Towards Robustness of Data-Driven Predictive Control for Building Energy Flexibility Applications

Anjukan Kathirgamanathan[1,3], Mattia De Rosa[1,3], Eleni Mangina[2,3] and Donal P. Finn[1,3]
[1]School of Mechanical and Materials Engineering, University College Dublin, Ireland
[2]School of Computer Science, University College Dublin, Ireland
[3]UCD Energy Institute, O'Brien Centre for Science, University College Dublin, Ireland

Abstract

Identifying physics-based models of complex dynamical systems such as buildings is challenging for applications such as predictive and optimal control for demand side management in the smart grid. Data-driven predictive control using machine learning algorithms show promise as a more scalable solution when considering the greater building stock. The robustness of these algorithms for different climate data, building types, quality and quantity of data, is still not yet well understood. The objective in this study is to investigate model identification and the resultant accuracy for these various contexts using the 'separation of variables' technique (DPC-En) and the consequent performance implications of the data-driven controller. The DPC-En controller is tested using a closed-loop simulation testbed of a 'large office' archetype building. The results show that the technique is relatively robust to missing data and different climate types and delivers promising results using limited training data without the need for disruptive excitation measures. This work contributes to enabling a greater proportion of the diverse building stock to be utilised for demand side management by harnessing their inherent energy flexibility potential.

Introduction

Demand Side Management using Buildings

Utilisation of energy flexibility in buildings, within a demand side management (DSM) context, is increasingly being considered as a vital contributor to the future smart grid given increasing penetrations of variable renewable energy sources. Gellings (Electric Power Research Institute, USA) coined the term 'DSM' defining "DSM activities as those which involve actions on the demand (i.e., customer) side of the electric meter, either directly or indirectly stimulated by the utility" (Gellings (1985)). As significant end-use consumers of electricity (Economidou et al. (2011)), buildings have significant energy flexibility potential through the use of their passive thermal mass for shifting consumption (Reynders et al. (2013)).

Model predictive control (MPC) provides an automated way to harness building energy flexibility effectively given its inherent self-optimisation, subject to a model capturing the building thermal dynamics and constraints, and ability to include predictive disturbances of the building in the model formulation (De Coninck et al. (2015)). MPC generally requires a linear model to guarantee a convex optimisation problem. However, a fully linear model is often not feasible with simplified building models, especially if they include ventilation models where bi-linearities are introduced due to modelling the product of temperature and mass flow rate (Sturzenegger et al. (2014)). Further, no one building is the same, being constructed, designed and operated differently to others, meaning that one building model cannot simply be employed on another building in a plug-and-play manner. The challenges of generating a suitable and accurate model for MPC have been highlighted as one of the most significant limitations of MPC when applied to building energy management and utilisation of energy flexibility (Sturzenegger et al. (2016); Henze (2013)). The "Internet of Things" revolution has led to the rapid proliferation of sensors in building control and availability of building data and coupled with the advances in machine learning, data-driven control approaches are proving to be a viable alternative to traditional MPC.

Data-Driven Predictive Control (DPC) for Building Energy Management and 'Separation of Variables'

Data-driven models, together with representing the building as a cyber-physical system, show promising potential in harnessing energy flexibility in commercial buildings, where deriving a physics-based control oriented model is challenging. A comprehensive literature review of prior work, on the application of data-driven methods for DSM, has shown that most studies apply such methods to an individual case study (often simulated), hence questioning the results achieved, its replicability and further application over many different case studies.

