	F-P-RF																							
Wrapper - RFE	W-RFE-LR	W-RFE-SVR	W-RFE-RF																							
Wrapper - GA	W-GA-LR	W-GA-SVR	W-GA-RF																							
Embedded			E-RF																							

Figure 1: Feature Assessment Procedure

extra useful information to the dataset. The Pearson correlation coefficient is used with a threshold value of 0.9, for feature pairs that are correlated with a value greater than this threshold, one of the features is removed from the dataset.

Steps 5 & 6. Feature Engineering

In these steps, features are generated from the raw data that is output from EnergyPlus. An example of this is proxy variables such as hour of the day and day of week. EnergyPlus outputs a timestamp variable that is not suitable for using as a feature in a predictive model directly. This timestamp variable is processed to extract variables such as the hour of the day and day of the week which can have significant predictive power given the periodic nature of building occupancy and energy consumption. Lag terms are also introduced here for the response variable. Normalising or scaling the training data is generally considered to be good practice in machine learning problems, especially for linear models where features with large ranges induce high variance and may become unnecessarily important. The approach used here is to scale the data to be in the range of [0,1].

Step 7. Apply & Evaluate Feature Selection Algorithms

Four feature selection algorithms are compared in this study, namely, one filter method based on the Pearson correlation coefficient, two wrapper methods (RFE and GA) and one embedded method (random forest). The Pearson correlation coefficient is used to rank all the features with respect to the response variable and the features that have a value of greater than 0.6 are selected. RFE is a backward selection technique where the model is initially fitted using all features and then at each iteration a specified number of features that are the weakest are removed. To find the optimal number of features, this method is used with cross-validation to score different subsets, utilising the Python Sklearn package (Pedregosa et al. (2011)). The GA is a heuristic optimisation method inspired by evolution, where the genes of organisms tend to evolve over successive generations to be more successful to the given environment. It is a stochastic method that can be used for feature selection on a given predictor that works based on the mechanisms of natural genetics and biological evolution found in nature with the three major steps being selection, crossover and mutation (Chtioui et al. (1998)). In the case of feature selection, each individual of the population represents a candidate subset of features with each individual being assigned a fitness value based on a fitness function (here being the prediction error on the cross-validation sets). In selection, the individuals with a high fitness value are given more chance to be selected for reproduction. During crossover, portions of the parent solutions are exchanged and finally in mutation, one or more components of the child individual are randomly changed. The random forest has an intrinsic feature importance variable (either the mean decrease impurity or mean decrease accuracy) that is used in constructing the model making this an embedded method.

All pre-processing of data, feature engineering and implementation of the above feature selection algorithms was carried out in Python on a server machine with an Intel(R) Xeon(R) CPU E5-2697 v2 2.7 GHz and 256 GB of RAM.

To evaluate the performance of the various feature selection algorithms, two metrics were used: (i) the root mean squared error (RMSE) of the prediction on the test set and (ii) the evaluation time of the model with the selected features on the test dataset. The RMSE is defined in Equation (1) as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}} \quad (1)$$

where \hat{y}_i is the predicted output value and y_i is the actual output variable for the i^{th} sample in the testing subset. N is the number of samples in the testing subset.

Results and Discussion

Comparison of Feature Selection Algorithms

The results of the various feature selection algorithms are compared first with three different predictors: linear regression, SVR and random forests. The two response variables considered in this feature selection study are the core mid zone temperature and building total power consumption. For reference, a naive prediction based purely on the lagged term from a week prior results with a RMSE of 0.56°C for the core mid zone temperature and a RMSE of 6.0 kW for the building total power consumption.

A comparison of the RMSE achieved by the SVR and run time for the different FS algorithms for the response variable of the core mid zone temperature is given in Figure 2. As this figure shows, all feature selection techniques result in significant reductions in the model evaluation time (over 93% reduction for the F-P-SVR model) with very minor gains in the prediction RMSE. A comparison of the RMSE achieved by the random forest predictor and run time for the different feature selection algorithms is presented in Figure 3 for the response variable of the building total power consumption. This figure shows that all feature selection techniques are able to reduce the evaluation time of the random forests model significantly (between 50% and 92% reductions) but given that the random forest predictor is inherently an embedded method of feature selection, the addition of the other feature selection algorithms does not significantly reduce the RMSE and, in some cases, even has a detrimental effect (such as using the F-P-RF model). The results presented for the embedded case (E-RF) is essentially the same as the "All Features" scenario (ALL-RF) except that hyper-parameter tuning has been employed to optimise the random forest structure. The random forest predictor is a special case, as an embedded method of feature selection is

