Working Title: Generalizing building electricity demand forecasting by means of hybrid model and spatial data enhancement

Cheng-Kai Wang
C.Wang-5@student.tudelft.nl

January 15, 2018

1 Introduction

The building sector plays an indispensable role in achieving low-carbon future as buildings alone account for nearly one third of total final energy use and CO2 emissions. Beside that, energy consumption in the building sector has grown by 1.3% per year in OECD and 2.1% per year in non-OECD countries from 1990 to 2010 [7]. Similar statistics are also given by the *World Business Council for Sustainable Development* [20]. Additionally, the rapid urbanization trend worldwide poses another stress in supporting increasing energy demand while achieving decarbonization target [15].

Although many challenges lie ahead, it also indicates a huge energy saving potential [5] and a golden opportunity to develop energy saving technologies and strategies to cope with such urgency in the built environment. Potential pathways include demand response and load shifting [17], fault detection and retrofit strategies [12, 13] as well as future scenario planning, building energy and electricity consumption signature and forecasting model at different spatial scales serves as a foundation to enable these applications.

1.1 Objective

There are various technologies ranging from smart meter analytics to building energy demand simulation that try to characterize building and urban scale energy use patterns and make accurate predictions accordingly. However, due to complex interactions between dynamic outdoor and indoor conditions, stochastic human behavior, diverse building thermal characteristics and energy systems including HVAC and lighting [2, 3, 13], ways to robustly and accurately predict energy consumption with minimum requirement for privacy-related data remains a challenging and active field [13]. The scope of this research is to explore how to make use of the ever-growing energy relevant spatial data, particularly building geometry and semantic information, energy simulation models at the district scale, and recorded smart meter data to improve building electricity demand forecasting model at residential district as well as to understand the significance of these influencing factors.

1.2 Outline of the proposal

The proposal will first start with literature review. This section illustrates the importance and value of building energy consumption signatures and also presents frequently applied forecasting models and approaches. Research questions will be presented in section 3. Section 4 explains a possible methodology to tackle the presented problem, while section 7 section and 6 focus on project planning, datasets and tools that are used in this research.

2 Related work

2.1 Energy saving potential enabled by consumption pattern

Accurate building energy consumption patterns and demand estimation is a necessary enabler for many energy saving strategies such as energy automation [2], supply and demand side response in smart grid systems [1, 17], user behavior change [11], building operation management and appliance fault detection [12, 13] and energy-driven urban planning [15].

Frankel et al. (2013) have reported 16% to 20% savings of the US residential-energy demand might be benefited from behavioral adjustment [5], or alternatively via energy automation, which directly relies on precise consumption signatures and accurate demand estimation. In terms of demand side response (DSR), utilities change electricity pricing based on the estimated *baseline* model and provide incentives to the consumers in order to adjust consumption behavior and enable load sifting [17]. Additionally, long-term climate change might significantly change weather pattern in specific areas [16]. As buildings have comparatively long lifetime and low retirement rate, consumption signature and demand estimation in the long-run should be considered during the planning phase of the new building design.

Many studies have researched different approaches for modeling consumption pattern and demand forecasting. These can be briefly categorized into three types: *i) Deterministic model approach*, *ii) Data-driven approach*, and *iii) Hybrid model approach* and are described in sequence below.

2.2 Deterministic model approach

Deterministic techniques simulate time-series energy consumption and performs forecasting based on the science of building physics [3]. Accuracy of such approach is often determined by the completeness and resolution of building thermal properties data and the underlying thermal physics model. Wate and Coors (2015) have concluded relevant building thermal properties affect temporal prediction accuracy in different level of details (LOD) [18].

On the other hand, there are many existing software packages such as CitySim, EnergyPlus, City Energy Analyst and TRNSYS used by engineer and designer to simulate energy consumption of the building at different spatial scales given different design scenarios. Each of these softwares targets different use cases from urban scale simulation using CitySim [15], district level planning and system simulation using City Energy Analyst [4] to the detailed and sophisticated simulation engine EnergyPlus. The advantage of such physically based model is that it does not require training data. It simulates non-linearity of energy demand given different building properties and environmental conditions and thus easier to generalize for the changing conditions. However, data unavailability and discrepancies of building thermal properties and difference of the underlying thermal simulation engine often cause such approach fall short of enough accuracy [3]. Furthermore, stochastic human behavior is difficult

to model and is often recognized as one of the main causes why simulation softwares have higher uncertainty.

2.3 Data-driven approach

With increasing number of environmental monitoring sensors deployed in the built environment, huge volumes of recorded data become valuable resources for data-driven modeling approaches. The technique depends on recorded smart meter data or even real-time energy data and attempt to learn the temporal consumption trends based on previous usage patterns and extrapolate to future scenarios [3]. Many supervised and unsupervised machine learning algorithms are developed to enable such forecasting models and increase prediction accuracy. These include: Regression fitting, Support vector machine (SVM), Artificial neuron network (ANN), Random forest, Bayesian Networks and so on. Details of these algorithms and associated performance are reviewed by [1, 3, 12, 13]

Many studies have shown data-driven approach results in better prediction accuracy than deterministic model [3, 9]. However, the downside of such method is that it often relies on site-specific recorded data and thus more difficult to generalize when conditions have dramatic changes or unavailability of training data. Additionally, the complex interaction between input and output data is often sophisticated and difficult to interpret and thus has less transparency [3].