One such promising technique is 'Separation of Variables' where regression trees are built using training data ensuring that the control inputs (variables to be optimised) are excluded (Behl et al. (2016)). This control input data is then used to fit affine convex models under each leaf of the regression trees leading to a convex optimisation problem. The initial study by Behl et al. (2016) only provided a one-step look ahead prediction and hence was not suitable for a receding horizon problem such as MPC. Jain et al. (2016) solved this problem and extended the initial work using multi-variate regression trees with output corresponding to each step of the prediction horizon. They were able to show that their technique had only a small justifiable additional cost compared to a traditional linear MPC approach. The approach was further improved by Smarra et al. (2018) where each regression tree was replaced with a random forest (ensemble of regression trees) in order to reduce the variance in prediction and mitigate the tendency of regression trees to overfit. The authors termed this technique 'Data-Driven Predictive Control with Ensemble Methods' or DPC-En. The technique is summarised graphically in Figure 1. This figure shows how the training data is composed of the disturbance variables (d) which includes both autoregressive terms (δ_d), predicted disturbances ($\bar{d}$) and autoregressive terms of the state variables (x). From this training data, random forest models are constructed over the prediction horizon (j) predicting the evolution of the state variables. Each forest is composed of n regression trees with an ensemble θ being taken of all the predictions. In every leaf node of each of the trees, a linear model is fitted from the affine combination of the control inputs to refine the state prediction. In this case, the ensemble of the regression model coefficients is used. The reader is referred to Smarra et al. (2018) for more details on the algorithm used to train the data-driven model and the mathematical definitions.
Training data collected either from real buildings or synthetic simulation data is often based on buildings employing rule-based controllers. The data may not be rich and varied in terms of explaining the relationship between input and output variables. Real data from buildings is often plagued with missing data and the effect of poor data quality on the performance of a data-driven controller is currently not well understood. Bunning et al. (2019) presented a real-life implementation, the only one to the knowledge of the authors, to a residential apartment although they did not investigate the quality and quantity of training data in their study. The authors used real data from a 10 month period to train the models with no excitation required and their controller was able to deliver significant energy consumption and comfort violation reductions compared to the reference operation. The question of data quality was also investigated by Jain et al. (2018) who proposed the use of Optimal Experiment Design (OED) for generating the training data. The authors developed a procedure using the dynamical system in closed-loop under limited system availability and operational constraints and found that OED provided a faster learning rate than uniform random sampling or Pseudo-Binary Random Signal (PBRS) sampling, thereby reducing the duration of functional tests by up to 50%.

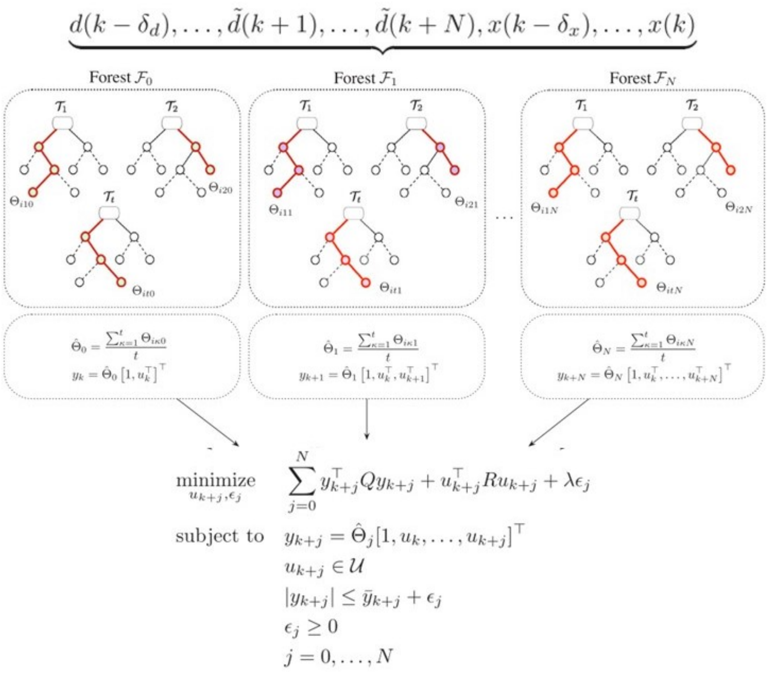

Figure 1: Schematic representation of 'separation of variables' technique using random forests (Smarra et al. (2018))

In existing literature utilising the 'separation of variables' technique, the problem of model identification has not received adequate importance unlike what more traditional model-based techniques such as Resistor-Capacitor (RC) approaches have seen in the literature Prívara et al. (2013). There are also several hyperparameters involved in training the random forest ensemble models that have not been investigated in studies to date. These hyperparameters include the number of regression trees making up the ensemble and the number of data samples in the leaves of each tree. The efficacy of the data-driven controller in building operation, and hence the efficacy of harnessing building energy flexibility, depends upon the accuracy and suitability of the given model. Hence, the objective of the current paper is to test the robustness of these models to a highly heterogenous and diverse building stock and the subsequent controller performance for building energy flexibility objectives in terms of the quality and type of data used for training.

