PREDICTING BUILDING ELECTRICITY USAGE AND OCCUPANCY FROM AMBIENT TRANSPORTATION DATA

A PREPRINT

Xuan Li
Carnegie Mellon University
Pittsburgh, PA 15213
xuanli1@andrew.cmu.edu

Xuesong (Pine) Liu
Carnegie Mellon University
Pittsburgh, PA 15213
pine@cmu.edu

Sean Qian
Carnegie Mellon University
Pittsburgh, PA 15213
seanqian@cmu.edu

March 22, 2021

ABSTRACT

The day-to-day operations and schedules of different urban systems are highly interdependent and influenced by system users' motion during different time of day and at specific locations. One such spatial-temporal correlation might exist between building utility usage and ambient traffic condition, which are sequentially determined by individuals' daily mobility. This project explores the potential of the electrical load forecasting in buildings with high-resolution transportation data. Precisely, real-time public transit information, such as travel time and occupancy on bus, congestion status and traffic flow between intersections, are deliberately calibrated to account for potential occupancy flow entering buildings. Integrated with other historical confounding factors, we proposed a framework to quantify the influence of traffic related features at different scale and time of day. Significant improvements are observed during weekends when the electrical loads are more irregular and random. Besides, key roadways and bus routes are identified to have time-varying impact on the electrical status as well.

1 Introduction

According to the survey from EIA[1], 40% of the total energy was consumed in building systems, while more than 60% was attributed to electricity usage. A precise prediction and forecast of building utility demand, which is highly correlated with building occupancy level, has the potential to improve energy efficiency on multiple scales. For individual buildings, a timely prediction of utility or occupancy can prevent over-conditioning or under-conditioning operations for systems including Heating, Ventilation, Air-Conditioning (HVAC) and lighting. On the other hand, when block of buildings are aggregated as microgrid, the forecasting of building demand with high precision is essential for decision makers on the grid to optimize power generation and transmission.

In this EAGER project the high-resolution real-time data from transportation are used to predict electricity and occupancy for buildings up to one hour ahead. As illustrated in Figure 1, the rationale lies in the spatial-temporal correlations exist between those two urban systems. Specifically, the precedent status of ambient traffic flows are correlated with the occupancy flow towards the buildings, which further determines the level of energy consumption. The discovery of these patterns, would bring new horizons for cross-system demand predictions and management in an effective and efficient manner. Thus, a data-driven approach is proposed and implemented in this research to investigate interrelationships between transportation and building systems. Specifically, the marginal improvement of building electricity and occupancy prediction is justified by the adopting of high-resolution transportation features. The performance would be evaluated in terms of accuracy as well as consistency. The report is constructed as follows: the necessity of building utility prediction, current approaches, gaps are illustrated in Section 'Background'; the proposed approach is introduced in Section 'Method'; the description of data, and pre-processing is explained in Section 'Data Source and Processing'; the result and analysis are given in Section 'Experiment'; and finally the major findings and future work are listed in Section 'Conclusion'.

[1] U.S. Energy Information Administration

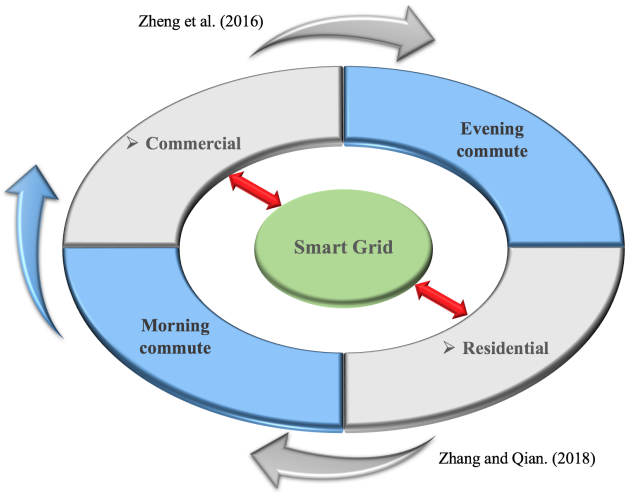

Figure 1: Spatial-Temporal Relationships between Infrastructure, Transportation and Grid

2 Background

2.1 Benefits of electricity prediction at building level

The prediction of of building electricity consumption as well as occupancy status has incentives for energy efficiency at different levels. In the first place, it is the right information that facilitates effective communications between grid operators and building managers regarding on negotiation of supply and demand. Electric load forecasting (prediction) has significant impact on the operations for power systems at different time scales. Based on the length of forecasting horizon, it can be categorized into short-term (one hour or one week ahead), medium-term (one week up to one year), and long-term (multiple years) forecasting. Basically, medium-term and long-term prediction are essential for equipment maintenance and fuel supply planning [1]. On the other hand, since electricity cannot itself be stored, an accurate short-term load prediction can not only reduce the costs of over or under power generation, but also help to make optimal decision such as fuel purchasing and resource allocation [2, 3, 1, 4]. An accurate load prediction on lower level (feeder, buildings) would also bring benefit to power operation on the grid as well. As pointed out in [5], the participation of small customers in demand side is increasing due to the growing adoption of smart meters, home energy management and automation system. This potential has been emphasized by the fact that an increasing proportion of electricity are generated using stochastic and intermittent renewable energy resources such as wind and power. However, a proper utilization of such resources remains the domain challenge for the grid due to its significant impact on the stability of the power systems [6]. Accordingly, an active participation of buildings on can help to smooth out the fluctuating loads exist between demand and supply. Basically, forecasting made on this level that accounts for neighboring buildings, i.e., a whole university campus as a microgrid, can help decision makers in grid to optimizes the daily generation and distribution of electricity utility, which leads to demand response strategies for downstream consumers to customize their operation for economical incentive. On the contrary, an anticipation of electricity profile for individual building provides the baseline that helps facility managers to evaluate the possibility of participating a specific demand response incentive. ref and discussion on buildings to participate in demand response

(discussion on the importance of variations in prediction)

MMC. In this project, however, the performance of the the proposed model enhanced with transportation features would be evaluated in terms of accuracy as well consistency. Specifically, we will compare and analyze the maximum variations at peculiar time of day and day of week with respect to the baseline model.

Besides the aforementioned benefits, given the observed correlation between building electricity and occupancy [7, 8, 9, 10], an improved prediction of building electric load can further enhance the estimation of building level occupancy, which in return improves energy efficiency in building HVAC systems by reducing unnecessary conditioning during actual unoccupied periods. As summarized in [10], 40% to 60% of the total electricity consumption in campus and office buildings occurs during non-working hours, while 15%–40% of the total building energy consumption can be saved with intelligent occupancy sensing and prediction. However, building level occupancy is not adequately considered in current HVAC control. For example, the volume of fresh (outdoor) air in Air Handling Unit (AHU) are determined by the ratio of occupancy in buildings. Due to a lack of perspective for actual occupants, however, a fixed

occupancy schedule is used to determine minimum outdoor air flow, which requires extra conditioning from the system, especially during hot summer and cold winter. Additionally, knowing the occupancy in advance can help to adjust deadband and setpoint during occupied period for temperature control, peculiarly in weekends when occupancy patterns are more irregular and random. The relaxing the deadband for even 1 $^\circ F$ has huge potential for energy saving. As showed in Table 1, by running a year-long simulations for small, medium and large commercial building across eight different cities in the U.S. using Energyplus, it is found that 1 $^\circ F$ increase in zone heating/cooling deadband would lead to 10% total energy saving in HVAC systems. The similar observations have been reported in [11, 12, 13, 14] as well.

Table 1: yearly energy saving under varying deadband

deadband/building	small	medium	large	**average**
4	12.28%	10.86%	11.51%	11.55%
5	23.13%	18.67%	19.05%	20.28%
6	32.80%	24.27%	23.89%	26.99%

The potential of electricity as well as occupancy prediction is not limited to the practice control of HVAC system. Essentially, it also facilitate the implementation of model predictive control (MPC), whose superior performance has been verified in simulations and field experiment. As proposed by DOE, MPC is a promising strategy to attain the 10% energy reduction by 2030. To ameliorate the inefficiency of current rule-based and reactive control methodology which lacks long-term and global design, MPC seeks to optimize the operation in an adaptive manner with the projection of various factors including occupancy, internal thermal loads, outdoor weather condition. The superior performance of MPC has been validated both in field experiments [15, 16, 17, 18] as well as in simulation verifications [19, 20, 21, 22], with the potential of energy saving ranging from 17% to 50%. Above all, the predictions of electricity and occupancy are the essential components for MPC design. On one hand a proportion of the predicted electric load directly accounts for the internal heat gain (through lighting, plugin, equipment) that MPC needs to include for zone level modeling. On the other hand, the prediction of occupancy contributes to several aspect of MPC as well, including the level of occupant induced internal load, range of zone temperature deadband as constraints.

2.2 Current approaches

The benefit of electricity prediction has involved a number of researchers in the past several decades to implement sophisticated forecasting models with a rich amount of features. Although a relatively low MAPE (mean average percentage error) of $1\% - 3\%$ on grid level can be attained for next day load prediction [4], however, a marginal improvement can still lead to significant economical benefit. As reported by Bunn and Farmer [23] in 1985, an increase of 1% in forecasting error would lead to 10 million pounds expense per year for an electric utility company in UK. Motivated to further improve prediction accuracy, a bunch of different approaches have been proposed and implement for short-term load forecasting (STLF), including time series models (ARMA, SARIMA), regression model (linear, kernel), state-space model with Kalman Filter, and black-box models such as neural networks. However, as pointed out in [3], no sophisticated method has been proved to be clearly better than others. Despite the variations in forecasting models, the following features are commonly selected and adopted by most of the models, including

- weather related: hourly temperature, wind speed, wind chill temperature [2]; THI (temperature-humidity index), WCI (wind chill index) [4]
- time and date related: day of week, hour of the day; weekday, weekends
- historical load related: hourly load for previous hour, the previous date, same day of the previous week [1]
- class of sector related: commercial or residential or industrial

While most of the forecasting approaches work well on city or country scale, the performance can degrade as we shift the horizon to building level. In [2, 5], it is found that the MAPE would increase as number of meters aggregated decrease. reference on this aspect

2.3 Rationale for transportation features

Motivated by the above observations, there still exist a need to improve on the accuracy of building load forecasting, especially at building levels. However, given the state-of-the-art performance is saturated by the choice of available features and approaches, the inclusion of other interdependent covariates, such as ambient transportation status, might potentially make advance in a different perspective. Essentially, in the project we explored whether features from transportation, a highly interactive system with buildings, can be adopted to further refine load prediction accuracy.

With available electric meters and ambient transportation information on campus, we aim to build the a prediction model on feeder level, aiming to discovering how much improvement we can make given the temporal-spatial correlated traffic features. Also, as traditional demand response is transitioning towards dynamic demand response that needs the prediction at hours ahead [5], in this paper we restrict the prediction horizon within an hour. The intuition is that infrastructures and transportation are highly spatial-temporal correlated systems given the intermediary activity of users. For example, people travel from home to office back and force in the morning and evening according to schedules and event. Most of the time daily electricity is highly predictable using historical information since it is operated according to fixed schedule. For campus buildings the occupancy patterns are more influenced by semester or class schedules, which indicates the deterministic nature of the accurate prediction based on historical observations. However, the clear patterns become vague especially during weekend, which might not be captured well by historical features solely. Using external features, such as ambient traffic, might account for this part and further enhance the performance. Inspired by the promising performance of traffic prediction with building information investigated in [24, 25], the inverse relationship would be investigated in this paper for the first time.