2.4 Hybrid model approach

Deb et al. (2017) defines the hybrid model as a combination of more than one machine learning techniques. By combining different methods, hybrid model is capable of modeling more complex autocorrelation structure of relevant features and thus improves accuracy [3].

Koponen et al. (2014) studied another hybrid model which not only combines different machine learning algorithms but includes partly physically based model and a Kalman-filter based predictor to predict energy demand [9].

Table 1: Comparison between data-driven and deterministic forecasting model [3]

Technique	Advantage	Disadvantage
Data-driven	- Very fast in computation with real-time data	- Requires past recorded data
	- Suitable for non-linear modeling	- Non-transparent and confined
	- Often more accurate than deterministic models	- Difficult to generalize
Deterministic	- Based on the science of building physics	- Difficult to model real scenarios
	- Transparent and no training data needed	- Data unavailability of building properties
	- Easy to generalize	- Not very accurate

2.5 Performance indices

The performance of the forecasting model is often measured by comparing the forecast results with actual load data during validation period [6, 9]. Commonly used performance indices include:

- MSE: Mean Squared Error = mean(e_t^2)
- RMSE: Root Mean Squared Error = $\sqrt{MSE}$
- MAE: Mean Absolute Error = mean($|e_t|$)

- MAPE: Mean Absolute Percentage Error = mean($|p_t|$)

Where e_t refers to the forecasting error at time t; while $p_t = 100e_t/y_t$ and y_t is the observation at time t.

3 Research questions

Wate and Coors (2015) and Miller (2016) have pointed out relevant influencing factors of energy consumption in the reports [11, 18]. Koponen et al. (2014) also reports significant accuracy improvement, for the partly physically based method it improved the the MAPE of hourly power forecast from 7.37% to 4.57%, when weather forecast data is included in the forecasting model [9]. However, often due to unavailability of data and time-consuming work of processing and organizing heterogeneous spatial data, the influence of spatial data such as building geometry, building semantic information on the short-term electricity consumption pattern and demand estimation is less studied. Consequently, the research scope of this project will explore:

- To what extent with addition of spatial features such as building geometry, building semantic information and physically based simulation results as model inputs can improve sub-hourly prediction accuracy of building electricity consumption at district scale.

To answer this question, the following sub-questions should be included in the research scope:

- What is the suitable hybrid model for short-term consumption forecasting which considers smart meter measurement, environmental data, building geometry, information and simulation result as model inputs.

Since unavailability and discrepancy of data are common, it is also interesting to look at which features have most significant influence on energy consumption in the ultimate model, so feature weights can be initialized accordingly.

- Among additional spatial features, which are the significant influencing factors of building electricity consumption at district scale in the real world practice.

4 Methodology

This section explains how the proposed methodology can be used to tackle the research questions in the following sequence. The overview of the expected workflow is presented in Figure 1

4.1 Feature engineering and selection

Except smart meter data, the research also considers the following spatial parameters in the final model inputs:

- Ambient condition data: local temperature, humidity, solar gain
- Building geometry: volume, floor, wall, and roof areas, glazing ratio
- Non-geometric building data: envelope thermal properties, infiltration rate of the building, thermal system, number of occupants, occupant profile and occupancy schedule, building usage type, built year, and refurbishment year