Motivation

Identifying physics-based models, capable of operation for energy flexibility applications, for the diverse building stock comprising of different construction types, varying data availability and subject to changing climate types and occupancy patterns is challenging exercise. Whilst data-driven approaches show promise as a more scalable technique, their robustness and practicality given varying quality and quantity of training data is still yet unproven. To the best knowledge of the authors, this is the first application and test of data-predictive control using the multi-output random forests with 'separation of variables' technique to such a large diverse number of training datasets comprising different climates, levels of missing data, training durations and levels of data excitation. Furthermore, the influence of the various hyperparameters in the model is also studied in detail. This work aims to further enable buildings, from a diverse and varied building stock, to be grid interactive and utilised with an enhanced potential in a demand side management context.

Methods

Building Model and Energy Systems

A modified EnergyPlus model of the US Department of Energy 'Large Office' archetype is utilised as the testbed building for this study (Deru et al. (2011)). This building is 12 storeys high with a floor area of 46,000 m^2. The building has a gas boiler for heating, two water-cooled chillers for cooling and a multi-zone variable air volume system for air distribution. A thermal energy storage (TES) tank of volume 100 m^3 was added to the existing model to include a further flexibility source to the building. The two chillers are operated in parallel configuration; the primary chiller directly meets the cooling load whereas the secondary chiller is used to charge the TES tank. The reference strategy is for the secondary chiller to charge the TES during the unoccupied hours and the TES discharges during the day, thereby reducing the primary chiller cooling load. The building complies with the minimum requirements of ASHRAE Standard 90.1-2004. The EnergyPlus model uses a simulation time-step of 15 minutes and a Rule-Based Controller (RBC) is utilised as the default controller.

Co-simulation Framework

In the absence of a real building, the virtual and high-fidelity EnergyPlus model of the aforementioned archetype building is used to test the data-predictive control in a closed-loop simulation through the use of co-simulation. The PyEp python module is used for communication between EnergyPlus and Python by means of an Open Platform Communications (OPC) bridge as outlined in Jain et al. (2018). Perfect forecasts are assumed and forecast uncertainty was not considered within the scope of this study (the reader is referred to Smarra et al. (2018) for treatment of this issue).

Data-Driven Predictive Control with Ensemble Methods (DPC-En)

The building 'core mid' zone temperature 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, as a result of the symmetry properties of the simulation. Note that the technique of 'separation of variables' has been shown to be suitable for multi-zone buildings and control of multiple zones by Jain (2018). The other response variable of interest is the building total power consumption which is relevant from a DSM perspective. A variation of the DPC-En approach employed by Smarra et al. (2018) is utilised in this study. The predicted electricity consumption at the j^{th} step is given by fitting a regression model on the samples in the leaf with the affine sum of the control inputs being the dependent variable as follows:

$$\hat{Y}_{electricity,j} = \hat{x}(k+j) = \gamma_j[1, u(k), ..., u(k+j-1)]^T \tag{1}$$

Similarly, the predicted temperature at the j^{th} step is as follows:

$$\hat{Y}_{temperature,j} = \hat{x}(k+j) = \alpha_j[1, u(k), ..., u(k+j-1)]^T \tag{2}$$

The γ_j and α_j terms are calculated from the average of the regression coefficients from all the trees of the forest. The reader is referred to Smarra et al. (2018) for an in-depth description of the derivation of the above model.

Similarly to what Bunning et al. (2019) found, the dimensionality of these coefficients is reduced to 2 for each j due to poor prediction performance for larger horizons resulting from the high dimensionality for the model fitting process.
Once the models for describing the building dynamics are found for the pertinent prediction horizon, they can be integrated into a traditional MPC-like receding horizon problem. Given a grid signal $e(t)$, e.g., real-time pricing, the corresponding linear optimisation problem is:

$$\min_{u,\epsilon} \quad \sum_{j=1}^{N} e(t)P_{k+j}^{grid} + \lambda\epsilon_{k+j} \tag{3a}$$

$$\text{subject to} \quad x_{k+j} = \gamma_j[1, u(k), ..., u(k+j-1)]^T, \tag{3b}$$

$$t_{min} - \epsilon_{k+j} \leq x_{k+j} \leq t_{max} + \epsilon_{k+j}, \tag{3c}$$