In short, as showed in Figure 2, the inclusion of transportation features in prediction can potentially factor out the uncertainty introduced in the set of confounding variables. Assume y, x, $x^\star$ represents the response, historical variables and traffic related variable. Showed in 1, ideally the inclusion would reduce the original uncertainty ϵ, such that $\epsilon^\star \leq \epsilon$.

$$y = f(x) + \epsilon; \qquad \Longrightarrow \qquad y = f(x^\star, x) + \epsilon^\star \tag{1}$$

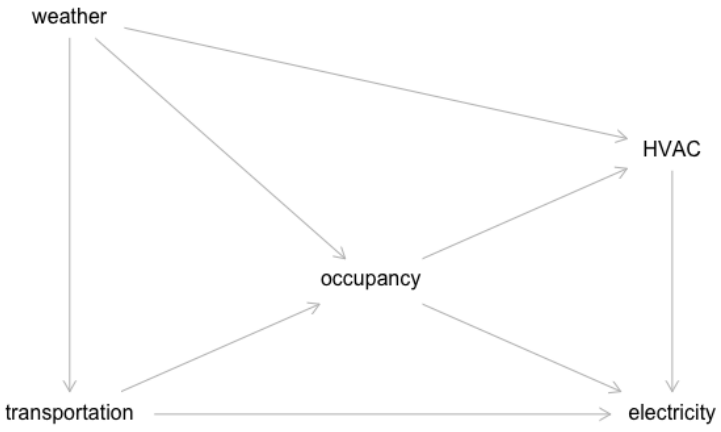

Figure 2: casual relationships between transportation and electricity

Although using historical data itself already achieves a very low prediction error, during data processing there exist a great proportion of meter reading error, which may lead to misleading predictions. However, external features such as ambient traffic conditions can be the backup feature sets if the above circumstance happens, which improves the robustness of the prediction.

3 Method

3.1 linear regression model for electrical load prediction

Since no sophisticated method has been proved to be clearly than others for electrical load prediction, in this project we restrict our main emphasis on studying the influence of transportation factors and choose to use linear regression model enhanced with feature selection. The available data at hand would be equally splitted into training and testing set. An optimal subset of features would be selected using five cross-validation in terms of lasso regression on the training set. The parameter of the model would be fitted with the selected features on the training set using ordinary linear regression, and finally the performance would be reported on the testing set. The advantages of this approach over other more complicated methods are two folds. Firstly, the p-value for each of the features provides an inference to justify the significance of the features. On the other hand, the sign and magnitude of each coefficient allow the quantification of the impact from each features included in the model.

In general, to predict the electrical load for building at time t, there are three types of features that are considered in our model: the historical electrical load X_h, the ambient traffic status X_m at $t - n\Delta t$, and the historical weather conditions X_w. $n\Delta t$ is the prediction horizon, and Δt is sampling rate of the features. The regression model is showed in Equation 2.

$$Y(t) = \sum_{i=1}^{N_h} \beta_i \cdot X_h(t - (i+n)\Delta t) + \sum_{j=1}^{M} \beta_j \cdot X_{m,j}(t - n\Delta t) + \sum_{k=1}^{N_w} \beta_k \cdot X_w(t - (k+n)\Delta t) \tag{2}$$

4 Data Source, Preprocessing

4.1 Building Energy Dataset

The building electrical consumption data are retrieved from the electrical meters maintained by the Facility Management Services (FMS) department of Carnegie Mellon University, including all academic and official buildings on main campus at Carnegie Mellon University (CMU). The data are queried from 01/01/2017 07/31/2019, collected by the electricity meters (cumulative) in each building. After consulting with the facility managers on campus, we have confirmed that the meters account for all electricity consumed within the building, including lighting plug-in load, electricity that drives AHU-VAV system, etc. However, what these meters do not account for is the energy required to generate and circulate chilled water/steam that is provided to the building from the campus central plant. The original data were randomly sampled and recorded in different interval scale, ranging from every 5 minute to 20 minute. Since we are more interested in evenly sampled data, we interpolated from the raw data at 15 minutes interval, which is a commonly scale used in building energy prediction.

Initially we sampled all the data for 20 main buildings on CMU campus, However, we have observed three types of data quality issue that restricted our scope to 6 buildings in this project. One of the most commonly seen issue is missing data, that at some time slot no numerical value was measured. Although the data as marked as 'I/O Timeout', 'No data', 'Failed', the missing data might last for days or weeks, which makes it difficult to apply missing data imputation methods as well.

4.1.1 Data visualization

To have a better sense of the pattern for the data we use the processed electricity consumption of Gates Hillman Center (GHC) in year 2017 for illustration. The general patterns are showed in Figure 3, where the upper plot covers full year and the bottom covers four weeks of January. It is observed there exist a distinct daily pattern for weekdays, where the peak usage happens around noon most of the time. On the other hand, the range of data is relative constant, which indicates that seasonality might not be an issue for our analysis.

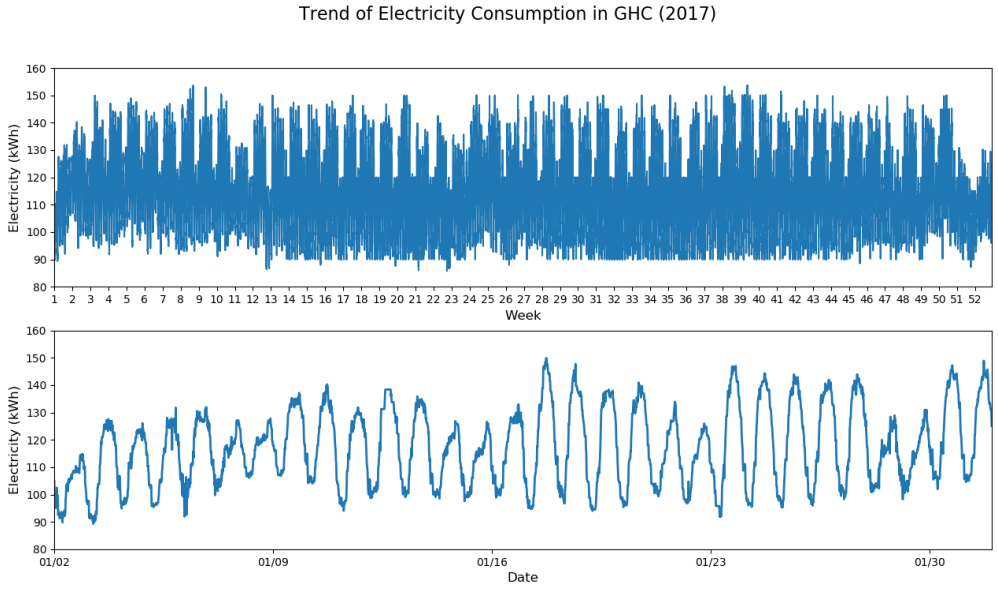

Figure 3: GHC Energy Patterns

To further analyze the weekday patterns we segment the yearly data according to weekday index, and compute the correlation coefficients, which is show in Figure 4. The results reflect there exist high correlations within weekdays, while for weekends the correlation is less significant. The above observation motivates us to come up with different model for weekday and weekends in later analysis.

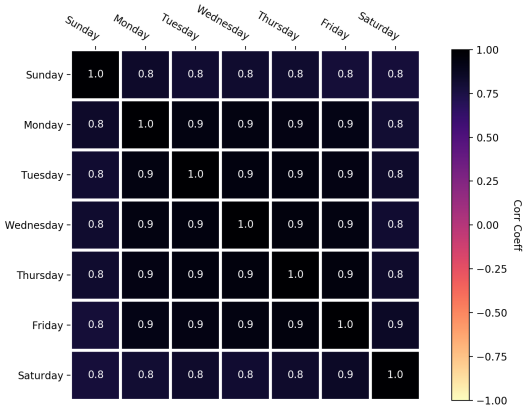

Figure 4: GHC day of week correlation

4.1.2 Feature processing

The response $y(t)$ and historical features X_h are generated from the processed building electrical loads. Two types of historical features are included in our model: the past N_h electrical loads up to $t - n\Delta t$, and past N_d weeks observation at the same time of day and time of week.

4.2 Transportation: public bus

The original data are downloaded from Port Authority[2] of Pitsburgh. The arrival time, leaving time, number of people boarding, leaving and remaining at each bus stop are recorded for each bus route at in-bound and off-bound direction.

[2] `https://www.portauthority.org/`

Also, the longitude and latitude of each bus stop is provided as well. In this project we focus on twelve bus routes that have bus stop ambient to the CMU main campus. Using the closest bus stop as reference, the average travel time from each bus stop to campus is estimated for each hour, which will be used to determined the potential travel time to campus. Finally, given $n\Delta t$ as constraint, the occupancy information for each route the bus arrived during $[n\Delta t - \eta, n\Delta t + \eta]$ would be selected and aggregated as potential occupancy heading campus. A sample outlook of the features for one of the bus route is showed in Figure 5 with $n\Delta t = 15$min.

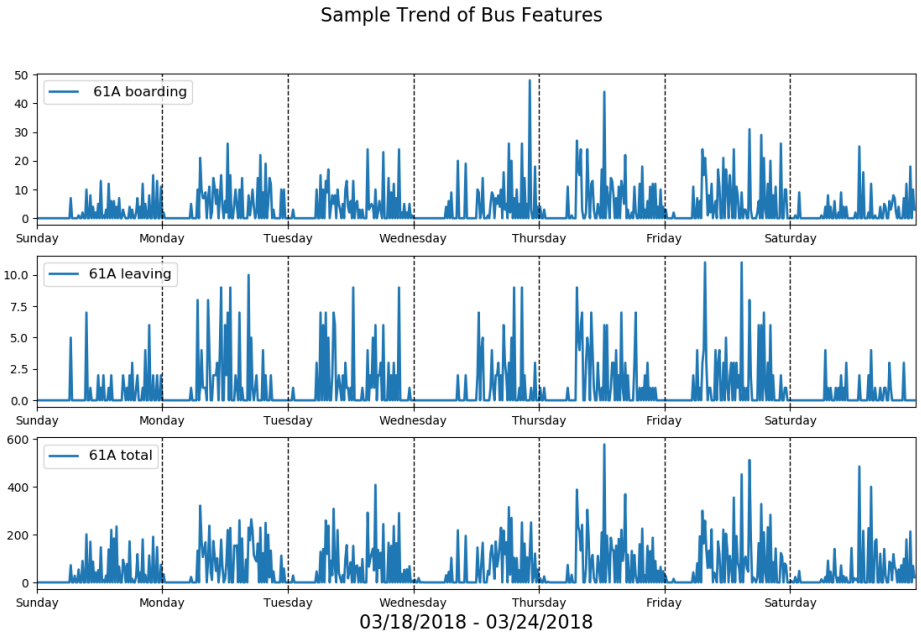

Figure 5: Sample Trend of Processed Bus Features

4.3 Transportation: inrix

The second set of transportation features are obtained from INRIX[3], where the high-resolution data sampled in one minute interval are provided. Basically, the dataset contains information of average speed and confidence score[4] of the traffic flow for both directions between two intersections of roadways. According to [26], the confidence score is denoted as 30 (report based on real-time data), 20 (report based on real-time data across multiple segments), 10 (report based primarily on historical data). Selecting the 72 roadways that are within five miles perisphere of CMU campus, we use the average speed, confidence score, local travel time (distance between the intersections / average speed), and estimated travel time (distance from the mid point of intersections to campus / average speed) at $t - n\Delta t$ as features for prediction at time t. A sample trend of the features on 'fifth and Shady Avenue, Eastbound' is showed in Figure 6.