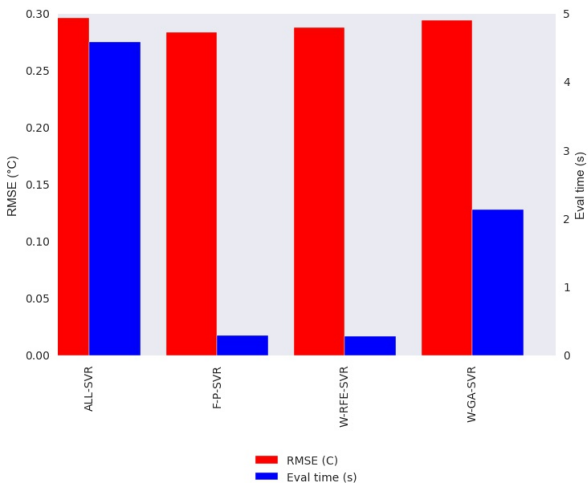


Figure 2: Comparison of FS Algorithms by RMSE and run time for the SVR Predictor with a response variable of the core mid zone temperature

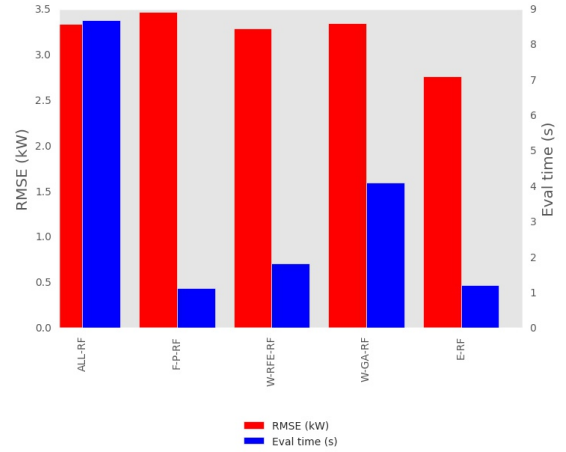


Figure 3: Comparison of FS Algorithms by RMSE and run time for the random forest Predictor with a response variable of the total building power consumption

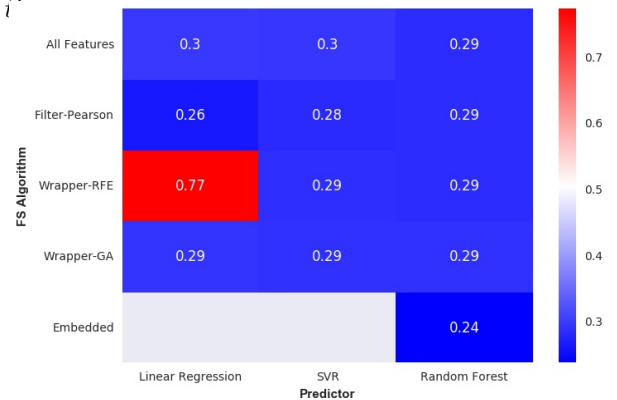


Figure 4: Comparison of FS Algorithms by RMSE with a response variable of the core mid zone temperature

inherently built into the algorithm which means that additional feature selection techniques are not necessary unless the model run time is a critical concern, e.g., if the control time step is in the order of seconds. Figures 4 and 5 summarise the RMSE values achieved by the selected features for all combinations given in Figure 1 and generally show the superiority of the random forest predictor with its embedded feature selection. It should also be pointed out that whilst not illustrated here, the wrapper methods of feature selection can be computationally intensive with the RFE and GA algorithms taking several hours to run. Given that model training should be conducted offline in applications utilising data-driven control, this is considered to be acceptable. Figure 6 illustrates the predicted and synthetic data values of the core mid zone temperature for varying n-steps ahead (from 15 minutes ahead to 1 day ahead) for a week in November (which is part of the test dataset) using the random forest with features selected through the inbuilt feature importance method (embedded method). Hyper-parameter tuning using a randomised grid search approach is used to tune the models. The figure shows that the predictive accuracy declines with increasing prediction horizons as

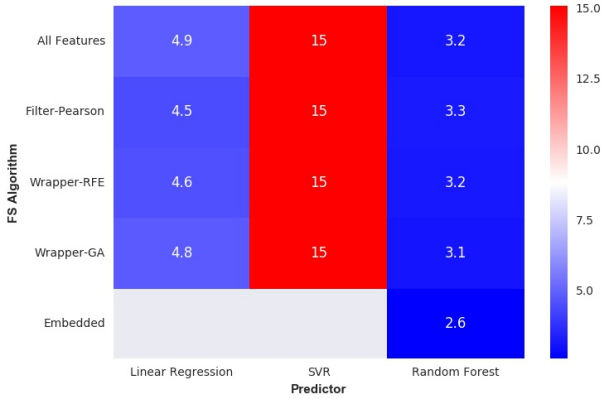


Figure 5: Comparison of FS Algorithms by RMSE with a response variable of the total building power consumption

expected (RMSE ranges from 0.24°C for 1-step ahead to 0.53°C for 96-step ahead).