$$u \in U, \tag{3d}$$

$$\epsilon \geq 0, \tag{3e}$$

$$j = 1, ..., N. \tag{3f}$$

where λ is the weighting term used to adjust the relative cost of comfort constraint violations, ϵ is a slack variable for the comfort constraint, t_{min} and t_{max} are the time-varying zonal temperature constraints, P_k^{grid} defines the power drawn from the grid at time-step k, and N is the prediction horizon. In this study, the cooling setpoint (indoor zone air dry bulb temperature for cooling seasons) is used as the decision variable, as it is easily interpretable and commonly used in building energy management systems as a feedback to end users/occupants. During occupied hours, the temperature constraints are set at $\pm 1°C$ from a reference temperature of $22°C$ and this is relaxed during unoccupied hours to $\pm 5°C$. A sensitivity study was carried out for the value of λ (weighting term for relative cost of meeting thermal comfort constraints) and it was set to 100 for this study. A prediction horizon of 20 steps, or 4 hours was used in this study. The real-time price used in this study was based on actual market data from Italy for 2017 (GME (2017)). It is observed whilst such dynamic pricing structures may not be currently available nor applicable to all electricity markets, it reflects a potential future grid signal. Furthermore, while the market data from Italy may not represent the multitude of different energy markets and climate types that exist, this study focuses purely on the modelling aspect and ability of the controller to react to a dynamic grid signal and hence the grid signal was kept consistent for all the simulations that were carried out.

Sensitivity Study Parameters

Data quality - Excitation

The synthetic data generated from the default RBC control (Excitation = 0) in EnergyPlus does not provide much variation in the value of the cooling setpoint (it is set to $22°C$ during occupied hours and $28°C$ during unoccupied hours for the entire year). It was expected that this would not provide sufficient excitation for training the data-driven models and hence two further scenarios were considered. The first alternative considered was excitation provided using a Pseudo-Binary Random Signal (PBRS) with changes in the cooling setpoint every two hours with control values constrained within the range of $21°C$ to $28°C$ (Excitation = 1). The second alternative considered was synthetic data generated from the data predictive control (DPC) itself to be used as training data (Excitation = 0-DPC). In this case, a preliminary model was trained from RBC data from the month of May and was applied in a co-simulation for a period of one month (June). This data, subject to the cooling setpoint varying according to the optimal predictive control from the preliminary model, was then used for training the final model.

Data quality - Missing data

Missing data was created in the dataset using the Missing Completely at Random (MCAR) approach (Rubin (1976)), where for every given variable, each timestamp has an equal probability of being assigned as a missing value based on the proportion of missing data for each variable. Missing proportions of data investigated were: 5%, 10%, 15%, 20% and 25%. The random forest algorithm is unable to deal with missing values and so a form of imputation is required for these values. As part of the preprocessing, in this study, backward filling (taking the next available value and propagating, i.e. copying, it backwards) is used to impute these values.

Data quantity - Training period and season

The quantity of training data was investigated through evaluating different training periods (1 month, 2 month, 3 month, 6 months) for the algorithm. Further, the season (winter, summer and both) of the training data was also taken into account (note that the controller was tested on a work week in summer).

Climate Type

The following different climate types (ASHRAE climate zones 0-8) were considered for the training and controller analysis:

1. 0A - Darwin, Australia (AUS_NT.D)
2. 1A - Miami, USA (USA_FL_M)
3. 2B - Cairo, Egypt (EGY_Cair)
4. 3A - Rome, Italy (ITA_Rome)
5. 4C - Vancouver, Canada (CAN_BC_V)
6. 5A - Dublin, Ireland (IRL_Dubl)
7. 6B - Datong, China (CHN_Shan)
8. 7 - Tampere, Finland (FIN_Tamp)
9. 8 - Fairbanks, USA (USA_AK_F)

The reference archetype model was simulated for a year for each climate type to generate the synthetic training data (with the different excitation schemes mentioned in the previous section). Note that sizing of the HVAC systems is performed according to the design day conditions of the relevant climate and site location for each of the simulations.

Model hyperparameters

In terms of the random forest model hyperparameters, both the number of trees, n (50, 100, 200, 400) and number of leaves in nodes (25, 50, 100, 200, 400) were investigated as part of this robustness and sensitivity study.

Measured Evaluation Metrics

For each model trained, the predictive accuracy of the models over the prediction horizon was measured on a test set (one work week in July). These models were then employed in a data-driven predictive control setting using the co-simulation environment, again for one work week in July. Various records were taken from each co-simulation run including:

1. average co-simulation step time (s),
2. energy purchased from grid (kWh),
3. energy cost (euros),
4. amount of deviation from comfort constraints (degree-hours).