[3] `http://inrix.com/`

[4] can be regarded as the volume of car on the road

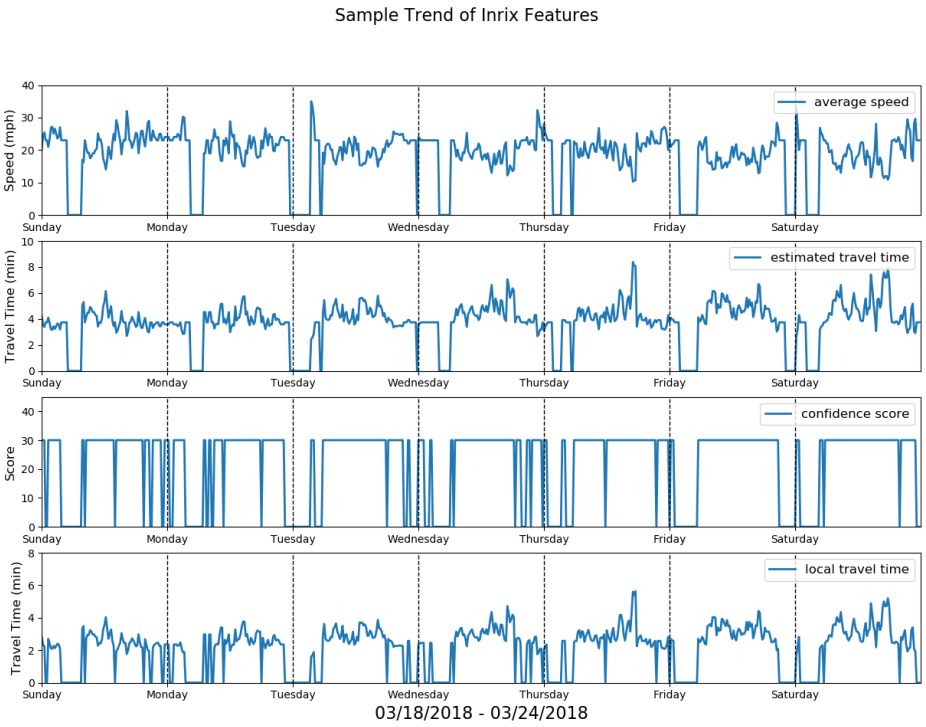

Figure 6: Sample Trend of Processed INRIX Features

4.4 Weather

The weather data was downloaded from Pennsylvania State Climatologist[5], recorded at hourly scale in Allegheny County. The features we choose are *average Temperature*, *average relative humidity*, *average wind speed*, which are linearly interpolated to 15 minute scale as well. As a consequence, the past N_w observations up to $t - n\Delta t$ are considered in the regression model.

4.5 Building Occupancy

For the second part of the analysis we investigated the impact of traffic features on the prediction of building occupancy. Due to the lack of occupancy sensors adopted on campus, however, we were only able to have access to the occupancy level of one building (GHC) on campus, with the reading recorded during two consecutive periods: from 9/15/2017 to 10/15/2017, and from 4/17/2019 to 6/14/2019. Basically, the information is extracted from PIR sensors for each of the 301 offices in GHC, with 'occupied/unoccupied' status recorded in the data log. Thus, the ratio of offices that are occupied at each timestamp can be obtained, which is defined as occupancy ratio in this analysis. The daily trend of the occupancy ratio and corresponding electricity loads are showed in Figure 7.

[5] `http://climate.met.psu.edu/`

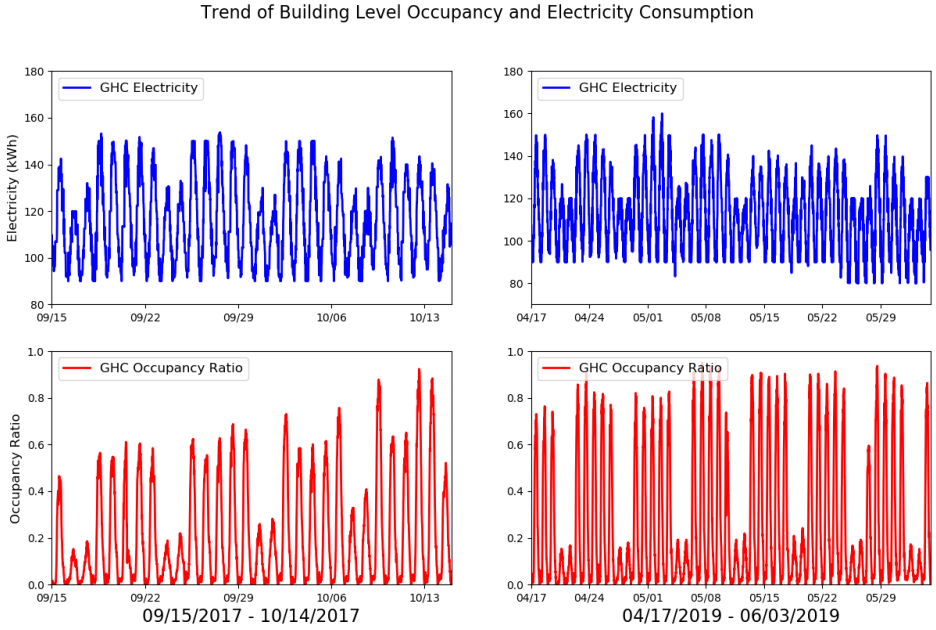

Figure 7: Trend of occupancy and electricity in GHC

5 Experiment and Results

The performance was analyzed for different time of day, time of week, number of buildings and prediction horizon in this project. Also, the performance is also compared with two kind of baseline models: the one without traffic features, and the one using nonlinear regression models such as random forest and multi-layer perception. The available data in 2017 as used as training set, and data in 2018 are used as testing set. Mean squared error (MSE), mean average percentage error (MAPE), and R^2 are used as evaluation metric for performance comparison. In specific, the general procedure for feature selection, model training, and prediction inference is summarized as follows:

- ten fold cross-validation is implemented on the training set to select the optimal α term in Lasso in each fold, which is marked as $\alpha^{\star}$[6]. Eventually, $\alpha^{\star}$ is applied to the full training set to select a subset of features $X^{\star}$.
- An ordinary linear regression model is fitted on the training set, with $X^{\star}$ as covariates and Y as response. The set of fitted parameters are marked as $\beta^{\star}$.
- The performance of the model is evaluated on the testing set, using the aforementioned metrics.

5.1 Performance for Monday and Friday ($\Delta t = 15$)

5.1.1 general performance

The results include the observation from Monday to Thursday, from 6:00 am to 7:00 pm. We segment the data into 4 periods, 6:00 - 9:00, 9:00 - 13:00, 13:00 - 16:00, and 16:00 - 19:00, based on the assumption that the pattern may vary according to different time of day. The results for prediction horizon as 15 minutes are listed in Table 2 below, with the electricity summed over all six buildings. It is observed that in general the inclusion of transportation features outperforms the baseline model. However, as the pattern of electricity during weekdays are more affected by academic schedules, the proposed method can only slightly improve the proportion explainable variances.

[6]roughly, for weekday $\alpha^{\star}$ is 4.0 and for weekend it is 3.0

Table 2: Model Comparison, **Linear**, weekday (Monday + Friday), $n\Delta t = 15$

Period	**baseline, OLS**			**Selected, LASSO**		
	MSE	MAPE (%)	R^2	MSE	MAPE (%)	R^2
6:00 - 9:00	82.96	1.43	0.86	57.99	1.26	0.90
9:00 - 13:00	59.09	1.07	0.92	47.91	0.95	0.94
13:00 - 16:00	48.61	0.96	0.93	37.48	0.85	0.95
16:00 - 19:00	65.91	1.15	0.93	40.44	0.92	0.96

Secondly, the performance nonlinear regression methods using the selected features are reported in Table 3. It is observed that for all of the time period the proposed method can outperform the more complicated regression method in terms of the three evaluation metrics. On the other hand, it partially justify the suitability of using linear regression model.

Table 3: Model Comparison, **Nonlinear**, weekday (Monday, Friday), $n\Delta t = 15$

Period	**Random Forest**			**Multilayer Perception**		
	MSE	MAPE (%)	R^2	MSE	MAPE (%)	R^2
6:00 - 9:00	56.54	1.22	0.90	147.68	2.05	0.75
9:00 - 13:00	84.62	1.18	0.89	137.94	1.66	0.74
13:00 - 16:00	57.54	0.98	0.92	72.83	1.20	0.90
16:00 - 19:00	45.66	1.00	0.95	80.68	1.36	0.91

5.1.2 significant features

We further narrow down our investigation on the impact of the traffic features in the proposed model. The correlation, coefficient, and p-value for the selected features for 6:00 - 9:00 are listed in Figure 8 and 9 below. It is observed the dominant factors are the historical electricity consumption. On the other hand, total occupants on bus, confidence score and average speed from the INRIX feature set are important factors selected by the model. On the contrary, however, the weather features set are seldom selected by the model.

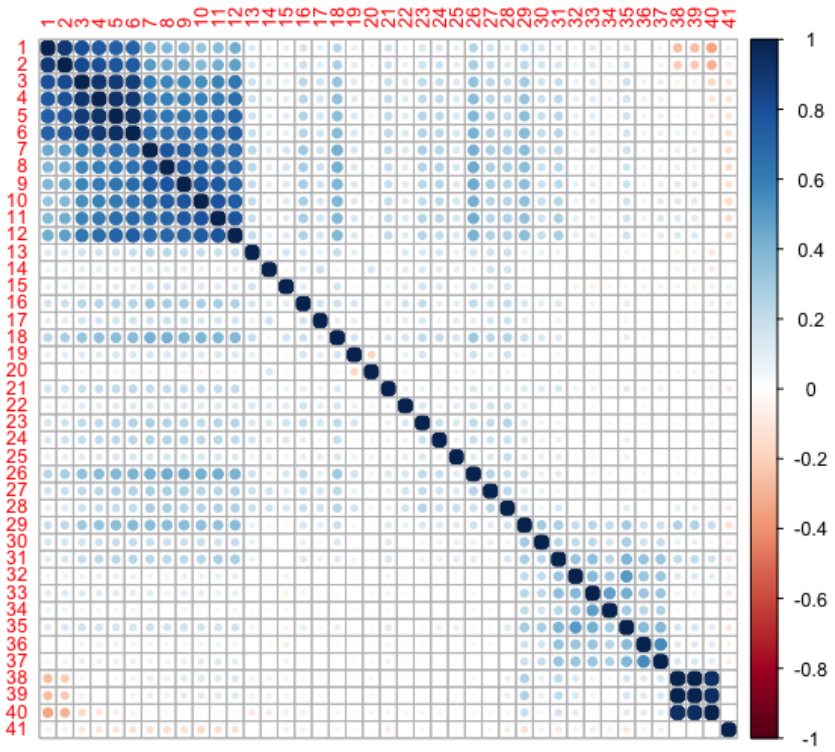

Figure 8: Feature correlations, Monday and Friday, 6:00 - 9:00, $\Delta t = 15$

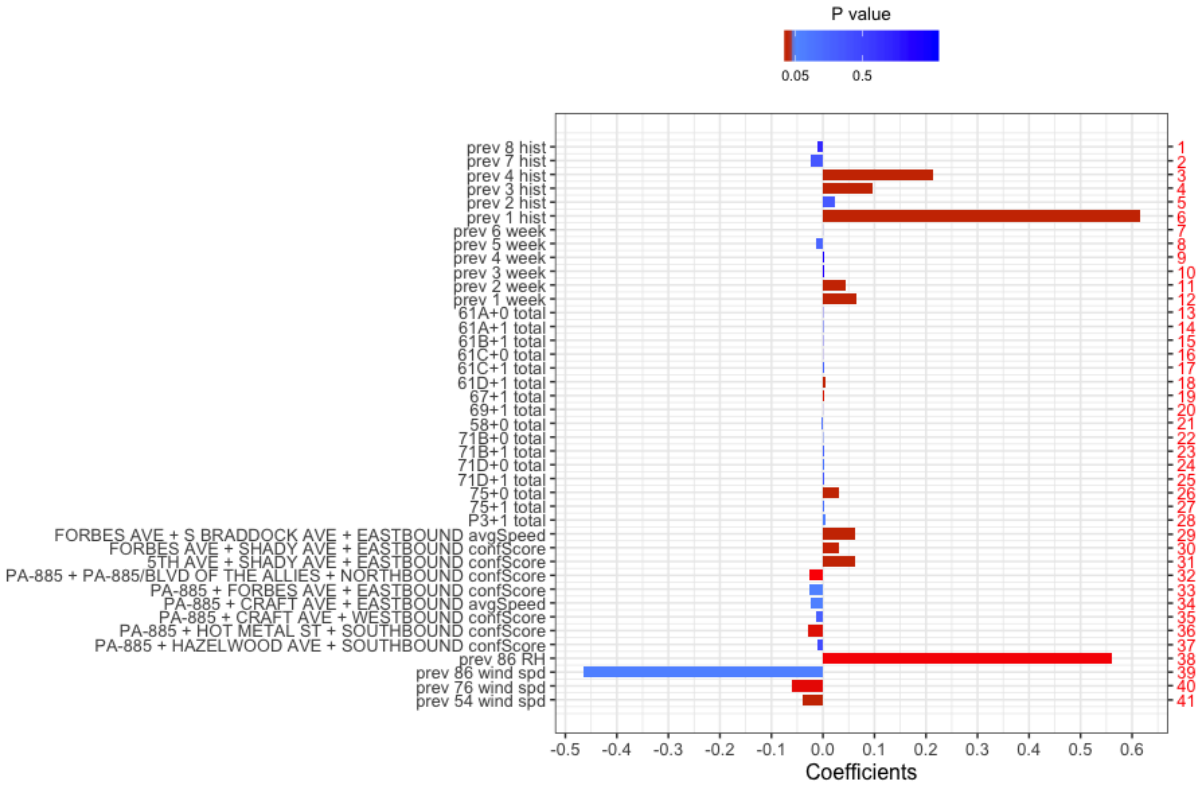

Figure 9: Coefficients and P value, Monday and Friday, 6:00 - 9:00, $\Delta t = 15$