Features Selected

This section investigates the actual features selected by the various feature selection algorithms. The following tables give a count of the top features selected by the various combinations of feature selection algorithms and predictors. As Table 1 and Table 2 show, the lag terms are significant for predicting both the zonal temperature and the building power consumption (1-step ahead). Specific to the core mid zone temperature, the floor temperature of the zone is the third most relevant feature signifying the physical thermal connection between the two. A similar connection exists between the basement and core bottom zones and the core mid zone temperatures. Considering the features selected for predicting the building total power consumption, aside from the lag terms importance as mentioned above, the power consumption of interior lights and equipment are the most significant contributors to the total power demand and this is highlighted in Kathirgamanathan et al. (2018) where the same building is used as the virtual DR testbed building. The influence of occupancy on the power demand is seen through the high relevance of the zone CO_2 concentration which can be taken to be a proxy measure for occupancy. Finally, the periodicity of power consumption in this building is seen through the importance of the 'hour' proxy variable. This is particularly applicable to commercial buildings where operation and hence occupancy and power consumption tends to follow 'standard' building operating hours as opposed to residential buildings. Interestingly, the control variables that were kept for the feature selection analysis (fan air speed and cooling setpoint schedule value) were not necessarily selected through the feature selection algorithms and found to be part of the most optimal subset. This could indicate that the building was not excited enough in the training data for the control features to offer much to the predictive model. The external weather features such as air dry bulb temper-

ature, relative humidity and solar radiation did not rank very highly for both 1-step ahead response variables, which is surprising given the amount of literature where external weather features are used (e.g., Fan et al. (2014); Kapetanakis et al. (2017); Drgona et al. (2018)). However, there are studies that have similarly found a lack of relevance of external weather

Table 1: Selected features from various FS algorithms and predictors for response variable of core mid zone temperature.

Features Selected	Count
CORE_MID:Zone Thermostat Air Temperature [C]	9
CORE_MID_ZN_5_FLOOR:Surface Outside Face Temperature [C]	8
CORE_MID:Zone Thermostat Air Temperature [C]-1	8
CORE_MID:Zone Thermostat Air Temperature [C]-5	7
CORE_MID:Zone Thermostat Air Temperature [C]-3	6
CORE_MID:Zone Thermostat Air Temperature [C]-4	6
VAV 2:Air System Outdoor Air Heat Recovery Bypass Minimum Outdoor Air Mixed Air Temperature [C]	6
GROUND FLOOR PLENUM WALL NORTH:Surface Inside Face Temperature [C]	6
CORE_BOTTOM: Zone Thermostat Air Temperature [C]	6
CORE_MID:Zone Thermostat Air Temperature [C]-2	6
BATTERY: Electric Storage Simple Charge State [J]	5
VAV 3 HEATC-VAV 3 FANNODE:System Node Temperature [C]	5
PERIMETER_BOT_ZN_2 VAV BOX OUTLET NODE NAME:System Node Current Density [kg/m3]	5
BASEMENT:Zone Thermostat Air Temperature [C]	5
Heating: Gas [J]	5
PERIMETER_MID_ZN_4 VAV BOX REHEAT COILDEMAND OUTLET NODE:System Node Current Density [kg/m3]	4

Table 2: Selected features from various FS algorithms and predictors for response variable of building total power consumption.

Features Selected	Count
Electricity:Facility [J]	9
Electricity:Facility [J]-3	8
Electricity:Facility [J]-4	8
CORE_BOTTOM:Zone Air CO_2 Concentration [ppm]	8
TOP FLOOR PLENUM:Zone Air CO_2 Concentration [ppm]	8
Electricity:Facility [J]-2	8
Electricity:Facility [J]-1	7
InteriorLights:Electricity [J]	7
Fans:Electricity [J]	7
VAV_3_FAN: Fan Air Mass Flow Rate [kg/s]	6
InteriorEquipment:Electricity [J]	6
CORE_BOTTOM VAV BOX OUTLET NODE NAME:System Node Relative Humidity [%]	6
WaterSystems:Gas [J]	6
CORE_TOP VAV BOX COMPONENT:Zone Air Terminal VAV Damper Position [J]	6
Electricity:Facility [J]-5	6