Results and Discussion

Training Period, Season and Excitation

Figure 2 and Figure 3 show the influence of training period (in months), season of the training data and type of excitation of the data, on the zonal temperature model and power consumption model, respectively. Examining both figures, it can be seen that the influence of excitement and quality of the control data have the most significant impact on the RMSE values. The RBC data (excitation = 0) shows poor model accuracy over the control horizon for both models with RMSE values growing significantly as the N-step ahead prediction horizon increases. Applying PBRS excitation (excitation = 1) results in the most accurate models in most cases. However, in practice, applying such excitation to an occupied building would be considered intrusive and disruptive, not to mention possibly highly sub-optimal in terms of economic and cost perspectives, and hence is considered highly disadvantageous. The DPC data (excitation = 0-DPC) provides a comparable accuracy over the predictive horizon to the highly excited data. Only one month of training data was considered for this scenario (Excitation = 0-DPC). The resulting control performance from the various models is investigated next. Figure 2 and Figure 3 generally show that increasing the training period of data increases the predictive accuracy of the model, e.g., increasing the training period from 1 month to 6 month for the excitation = 1 case in both figures. However, the season seen by the algorithm also makes a significant difference. A model trained using only one month (Period =1) of summer data achieves a much higher accuracy than a model trained using only one month of winter data when tested on a summer month (July) and almost as accurate as a model trained using six months (Period = 6) of data (including both seasons). This could highlight the need for several models based on the season or that the model needs periodic retraining to adjust to climate changes.

Missing Data

Figure 4 shows the predictive accuracy of the power consumption model (6/both/1 model from the previous section) as function of missing data. Small levels of missing data are observed to have a negligible effect on the model accuracy and it is only when significant proportions of missing data (up to 50%) are present that a tangible

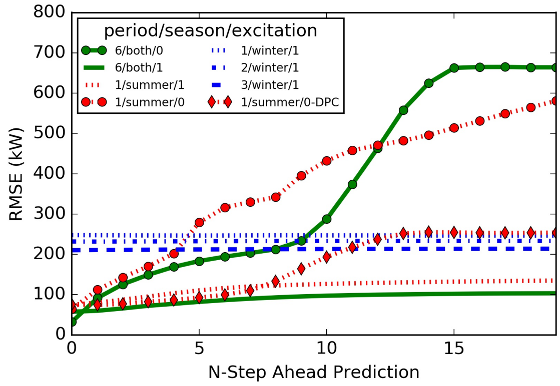

Figure 2: Predictive accuracy of power consumption model as a function of training data period in months/season/excitation

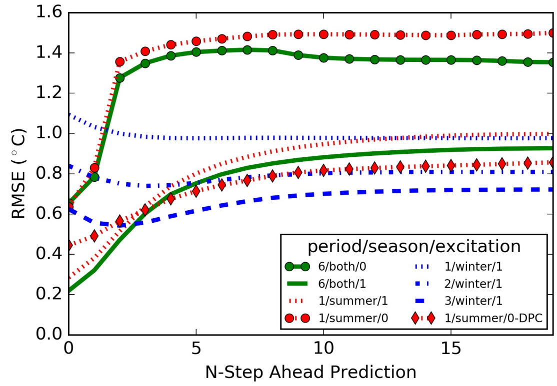

Figure 3: Predictive accuracy of zonal temperature model as a function of training data period in months/season/excitation

effect is evident. Applying the backpropagation technique to impute missing values essentially "smoothes" the data at least for low levels of missing data. Hence, this method is relatively robust to low levels (up to 25%) of missing training data, although it has to be noted that this is at random (MCAR) amongst the input features and long consecutive periods of missing data may have a more significant effect.

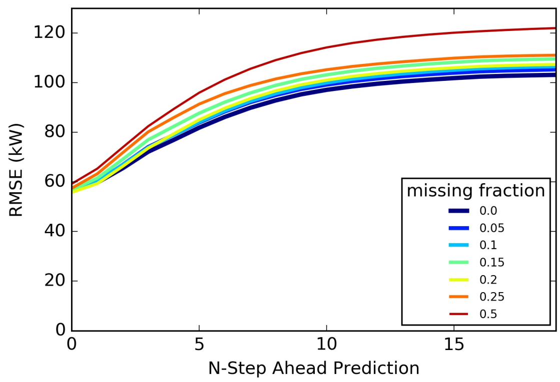

Figure 4: Predictive accuracy of power consumption model as function of proportion of missing data