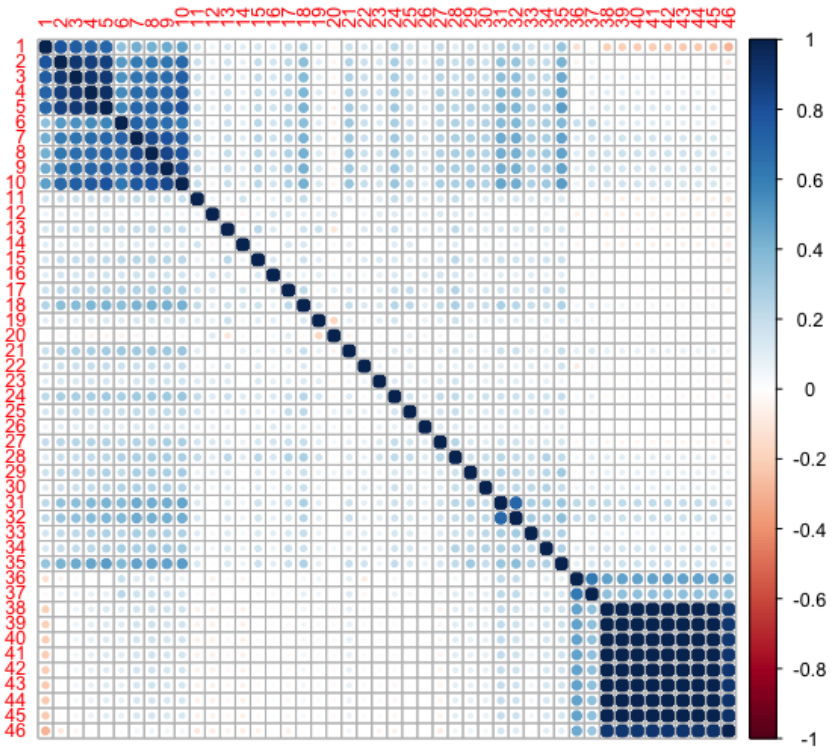

Figure 10: Feature correlations, Tuesday - Thursday, 6:00 - 9:00, $\Delta t = 15$

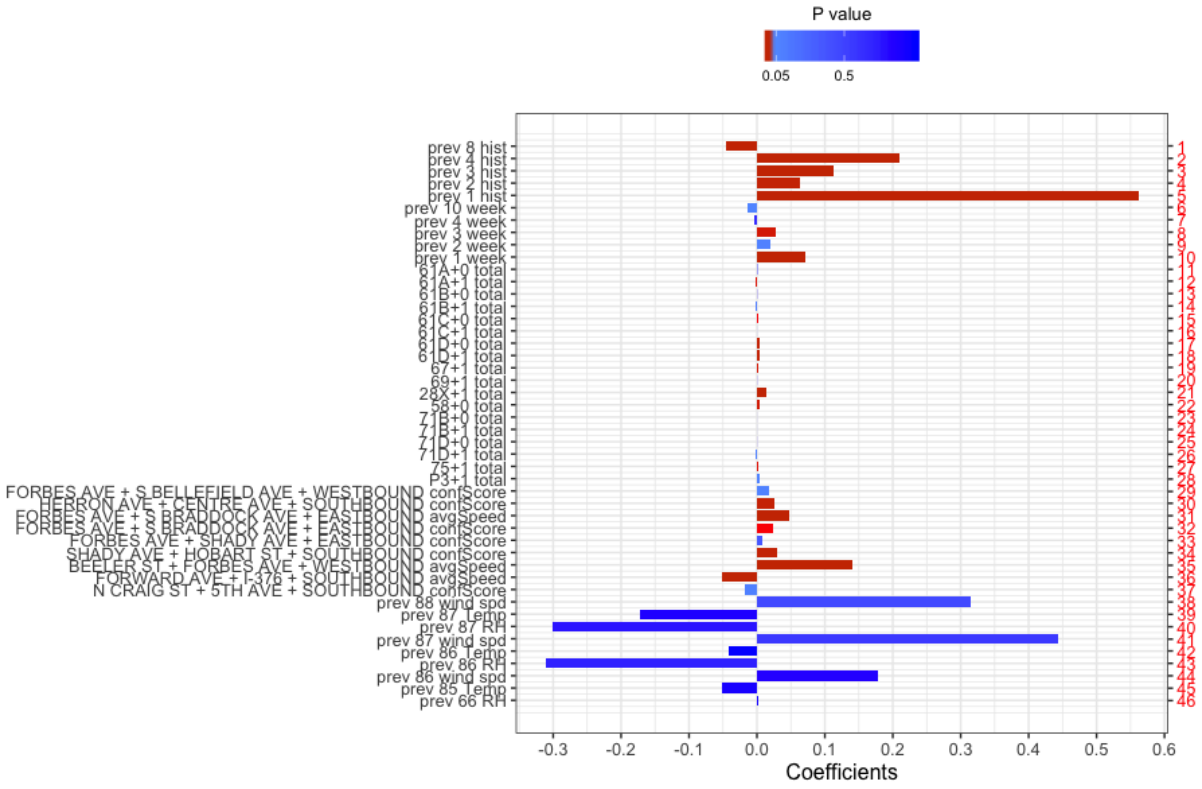

Figure 11: Coefficients and P value, Tuesday - Thursday, 6:00 - 9:00, $\Delta t = 15$

5.2 Performance on (Tuesday-Thursday) ($\Delta t = 15$)

Table 4: Model Comparison, **Linear**, weekday (Tuesday, Wednesday, Thursday), $n\Delta t = 15$

Period	**baseline, OLS**			**Selected, LASSO**		
	MSE	MAPE (%)	R^2	MSE	MAPE (%)	R^2
6:00 - 9:00	70.91	1.36	0.88	56.85	1.23	0.91
9:00 - 13:00	51.22	0.99	0.93	43.42	0.92	0.94
13:00 - 16:00	53.56	0.95	0.93	41.18	0.83	0.95
16:00 - 19:00	49.10	0.99	0.95	40.61	0.91	0.96

Table 5: Model Comparison, **Nonlinear**, weekday (Tuesday, Wednesday, Thursday), $n\Delta t = 15$

Period	**Random Forest**			**Multilayer Perception**		
	MSE	MAPE (%)	R^2	MSE	MAPE (%)	R^2
6:00 - 9:00	61.18	1.27	0.89	141.10	1.83	0.77
9:00 - 13:00	55.38	1.04	0.92	133.93	1.69	0.83
13:00 - 16:00	67.31	0.99	0.91	97.53	1.39	0.87
16:00 - 19:00	45.92	0.96	0.95	69.24	1.20	0.93

5.3 Performance on (weekends) ($\Delta t = 15$)

5.3.1 general performance

The same kind of analysis is conducted for weekends cases as well. The comparison with the two kind of baseline model are reported in Table 6 and 7. Significant improvements of using traffic features have been observed in both of

the cases. A possible explanation is that the pattern of electrical load during weekends are more stochastic so that only using historical features is not sufficient as in the weekday cases.

Table 6: Model Comparison, **Linear**, weekends, $n\Delta t = 15$

Period	**baseline, OLS**			**Selected, LASSO**		
	MSE	MAPE (%)	R^2	MSE	MAPE (%)	R^2
6:00 - 9:00	46.73	1.16	0.66	23.26	0.86	0.83
9:00 - 13:00	46.64	1.14	0.77	27.52	0.87	0.87
13:00 - 16:00	41.13	1.01	0.81	25.04	0.80	0.89
16:00 - 19:00	54.40	1.14	0.77	28.77	0.86	0.88

Table 7: Model Comparison, **Nonlinear**, weekends, $n\Delta t = 15$

Period	**Random Forest**			**Multilayer Perception**		
	MSE	MAPE (%)	R^2	MSE	MAPE (%)	R^2
6:00 - 9:00	34.85	1.01	0.75	62.83	1.37	0.54
9:00 - 13:00	29.89	0.91	0.86	87.41	1.61	0.58
13:00 - 16:00	26.46	0.81	0.88	106.30	1.66	0.52
16:00 - 19:00	29.06	0.85	0.88	64.09	1.27	0.73

We also performed an analysis of the residual, which is showed in Figure 12. The normally distributed residuals justifies the appropriateness of linear regression model in this case.

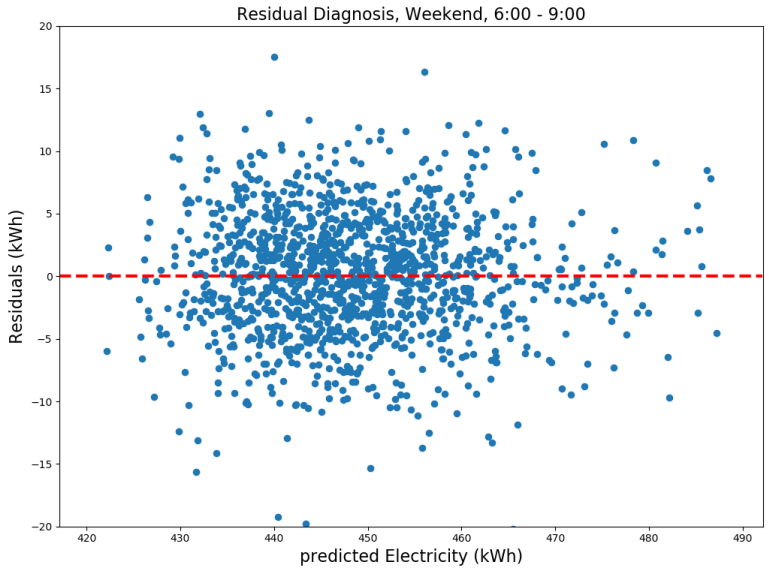

Figure 12: Residual Diagnosis, weekend, 6:00 - 9:00, $\Delta t = 15$

5.3.2 significant features

The correlation between selected features, coefficients and corresponding p-values for 6:00 - 9:00 are reported in Figure 13 and 14. Compared to weekdays, more features from bus and inrix feature set have been selected by the model. Furthermore, the set of transportation features have a larger impact given the increased magnitude in coefficients. Finally, compared to weekday models, it is observed that more weather related features have been chosen as well.