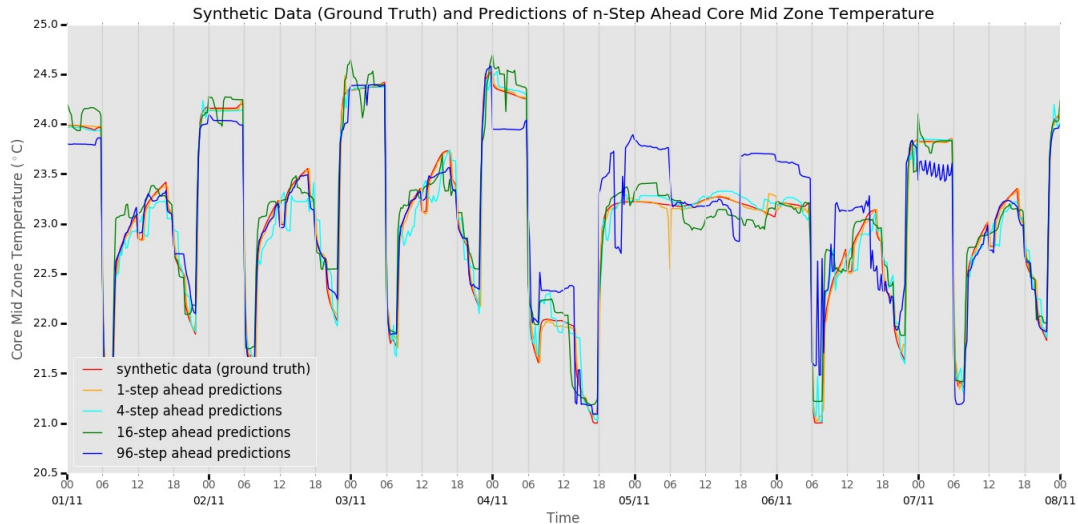


Figure 6: Comparison of predicted (for varying n -steps ahead) and synthetic data (ground truth) values from model 'E-RF' with a response variable of the core mid zone temperature for one week in test set

features (e.g., Zhang and Wen (2019); Nghiem and Jones (2017)). For the core mid zone temperature, this can be explained by the central location of this zone in the building damping the effects of external conditions and the significance of internal gains on the cooling and heating loads of this zone. For the building total power consumption, as explained above, the cooling and heating loads which are most sensitive to external weather features are relatively small compared to the lighting and equipment loads which dominate the total power consumption. As a method of validation of the feature selection method, the random forest was used to predict the core mid zone temperature for varying n -steps ahead (from 1-step ahead to 96-step ahead) with hyperparameter tuning (Figure 6). Analysing the features selected by this method, whereas lag terms are quite dominant for 1 and 4 steps ahead, for greater prediction horizons, this is not the case as expected and markers for periodicity and occupancy such as 'day' and zone CO_2 concentration are increasingly selected. Although not included in this study, for longer prediction horizons, it is expected that weather forecast variables will be more relevant features.

Conclusion

Feature selection algorithms commonly used in literature have been applied to a specific case of a synthetic dataset from a commercial building for use in constructing a predictive model for energy flexibility applications. The results show that the feature selection techniques generally offer significant reductions in the model evaluation time (ranging from 50% to 94% reductions) and for the predictors selected, did not make any significant difference in the predictor accuracy. The choice of feature selection algorithm should generally be made based on the predictor used, for example if one has selected the random forest as the predictor to be used, applying a wrapper feature

selection method may be considered to be unnecessary. The random forest model with embedded feature (E-RF) selection was found to give the best accuracy considering both the zone temperature and total power consumption response variables.

For the virtual building considered in this study, feature selection showed that the lag terms of both response variables were highly relevant and most feature selection algorithms ended up choosing up to five lagged terms. For the core mid zone temperature, the feature selection picked up on thermal connections to adjacent surfaces and zones with these features being selected frequently. For the building total power demand, the high portion of power consumption dedicated to interior lights and equipment was revealed and this also explains the importance of a proxy variable such as hour of the day. The relative unimportance of external weather features was found to be the case for this particular building. This work identifies the important features for development of a data-driven model to harness the energy flexibility available in this case study building. This methodology can be repeated with any building and allows the specification of what experimental data is required from a building.

Given that receding horizon control problems require sufficiently accurate predictions over the range of the prediction horizon, future work should consider the issue of balancing the differing optimal variables between short-term predictions and longer-term predictions. Further work is also required to determine the importance of weather predictions on longer prediction horizons. The approach also needs to be applied to real data from a building. Real data has stochasticity that this synthetic data used in this study is missing as well as data-quality issues that a feature selection methodology needs to be robust against.