Model Parameters

Increasing the number of trees leads to small gains in model accuracy as Figure 5 shows. However, a lower number of trees reduces the computational effort of the training and prediction and hence reduces the solve time of the optimisation problem at each time-step in online control. Figure 6 plots the effect of the minimum number of samples at each leaf in each regression tree in the random forest on model accuracy over the prediction horizon. Similarly, there is no significant trend, although in the case of the power consumption model, a sample number of 200 seems to be optimal. When the sample size is set too low, the depth of each tree increases making overfitting to the training data more likely. This leads to the model not generalising as well and loss in performance on new data. When the sample size is set too large, the influence of disturbance variables such as weather may not be captured in their entirety leading to a poor fit on the relationship between the control variable and the output response variables.

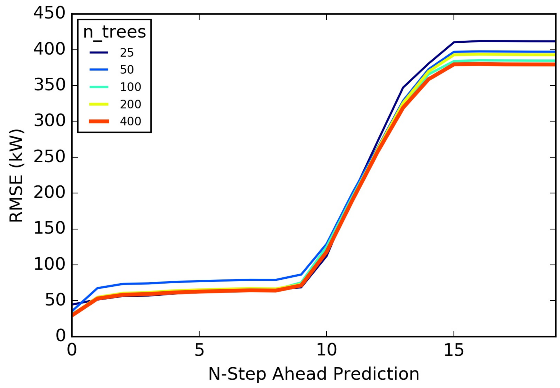

Figure 5: Predictive accuracy of power consumption model as function of number of trees in forest

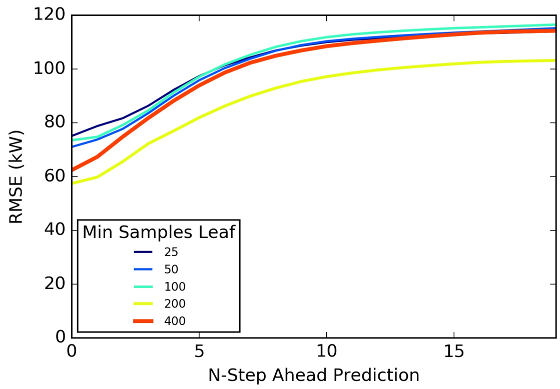

Figure 6: Predictive accuracy of power consumption model as function of number of samples in leaves

Climate Type

Figure 7 plots the predictive accuracy of the zonal temperature model for the different climates investigated. It shows that the models trained on the colder climates generally show a higher accuracy over the prediction horizon compared to the hotter climates for the test week in July. The disturbance features used for training were the same for all models and climates. The weather features utilised in particular were the outdoor drybulb temperature, wetbulb temperature, outdoor wind speed, and outdoor relative humidity. Different weather parameters may be more relevant for different climates, e.g., direct solar radiation for climates with significant solar irradiation. The importance of different climate features for the varying climates will be investigated as future work.

Controller Performance

Figure 8 plots the DPC-En controller performance (energy consumption, economic cost and comfort constraint violations as compared to the reference RBC control) for the different models with varying training periods and season and excitation (the training months/season/excitation notation is used to describe the models in the figure and in this section). The models trained on the RBC data with no excitation performed the worst and actually increased the energy consumption and economic cost compared to the RBC control with the exception of the model trained just on the summer month. It is interesting that for the three models with excitation and trained on winter, even though they seemed to have inferior model accuracy compared to 6/both/1 and 1/summer/1, they have comparable controller performance. Although when comparing the results between 1/winter/1 and 2/winter/1, although 1/winter/1 did not seem to have a significant loss in accuracy for the power consumption model, the controller performance difference is significant, with 1/winter/1 performing

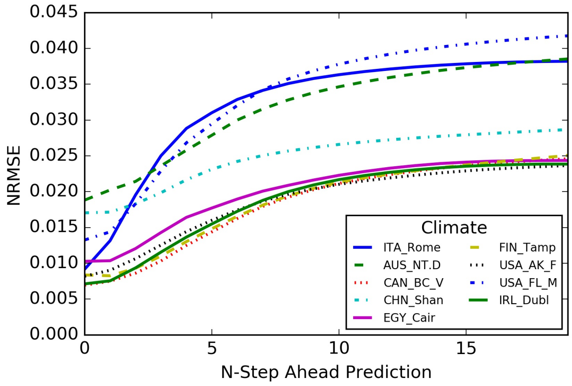

Figure 7: Predictive accuracy of zonal temperature model as a function of climate type

worse than the reference RBC for both energy consumption and cost metrics. In this case, the temperature models explain this discrepancy, with the model trained on 2 months outperforming the one trained only on 1 month.