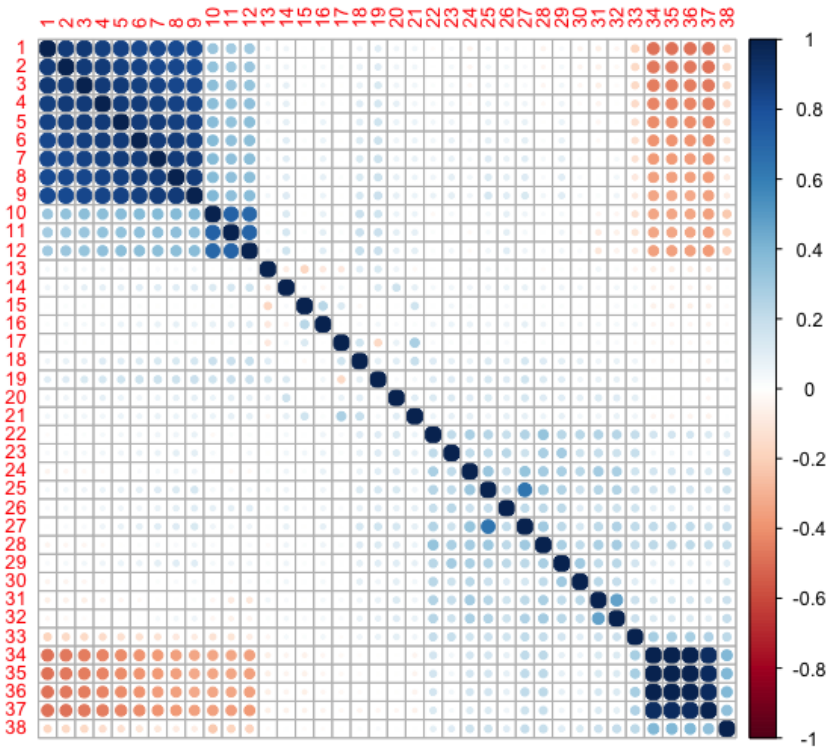

Figure 13: Feature correlations, weekend, 6:00 - 9:00, $\Delta t = 15$

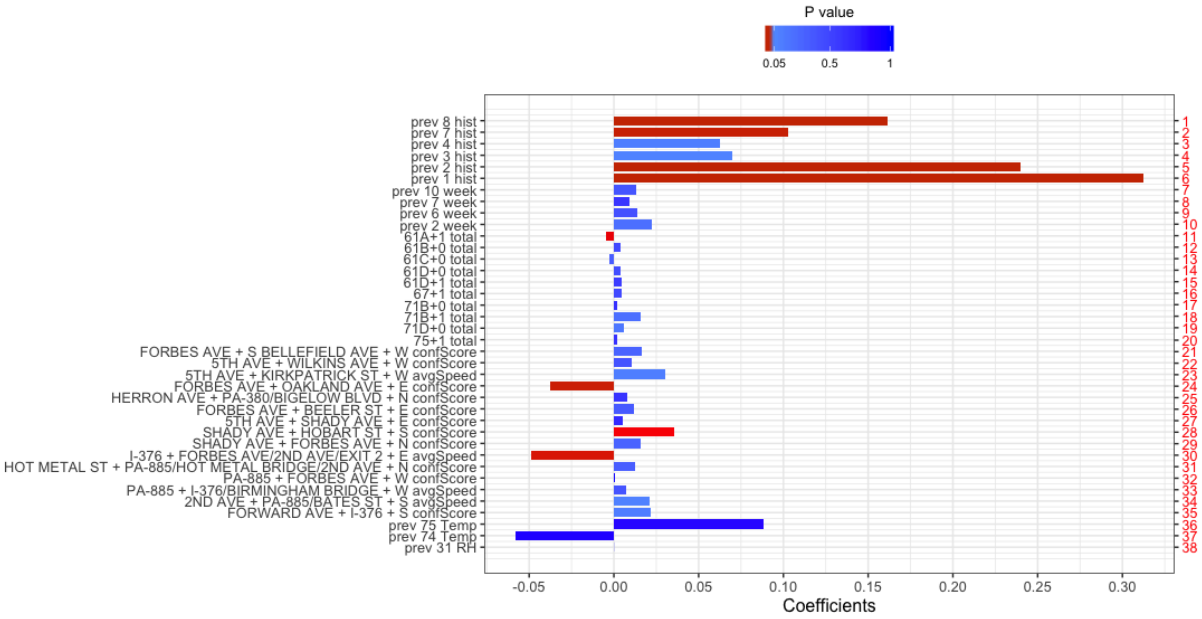

Figure 14: Coefficients and P value, weekend, 6:00 - 9:00, $\Delta t = 15$

5.4 Performance for different Prediction Horizon

We vary the prediction horizon from 15 minute to 30, 45 and one hour, to test if the performance is consistent. The results for one hour ahead predictions are listed in Table 8 and 9. The trend is showed in Figure 15 for time period 6:00 - 9:00. It is observed that on average the proposed model can outperform the baseline model for all of the prediction on

weekend. On weekday, however, the inclusion of transportation features have a negative impact. Results for other time periods are listed in the supplement document.

Table 8: Model Comparison, **Linear**, weekday, $n\Delta t = 60$

Period	**baseline, OLS**			**Selected, LASSO**		
	MSE	MAPE (%)	R^2	MSE	MAPE (%)	R^2
6:00 - 9:00	111.15	1.74	0.84	125.57	1.80	0.81
9:00 - 13:00	94.76	1.38	0.84	109.51	1.44	0.82
13:00 - 16:00	88.02	1.28	0.85	76.61	1.16	0.87
16:00 - 19:00	96.12	1.46	0.87	86.35	1.40	0.89

Table 9: Model Comparison, **Linear**, weekend, $n\Delta t = 60$

Period	**baseline, OLS**			**Selected, LASSO**		
	MSE	MAPE (%)	R^2	MSE	MAPE (%)	R^2
6:00 - 9:00	60.27	1.31	0.56	30.48	0.98	0.78
9:00 - 13:00	70.53	1.40	0.66	48.16	1.18	0.77
13:00 - 16:00	64.95	1.27	0.71	40.13	1.03	0.82
16:00 - 19:00	78.85	1.37	0.66	49.43	1.14	0.79

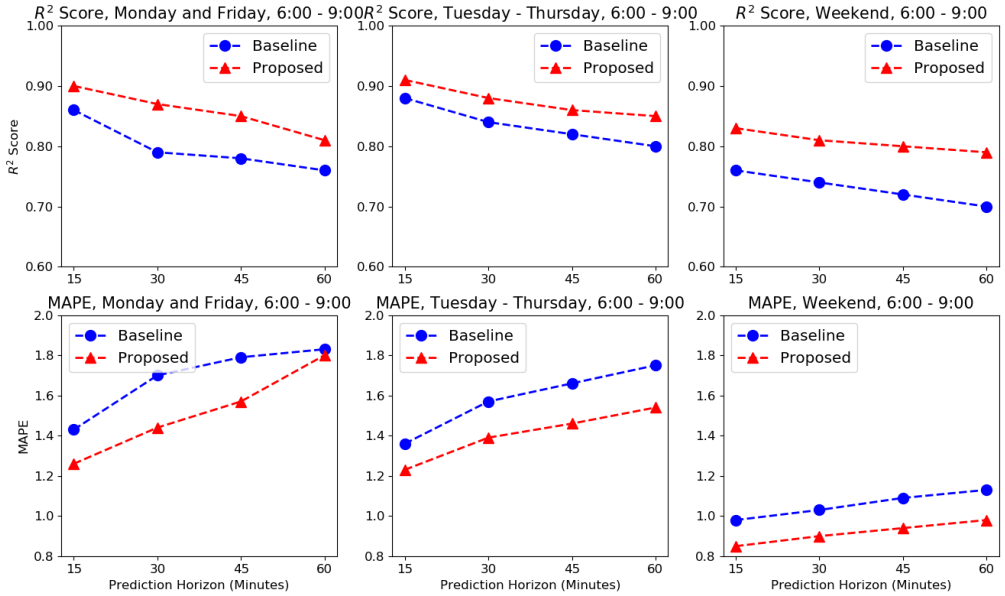

Figure 15: Comparison: R^2 and MAPE over different prediction horizon

5.5 Performance on individual buildings ($\Delta t = 15$)

We further analyze the performance on building level for each of the building. The performances for two periods, 6:00 - 9:00 and 9:00 - 13:00, are listed in Table 10. As expected, the improvement is consistent for individual buildings. However, compared with the aggregated model, we also notice a degradation of prediction accuracy, which is consistent with the observation in other papers mentioned in Background. The comparison for other period can be found the supplement.

Table 10: Model Comparison for different building, 6:00 - 9:00, weekends

Period	**baseline, OLS**			**Selected, LASSO**		
	MSE	MAPE (%)	R^2	MSE	MAPE (%)	R^2
GHC	18.20	3.03	0.57	13.26	2.61	0.69
Roberts	0.78	0.66	0.92	0.50	0.53	0.95
Purnell	1.11	1.61	0.85	0.65	1.21	0.91
Hamburg	2.17	2.36	0.78	1.45	2.00	0.86
MMCH	0.95	2.29	0.92	0.64	1.86	0.95
NSH	5.60	1.95	0.68	3.99	1.69	0.78

5.6 Post analysis: when would traffic feature help the most?

To discovery the recurrent patterns for the improvement we further compared the performance of proposed model on a daily scale. Basically, the days that the proposed model outperforms baseline are defined using the following threshold: (sum over all prediction in that day)

$$\frac{\sum_{i=1}^{n} err_b(i)^2 - \sum_{i=1}^{n} err_{prop}(i)^2}{\sum_{i=1}^{n} err_{prop}(i)^2} > r \tag{3}$$

The period for weekdays and weekends are 6:00 - 9:00, while the thresholds are 1.5 and 2.5. The comparisons are showed in Figure 16, 17 and 18 below. The improvement is not visually significant. However, it is observed that during final and orientation weeks. For weekends, the proposed model has a superior performance for the same event days as well, i.e., 2018-05-12 as final week and 2018-08-18 as orientations. Furthermore, significant improvements occur during season transition periods, such as 2018-03-24 and 2018-06-17.