Acknowledgment

This work has emanated from research conducted with the financial support of Science Foundation Ireland under the SFI Strategic Partnership Programme Grant Number SFI/15/SPP/E3125.

References

- Breiman, L. (2001). Random Forests. *Machine Learning* 45(1), 1–33.
- Chtioui, Y., D. Bertrand, and D. Barba (1998). Feature Selection by a Genetic Algorithm . Application to Seed Discrimination by Artificial Vision. *J Sci Food Agric* 77.
- NREL (2011). *U.S. Department of Energy commercial reference building models of the national building stock*.
- Drgona, J., D. Picard, M. Kvasnica, and L. Helsen (2018). Approximate model predictive building control via machine learning. *Applied Energy* 218(February), 199–216.
- Buildings Performance Institute Europe (BPIE) (2011). *Europe’s Buildings Under the Microscope*.
- Fan, C., F. Xiao, and S. Wang (2014). Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy* 127, 1–10.
- Guyon, I. and A. Elisseeff (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)* 3(3), 1157–1182.
- Henze, G. P. (2013). Model predictive control for buildings: a quantum leap? *Journal of Building Performance Simulation* 6(3), 157–158.
- EA Technology (2012). *DSM / DSR: What, Why and How?*.
- Jain, A., M. Behl, and R. Mangharam (2016). Data Predictive Control for Building Energy Management. *Proceedings of BuildSys ’16* (May), 245–246.
- Jain, R. K., T. Damoulas, and C. E. Kontokosta (2014). Towards Data-Driven Energy Consumption Forecasting of Multi-Family Residential Buildings: Feature Selection via The Lasso. *Computing in Civil and Building Engineering*, 1675–1682.
- Jensen, S. Ø., A. Marszal-Pomianowska, R. Lollini, W. Pasut, A. Knotzer, P. Engelmann, A. Stafford, and G. Reynders (2017). IEA EBC Annex 67 Energy Flexible Buildings. *Energy and Buildings* 155, 25–34.
- Kapetanakis, D.-S., E. Mangina, and D. P. Finn (2017). Input variable selection for thermal load predictive models of commercial buildings. *Energy and Buildings* 137, 13–26.
- Kathirgamanathan, A., K. Murphy, M. D. Rosa, E. Mangina, and D. P. Finn (2018). Aggregation of Energy Flexibility of Commercial Buildings. In *Proceedings of eSim 2018, May 9-10, 2018*, Montreal, pp. 173–182.
- Kohavi, R. and H. John (2011). Wrappers for feature subset selection. *Artificial Intelligence* 97(97), 273–324.
- Lund, P. D., J. Lindgren, J. Mikkola, and J. Salpakari (2015). Review of energy system flexibility measures to enable high levels of variable renewable electricity. *Renewable and Sustainable Energy Reviews* 45, 785–807.
- Molina, L. C., L. Belanche, À. Nebot, J. Girona, and C. N. C (2002). Feature Selection Algorithms: A Survey and Experimental Evaluation. *Proceedings of ICDM 2002.*, 306–313.
- Narendra, P. M. and F. Keinosuke (1977). A Branch and Bound Algorithm for Feature Subset Selection. *IEEE TRANSACTIONS ON COMPUTERS C-26*(9), 917–922.
- Nghiem, T. X. and C. N. Jones (2017). Data-driven Demand Response Modeling and Control of Buildings with Gaussian Processes. In *2017 American Control Conference (ACC), Seattle, WA*, Seattle, pp. 2919–2924.
- Pedregosa, F., R. Weiss, M. Brucher, G. Varoquaux, A. Gramfort, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Perrot, and É. Duchesnay (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.
- Smola, A. J. and B. Scholkopf (2003). A Tutorial on Support Vector Regression . *Statistics and Computing* 14(3), 199–222.
- Sturzenegger, D., D. Gyalistras, M. Morari, and R. S. Smith (2016). Model Predictive Climate Control of a Swiss Office Building: Implementation, Results, and Cost-Benefit Analysis. *IEEE Transactions on Control Systems Technology* 24(1), 1–12.
- Zhang, L. and J. Wen (2019). Energy & Buildings A systematic feature selection procedure for short-term data-driven building energy forecasting model development. *Energy & Buildings* 183, 428–442.
- Zhao, H.-X. and F. Magoulès (2012). Feature Selection for Predicting Building Energy Consumption Based on Statistical Learning Method. *Journal of Algorithms & Computational Technology* 6(1), 59–77.