All other models were capable of outperforming the reference RBC control on all three controller performance criteria. Notably, the controller trained with only one month of the DPC training data achieved the largest reduction in comfort constraint violations and comparable savings for energy consumption and cost, highlighting the promising potential to be grid interactive and harness energy flexibility given a limited quantity of training data. Note that these performance targets can be considered to be an upper bound as the forecast uncertainty for pricing and weather was not considered in this study.

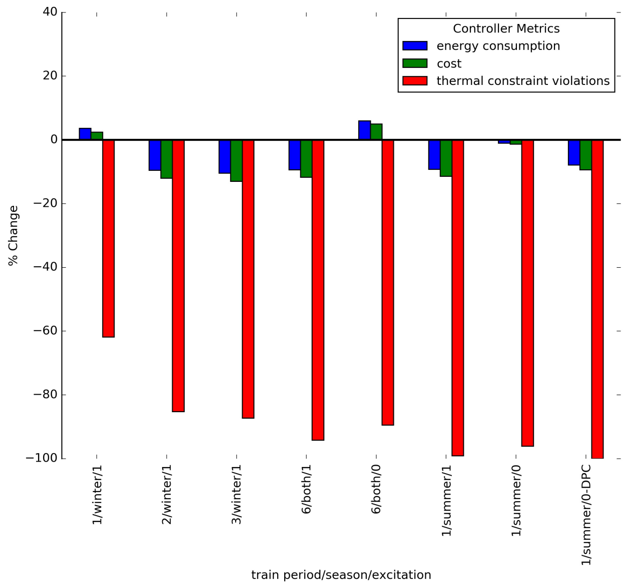

Figure 8: DPC-En controller performance compared to reference RBC control as a function of training data period in months/season/excitation

Figure 9 plots the controller performance (energy consumption and economic cost only) for the different models representing the different climates. Again, the controller performance is compared to the reference RBC. The figure shows varied performance of the controller across the different climates with some climates even reporting a greater energy consumption and cost with the predictive controller. There also seems to be no clear correlation between the predictive accuracy of various models (both zonal temperature and power consumption) and the resultant controller performance. This suggests the need for tuning the controller itself (e.g., comfort cost weighting function) to achieve a satisfactory balance between the different controller objectives. However, given that five of the climates were able to report significant savings with essentially a plug-and-play style controller with no extra feature selection analysis or tuning performed, this is very promising towards the robustness of the approach for generalising over different buildings in different climate contexts.

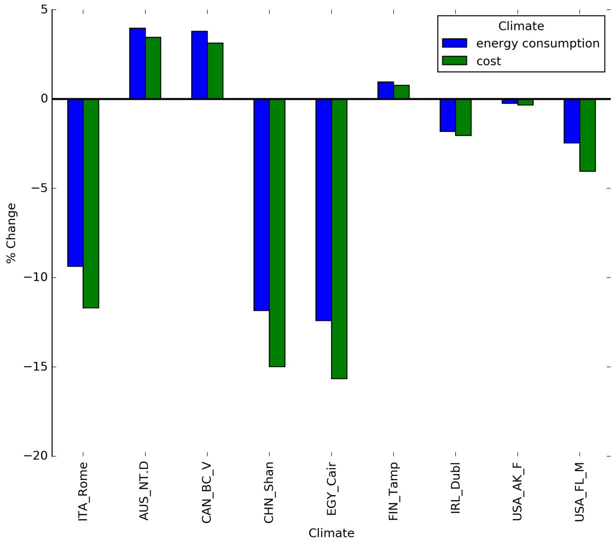

Figure 9: DPC-En controller performance compared to reference RBC control as a function of climate type

Conclusion

The Data Predictive Control (DPC-En) technique using 'separation of variables' together with random forests was shown to be robust to different quantities and quality of training data as well different climate types.
It was shown that it is possible to train this algorithm to achieve satisfactory performance with minimal operational data and without the need for disruptive or performance sacrificing excitation measures. It was further shown that the technique is robust to significant amounts (up to 25%) of missing data. Some level of tuning of the controller objectives is considered to be necessary, however, especially when dealing with a completely new building type or different climate types.
The most significant limitation of this study is the use of synthetic data for training. There have been very few applications of this novel technique on a real case study building. Future work will examine training models using this technique on a number of real-world datasets. Another limitation is that the process of feature selection, especially for weather variables, was not considered in depth for the range of climates studied.