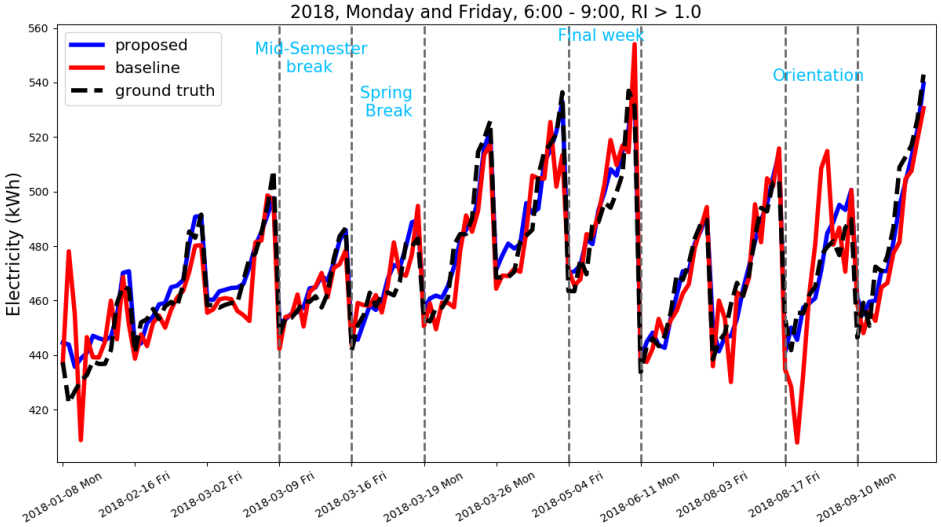

Figure 16: Example days, Monday and Friday, 6:00 - 9:00

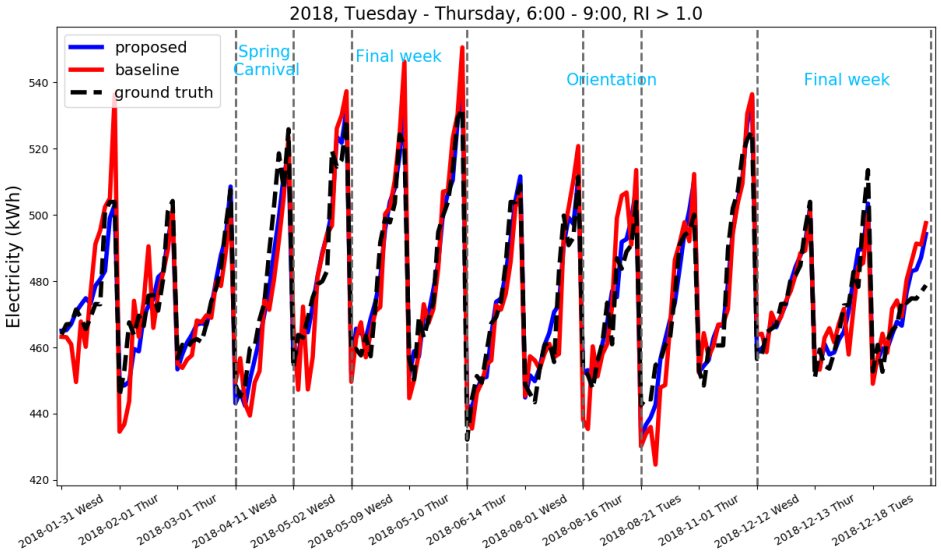

Figure 17: Example days, Tuesday - Thursday, 6:00 - 9:00

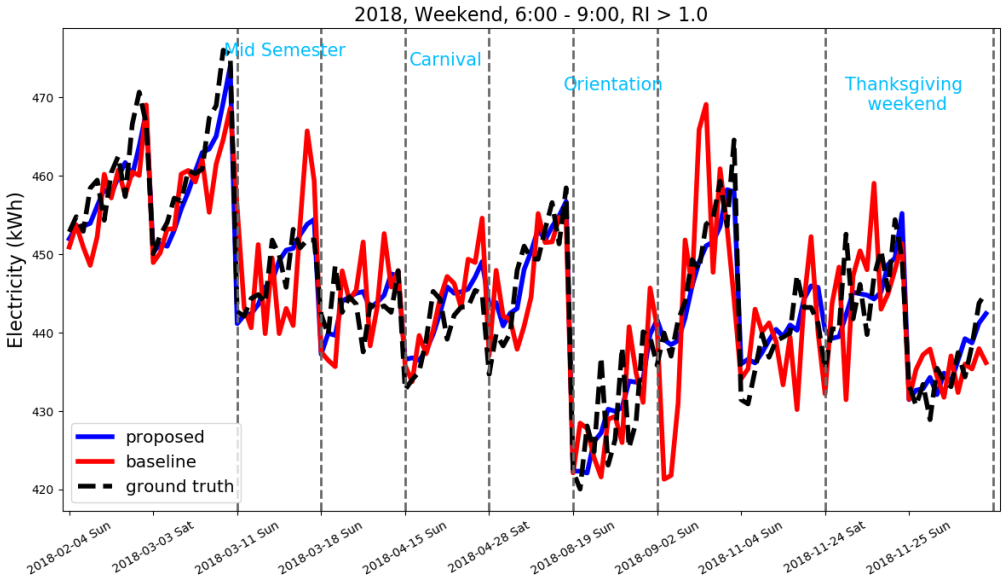

Figure 18: Example days, weekend, 6:00 - 9:00

The distributions for the relative improvement of all days are showed in Figure 19 and 21. It is observed that the proposed model outperform baseline model most of the time on weekdays.

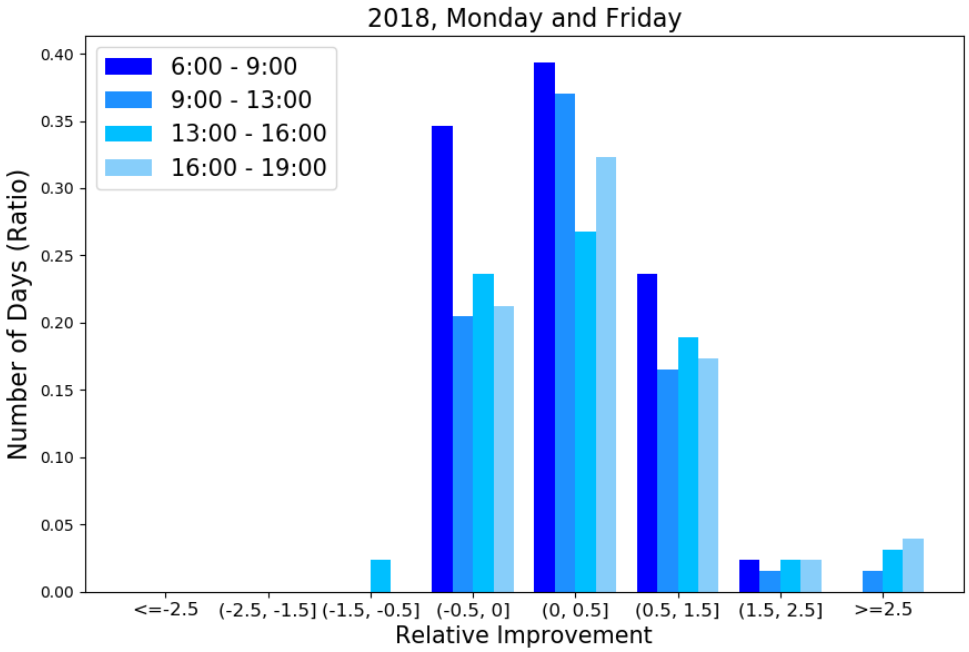

Figure 19: Relative improvement, Monday and Friday

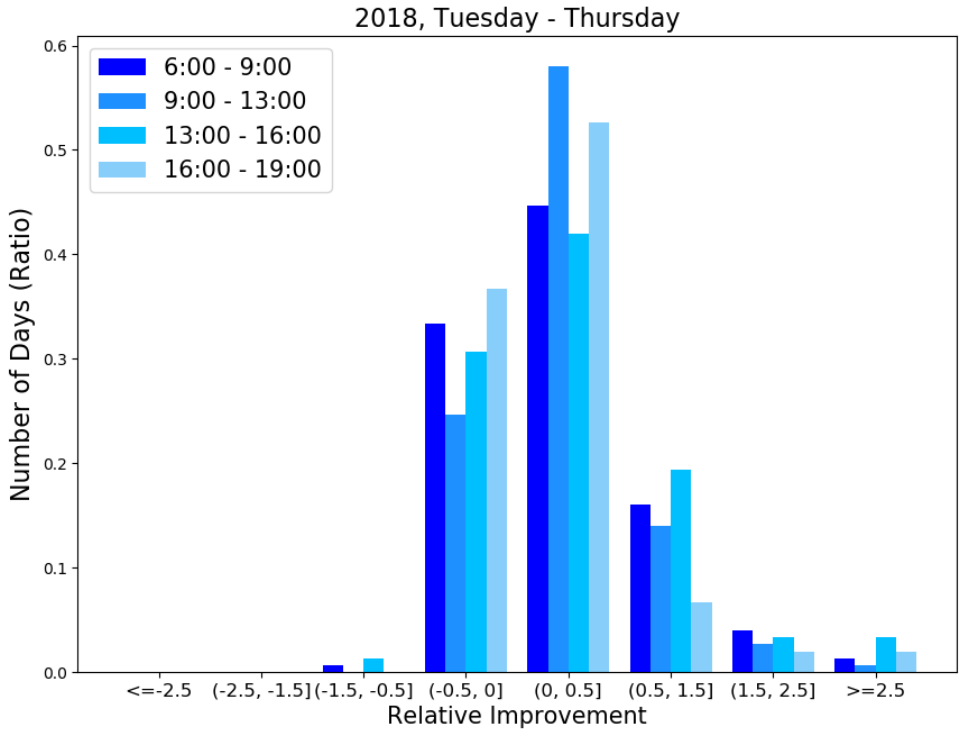

Figure 20: Relative improvement, Tuesday - Thursday

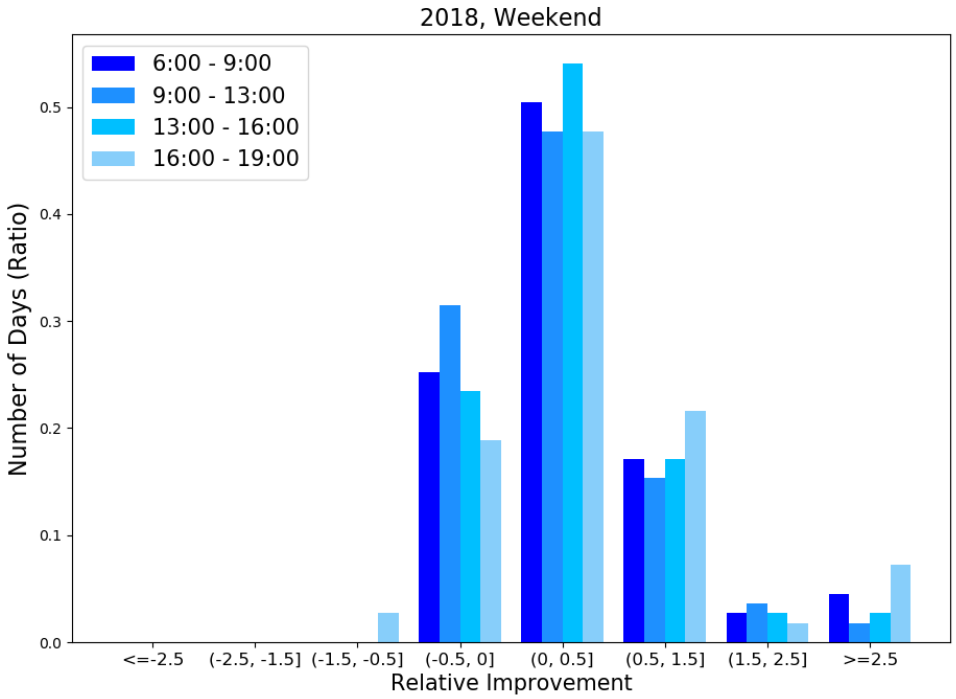

Figure 21: Relative improvement, weekend

5.7 Visualization: selected feature from inrix

Common features during 6:00 - 9:00, between weekends and Tuesday-Thursday:

- 5th Ave + Shady Ave, confScore
- Forbes Ave + Shady Ave, confScore
- Forbes + S Braddock Ave, AvgSpeed
- PA-885 + Hazelwood Ave, confScore

Common features during 16:00 - 19:00, between weekends and Tuesday-Thursday:

- Shady Ave + 5th Ave, confScore
- Forbes Ave + Schenley Dr, ConfScore
- Hot Metal ST + PA-837/E Carson ST, ConfScore
- Bigelow Blvd + Forbes, ConfScore

5.7.1 weekday: Monday and Friday

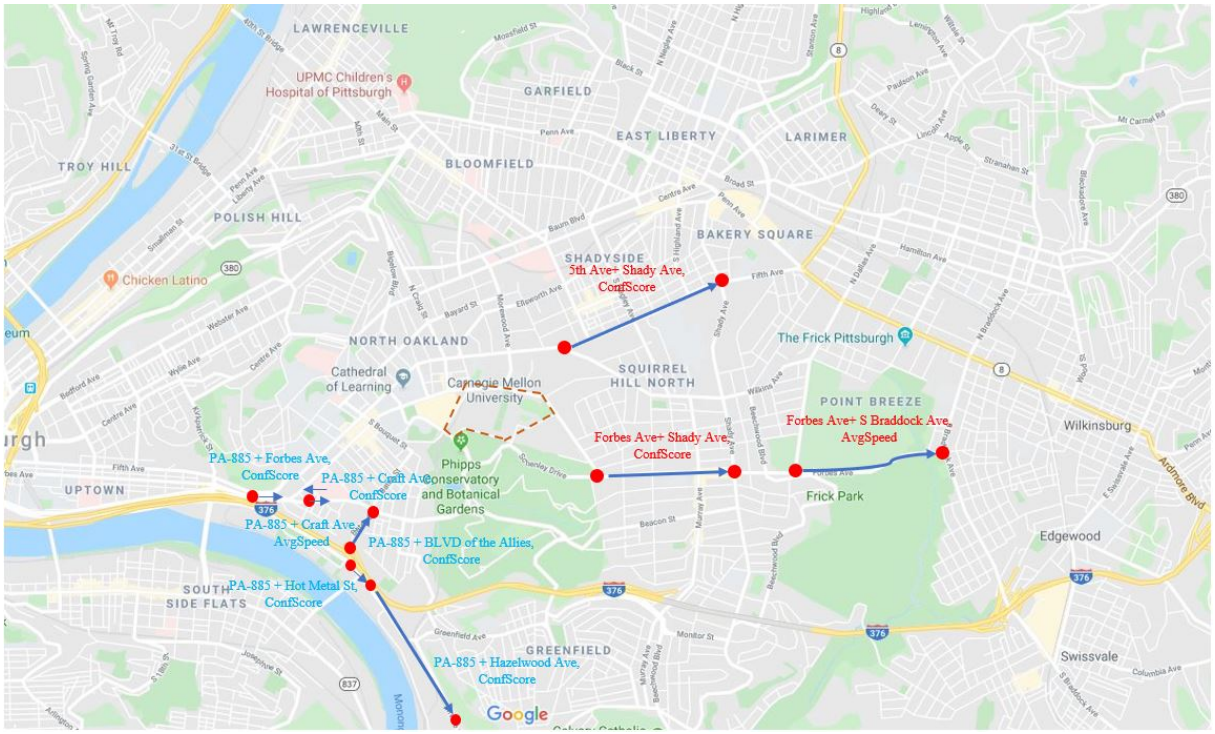

Figure 22: Visualization of inrix features, Monday and Friday, $\Delta t = 15$

5.7.2 weekday: Tuesday - Thursday

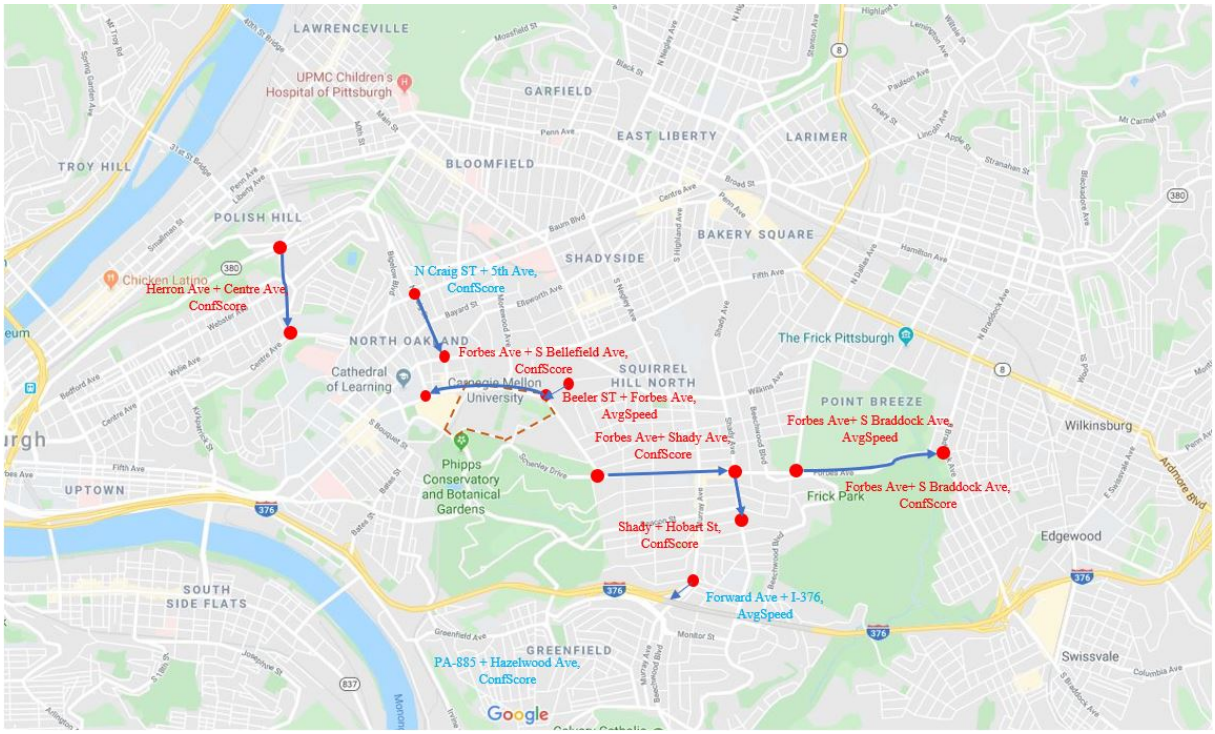

Figure 23: Visualization of inrix features, Tuesday - Thursday, $\Delta t = 15$

5.7.3 weekend

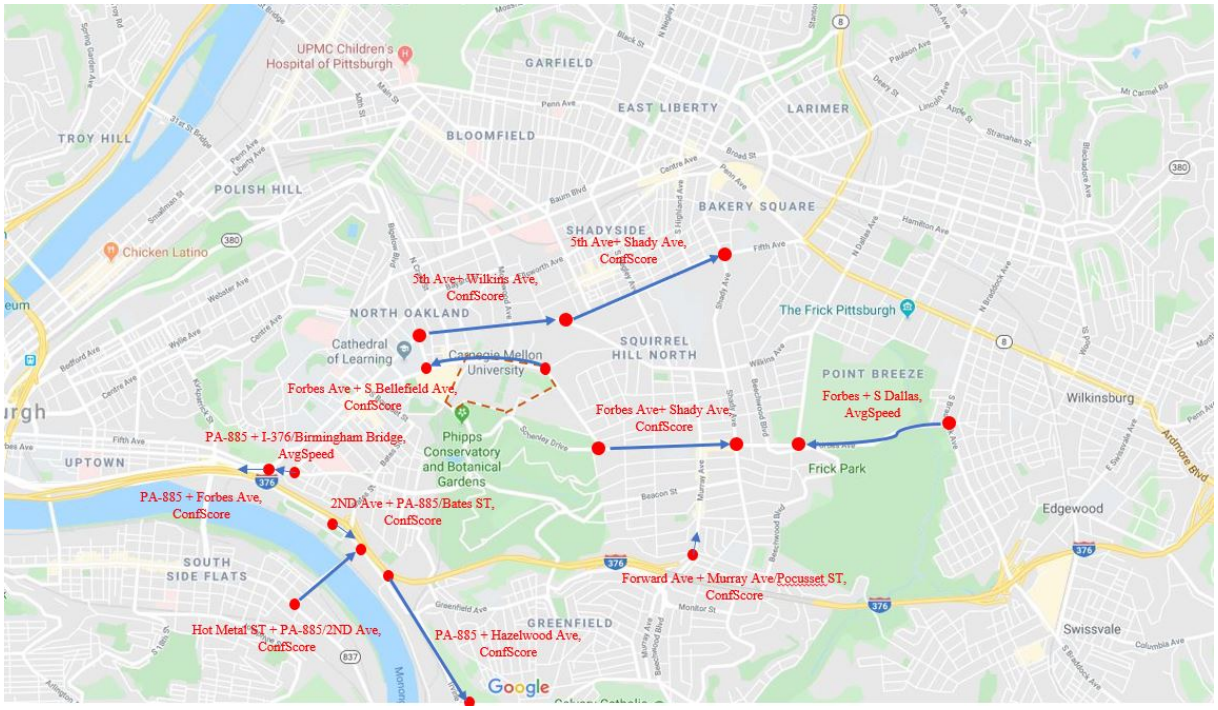

Figure 24: Visualization of inrix features, weekend, $\Delta t = 15$

5.8 Improvement for Occupancy Prediction

The same approaches used for electricity prediction are adopted to test whether transportation features are also beneficial of predicting occupancy in buildings. The response variable (Y) are the occupancy data sampled from 9/15/2017 00:00:00 - 10/15/2017 23:00:00, and from 4/17/2019 00:00:00 - 5/15/2019 00:00:00 at GHC [7]. The data was collected from all offices (301 in total) with occupancy sensor (PIR), with reading '1', '0' indicating occupied/unoccupied status. The occupancy status is defined as the sum over all offices during each time interval, aligned in time with that of the GHC electricity data (15 minutes interval). The range of occupancy status is between zero and 301. The trend is showed in Figure 7. The correlation coefficient between occupancy status and electricity is 0.8375 and 0.8145, which indicates these two observations are highly correlated.

5.8.1 performance at different time of day

To investigate the improvement of occupancy prediction given traffic features, the total occupied zones at time t is the response variable $y(t)$, its past n_{occ} (5, chosen from cross-validation) historical observations, past n_w (48) outdoor air temperature, previous inrix features (confScore, avgSpeed), previous bus features (total number on bus) are used as covariates. The processing of the features is the same as that used for electricity prediction. However, since the available occupancy information is limited, the length of initial covariates are shortened accordingly. Specifically, only the significant type of features in inrix and bus set are used in this part of analysis. Using the same feature selection pipeline, we denoted *proposed* as the model with featured selected from cross-validated lasso, and *baseline* as the model with historical features (occupancy and weather) only.

Due to the limited volume of occupancy data at hand, we first conduct a 15 minute ahead prediction for weekdays and weekend separately. The dataset collected in 2017 is used as training set and 2019 as testing. To overcome the scarcity of data the baseline model is trained using all data during 6:00 - 19:00. On the other hand, a unique proposed model is generated for period (each time of day). The results on the testing set are for weekday (Monday - Friday) and weekends listed in Table 11 an 12 below. It is observed that the proposed model can only outperform the baseline model during 6:00 - 9:00 for 15-minutes-ahead prediction.

[7] actually the data was sampled until 6/14/2019, but the bus features are only available until 5/15/2019

Table 11: Model Comparison for occupancy prediction, weekday (Monday - Friday), $\Delta t = 15$

Period	**baseline, OLS**		**Selected, LASSO**	
	MSE	R^2	MSE	R^2
6:00 - 9:00	51.03	0.96	23.67	0.98
9:00 - 13:00	154.09	0.97	181.70	0.97
13:00 - 16:00	65.36	0.99	58.69	0.99
16:00 - 19:00	58.01	0.98	57.22	0.98

Table 12: Model Comparison for occupancy prediction, weekends, $\Delta t = 15$

Period	**baseline, OLS**		**Selected, LASSO**	
	MSE	R^2	MSE	R^2
6:00 - 9:00	91.19	0.95	45.92	0.97
9:00 - 13:00	104.43	0.93	119.55	0.92
13:00 - 16:00	79.45	0.74	74.71	0.76
16:00 - 19:00	94.43	0.96	75.51	0.97

The correlations, coefficients and p-values for the proposed model during 6:00 - 9:00 at weekend are showed in Figure 25 and Figure 26 below. The most dominant factor is the previous 15 minute occupancy status. Also, it is observed there are 2/6, 6/14 significant factors from inrix and bus features set. The inrix feature '2ND AVE + BRADY ST + N, confScore' can be regarded as critical roadways since it is also selected by Lasso in the electricity analysis.