Acknowledgement

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

Behl, M., A. Jain, and R. Mangharam (2016). Data-Driven Modeling, Control and Tools for Cyber-Physical Energy Systems. *2016 ACM/IEEE 7th International Conference on Cyber-Physical Systems, ICCPS 2016 - Proceedings* (April).

Bunning, F., B. Huber, P. Heer, A. Aboudonia, and J. Lygeros (2019). Experimental demonstration of data predictive control for energy optimization and thermal comfort in buildings. *Preprint submitted to Elsevier*.

De Coninck, R., F. Magnusson, J. Åkesson, and L. Helsen (2015). Toolbox for development and validation of grey-box building models for forecasting and control. *Journal of Building Performance Simulation* (July), 1–16.

Deru, M., K. Field, D. Studer, K. Benne, B. Griffith, P. Torcellini, B. Liu, M. Halverson, D. Winiarski, M. Rosenberg, M. Yazdanian, J. Huang, and D. Crawley (2011). U.S. Department of Energy commercial reference building models of the national building stock. Technical Report February 2011, NREL.

Economidou, M., J. Laustsen, P. Ruyssevelt, and D. Staniaszek (2011). Europe's Buildings Under the Microscope. Technical report, Buildings Performance Institute Europe (BPIE).

Gellings, C. (1985). The concept of demand-side management for electric utilities. *Proceedings of the IEEE 73*, 1468–1470.

Gestore Mercati Energetici (2017). Online database. Data retrieved from, `http://www.mercatoelettrico.org/En.`

Henze, G. P. (2013). Model predictive control for buildings: a quantum leap? *Journal of Building Performance Simulation 6*(3), 157–158.

Jain, A. (2018). Data-Driven Model Predictive Control with Regression Trees—An Application to Building Energy Management. *ACM Trans. Cyber-Phys. Syst 2*(21), 1–21.

Jain, A., M. Behl, and R. Mangharam (2016). Data Predictive Control for Building Energy Management. *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments - BuildSys '16* (May), 245–246.

Jain, A., T. Nghiem, M. Morari, and R. Mangharam (2018). Learning and Control Using Gaussian Processes. *Proceedings - 9th ACM/IEEE International Conference on Cyber-Physical Systems, ICCPS 2018*, 140–149.

Jain, A., D. Nong, T. X. Nghiem, and R. Mangharam (2018). Digital Twins for Efficient Modeling and Control of Buildings An Integrated Solution with SCADA Systems. In *2018 Building Performance Modeling Conference and SimBuild, co-organized by ASHRAE and IBPSA-USA Chicago, IL, September 26-28, 2018*, Chicago, IL, USA.

Prívara, S., J. Cigler, Z. Váňa, F. Oldewurtel, C. Sagerschnig, and E. Žáčeková (2013). Building modeling as a crucial part for building predictive control. *Energy and Buildings 56*, 8–22.

Reynders, G., T. Nuytten, and D. Saelens (2013). Potential of structural thermal mass for demand-side management in dwellings. *Building and Environment 64*, 187–199.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Smarra, F., A. Jain, T. D. Rubeis, D. Ambrosini, A. D. Innocenzo, and R. Mangharam (2018). Data-driven model predictive control using random forests for building energy optimization and climate control. *Applied Energy* (February), 1–21.

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.

Sturzenegger, D., D. Gyalistras, V. Semeraro, M. Morari, and R. S. Smith (2014). BRCM Matlab Toolbox: Model generation for model predictive building control. *2014 American Control Conference*, 1063–1069.

Nomenclature

α Leaf coefficients for temperature model

δ_d Autoregressive order for disturbance variables

δ_x Autoregressive order for state variables

ϵ Comfort constraint slack variable

γ Leaf coefficients for power model

$\hat{y}_i$ predicted value of response variable

λ Comfort weighting term

$\overline{y}_i$ mean value of response variable

d Disturbance variables

$e(t)$ Grid signal, e.g. real-time price (Euros/kWh)

j $\in 1, ..., N$ steps in prediction horizon

N Prediction horizon

n Number of trees in ensemble

$P^{grid}(k)$	Power purchased from grid at time-step k
u	Control inputs
X	Measured data - inputs
Y	Measured data - outputs
y_i	actual value/ground truth of response variable