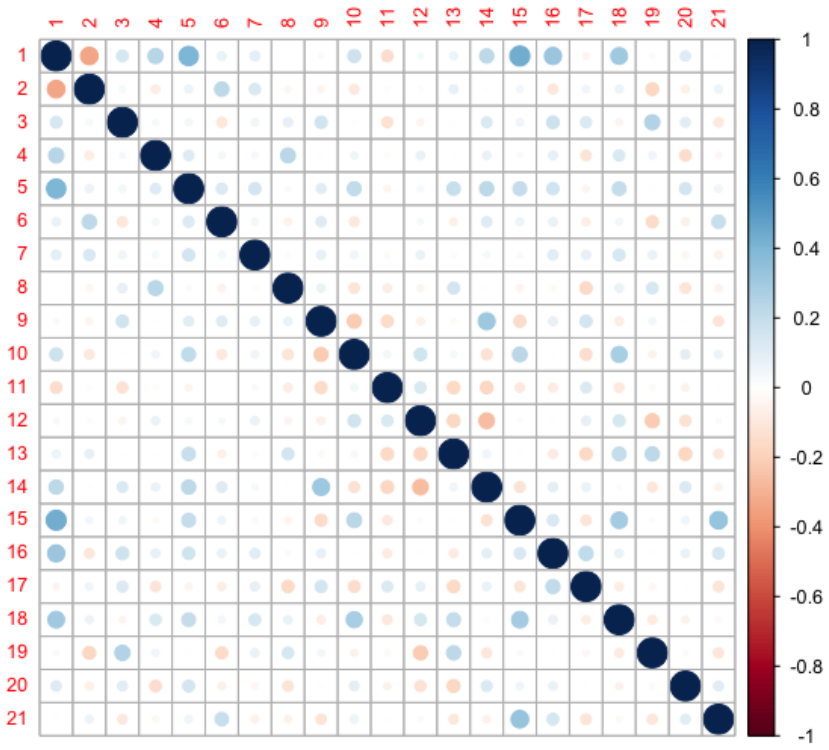

Figure 25: Feature correlations, occupancy prediction, weekend, 6:00 - 9:00, $\Delta t = 15$

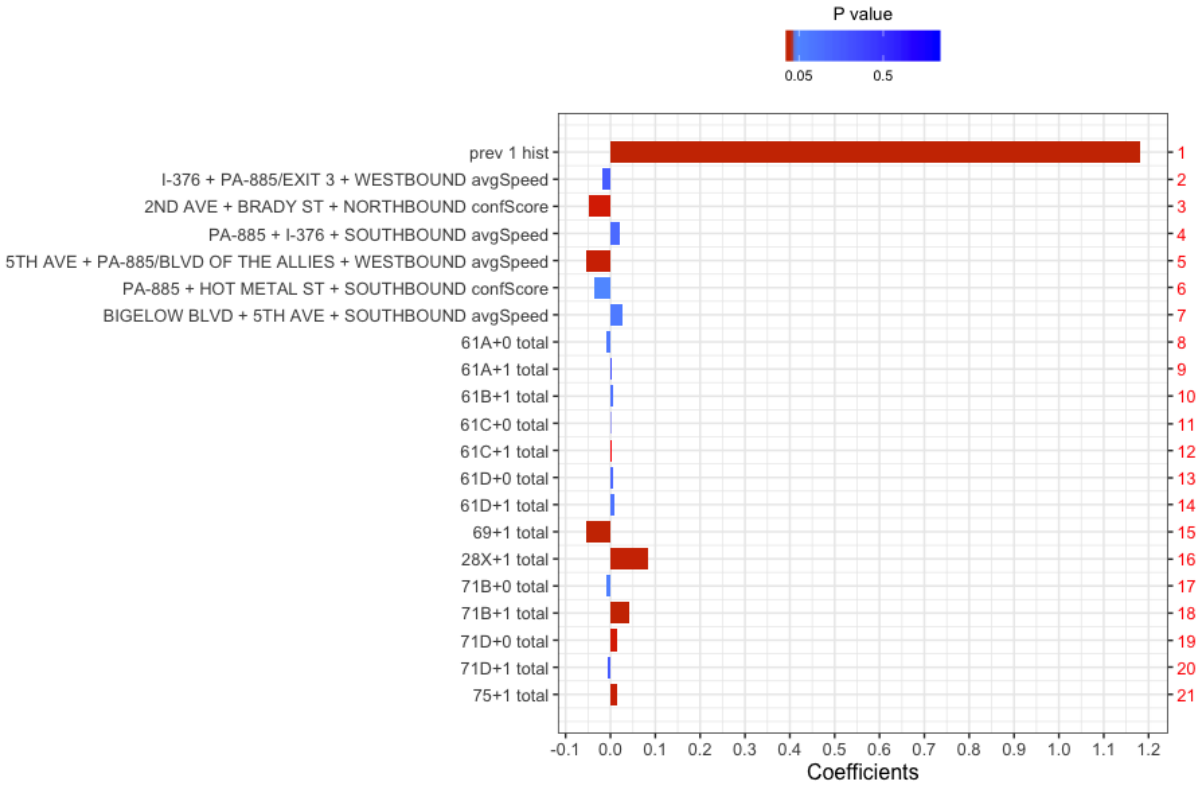

Figure 26: Coefficients and P value, occupancy prediction, weekend, 6:00 - 9:00, $\Delta t = 15$

5.8.2 performance at different prediction horizon

We further analyze if the improvement in prediction is consistent as prediction horizon increases. It is found that the improvement is only persistent during the period 6:00 - 9:00 for weekday and weekend. The trend plot is showed in Figure 27 below.

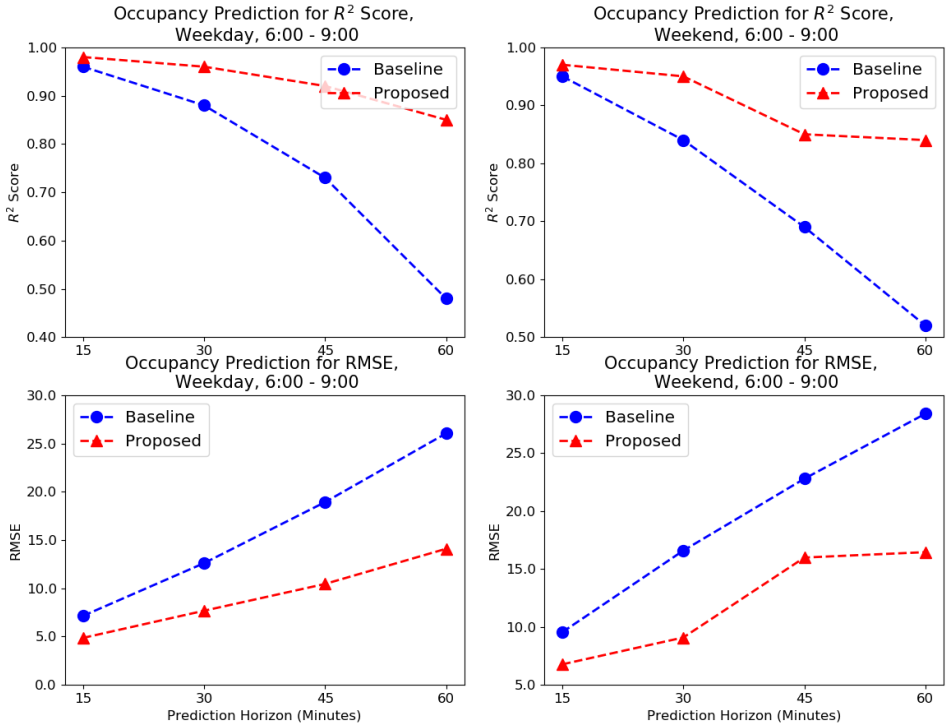

Figure 27: Comparison: R^2 and RMSE over different prediction horizon

6 Conclusion

To sum up, the influence of using real-time high-resolution traffic features for the prediction of electrical loads in buildings have been thoroughly explored in this project. It is revealed that the inclusion of such features improve the overall precision of prediction, especially during weekends when the electrical loads are more random than weekdays. The general conclusions are summarized as follows:

1. for 15 minutes ahead prediction, on average the proposed model can reduce the MAPE from 1.08 to 0.99, and 1.11 to 0.85 for weekdays and weekends respectively, compared when the baseline model with historical features alone. The most significant improvements are observed on days in final and recitation weeks for weekends, and for the morning on Mondays when the weekend-to-weekday transition happen.
2. The propriety of using linear regression model has been justified by the normally distributed residuals and superior performance in MSE, MAPE and R^2 when compared with nonlinear models such as random forest and Multilayer perceptions.
3. Within the traffic feature set, the total number of occupant on bus, confidence score, and average speed are identified to have significant impact on building electrical loads.
4. Finally, based on the relationship between building electrical load and occupancy, we have also validated that using transportation features can further reduce estimation error of building level occupancy.

References

[1] E. Kyriakides and M. Polycarpou. Short term electric load forecasting: A tutorial. *In Trends in Neural Computation*, 2007.

[2] Chen S. Ho T.K. Mirowski, P. and C.N. Yu. Demand forecasting in smart grids. *Bell Labs technical journal*, 2014.

[3] Sánchez-Úbeda E.F. Cruz A. Muñoz, A. and J. Marín. Short-term forecasting in power systems: a guided tour. *In Handbook of power systems II*, 2010.

[4] E.A. Feinberg and D. Genethliou. Load forecasting. *In Applied mathematics for restructured electric power systems*, 2005.

[5] Frincu-M. Chelmis C. Noor M. Simmhan Y. Aman, S. and V.K. Prasanna. Prediction models for dynamic demand response: Requirements, challenges, and insights. *In 2015 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2015.

[6] Boudreau-M.C. Helsen L. Henze G. Mohammadpour J. Noonan D. Patteeuw D. Pless S. Lawrence, T.M. and R.T. Watson. Ten questions concerning integrating smart buildings into the smart grid. *Building and Environment*, 2016.

[7] Hong-T. Wang, Z. and Piette M.A. Predicting plug loads with occupant count data through a deep learning approach. *Energy*, 2019.

[8] Airaksinen-M. Sekki, T. and A. Saari. Impact of building usage and occupancy on energy consumption in finnish daycare and school buildings. *Energy and Buildings*, 2015.

[9] Lee-D. Robinson P. Britter R. Martani, C. and C. Ratti. Enernet: Studying the dynamic relationship between building occupancy and energy consumption. *Energy and Buildings*, 2012.

[10] Y.S. Kim and J. Srebric. Impact of occupancy rates on the building electricity consumption in commercial buildings. *Energy and Buildings*, 2017.

[11] Zhang-K. Dutta K. Yang Z. Ghahramani, A. and B. Becerik-Gerber. Energy savings from temperature setpoints and deadband: Quantifying the influence of building and system properties on savings. *Applied Energy*, 2016.

[12] Arens-E. Hoyt, T. and H. Zhang. Extending air temperature setpoints: Simulated energy savings and design considerations for new and retrofit buildings. *Building and Environment*, 2015.

[13] Katipamula-S. Wang W. Huang Y. Fernandez, N. and G. Liu. Energy savings modelling of re-tuning energy conservation measures in large office buildings. *Journal of Building Performance Simulation*, 2015.

[14] Ucci-M. Marmot A. Lakeridou, M. and I. Ridley. The potential of increasing cooling set-points in air-conditioned offices in the uk. *Applied energy*, 2012.

[15] Borrelli-F. Hencey B. Coffey B. Bengea S. Ma, Y. and P. Haves. Model predictive control for the operation of building cooling systems. *IEEE Transactions on control systems technology*, 2012.

[16] Master-N. Taneja J. Culler D. Aswani, A. and C. Tomlin. Reducing transient and steady state electricity consumption in hvac using learning-based model-predictive control. *Proceedings of the IEEE*, 2012.

[17] Kara-E.C. MacDonald J. Andersson G. Vrettos, E. and D.S. Callaway. Experimental demonstration of frequency regulation by commercial buildings–part i: Modeling and hierarchical control design. *IEEE Transactions on Smart Grid*, 2018.

[18] Armstrong-P.R. Gayeski, N.T. and L.K. Norford. Predictive pre-cooling of thermo-active building systems with low-lift chillers. *HVAC&R Research*, 2012.

[19] Risbeck-M.J. Rawlings J.B. Wenzel M.J. Patel, N.R. and R.D. Turney. Distributed economic model predictive control for large-scale building temperature regulation. *American Control Conference (ACC)*, 2016.

[20] Matuško-J. Ma, Y. and F. Borrelli. Stochastic model predictive control for building hvac systems: Complexity and conservatism. *IEEE Transactions on Control Systems Technology*, 2015.

[21] Parisio-A. Jones C.N. Morari M. Gyalistras D. Gwerder M. Stauch V. Lehmann B. Oldewurtel, F. and K. Wirth. Energy efficient building climate control using stochastic model predictive control and weather predictions. *American control conference (ACC)*, 2010.

[22] V.M. Zavala. Real-time optimization strategies for building systems. *Industrial & Engineering Chemistry Research*, 2012.

[23] D. Bunn and E.D. Farmer. Comparative models for electrical load forecasting. *NA*, 1985.

[24] P. Zhang and Z.S. Qian. User-centric interdependent urban systems: Using time-of-day electricity usage data to predict morning roadway congestion. *Transportation Research Part C: Emerging Technologies*, 2018.

[25] Wang-D. Pei J. Yuan Y. Fan C. Zheng, Z. and F. Xiao. Urban traffic prediction through the second use of inexpensive big data from buildings. *In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016.

[26] S. Kim and B. Coifman. Comparing inrix speed data against concurrent loop detector stations over several months. *Transportation Research Part C: Emerging Technologies*, 2014.