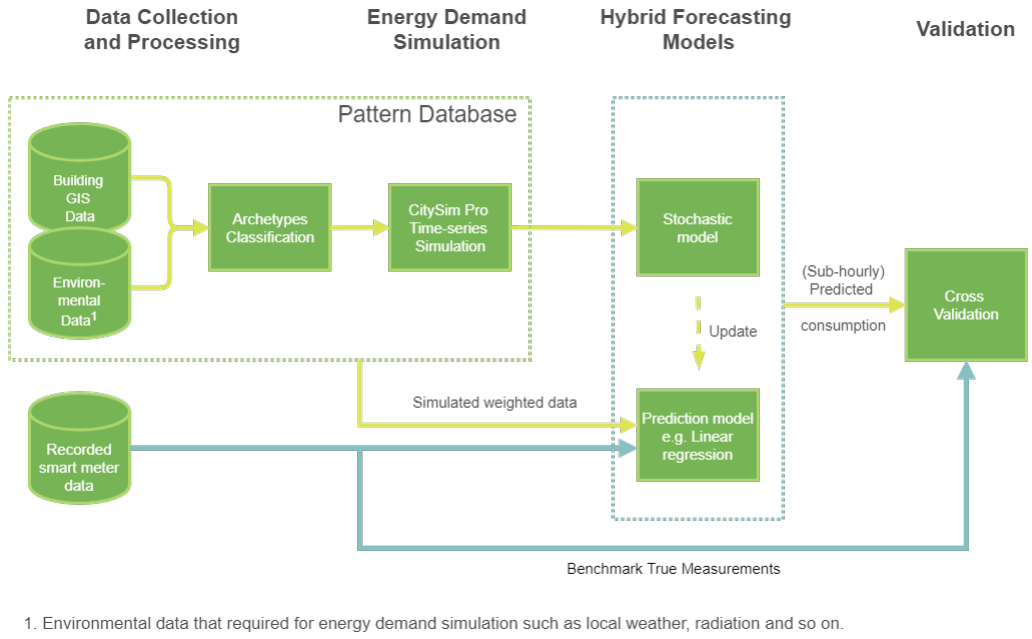

Figure 1: General workflow of the research methodology

- CitySim energy simulation results

Database software such as PostgreSQL and CityGML data model will be used to manage these various features.

Data processing and cleaning like outlier detection and interpolating missing values should be carried out before performing any modeling or analysis. On the other hand, to get an initial insight of data characteristic, Pearson correlation study between input features and recorded consumption can be applied in this step. Alternatively, principle component analysis (PCA) can serve as a dimension reduction method to identify major influencing features or initialize feature weight according to the feature eigenvalues for later use.

4.2 Archetype segmentation and characterization

One of the major challenges of using deterministic approach to model consumption pattern and perform demand forecasting is discrepancies and unavailability of detailed building thermal properties [3, 15], even Internet of Things (IoTs) has produced and accumulated huge volume of data in all sectors nowadays. The practical solution is thus abstracting building stocks into "building archetypes" based on the similarity of building usage type, built year, refurbishment year, thermal system and assign building thermal properties accordingly [15].

Segmentation step can be simply based on the defined decision rules or by means of many existing unsupervised segmentation algorithms such as k-means, k-nearest neighbor, random forest or t-SNE.

4.3 CitySim energy simulation

Data-driven approach is reported by many studies to have higher forecasting accuracy than deterministic model. However, as it relies on site-specific training data for the model fitting, it can not be easily generalized to other scenarios or perform future scenario simulation. The deterministic approach based on the science of building physics shows the advantage in this circumstance. The ideal here is thus using open source urban scale energy simulation software: CitySim to generate time-series simulation data as additional spatial features for model fitting and to compensate the downside of data-driven approach.

Preliminary workflow have been set up to enable importing building statistic data (for instance, number of occupants, thermal properties and so on), assigning building semantic information to the 3D model with Python script and FME, executing CitySim energy simulation and managing simulation output. Figure 5 illustrates the workflow for generating district level simulation with CitySim.

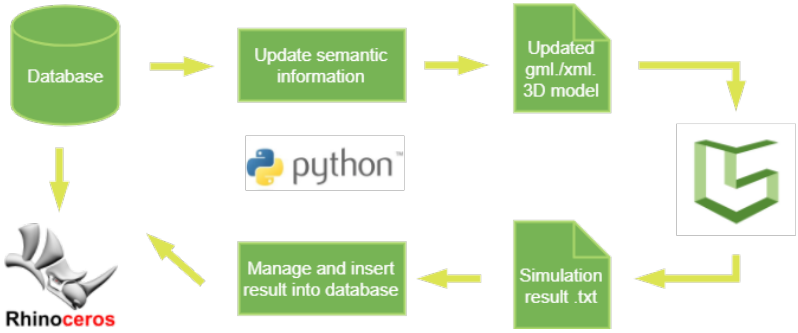

Figure 2: CitySim simulation procedure

4.4 Forecasting model building

The forecasting model to be adapted is based on a hybrid model, which is defined as using recorded data to train the "baseline" model while site-specific spatial features and simulation results of deterministic approach are considered as "spatial enhancement" used to model stochastic component of the future state, and can be written in the following form:

$$s_t^{(n)} = As_{t-1}^{(n)} + Bw_{t-1}^{(n)} \tag{1}$$

Where s_t is the consumption state at time t and is derived from the previous state s_{t-1}, which propagates based on the data-driven baseline model A, plus additional uncertainty w_{t-1} modeled by stochastic function. In terms of Sequential Monte Carlo (SMC) method, the state S is represented by the probability distribution of weighted sample set $S = \{(s^{(n)}, \pi^{(n)}|n = 1, ...N\}$, where π represents sample weight. The detail of the SMC model will be illustrated in section 4.4.2.

4.4.1 Regression model

Regression is one of the most commonly used model fitting techniques which is also applied in electricity load forecasting [14, 19], and can be a proper starting point for the baseline model. The simplest linear regression model has the general form as:

$$y = \boldsymbol{\theta}^T \boldsymbol{x} + \epsilon \tag{2}$$

Where y is the predicted consumption state, while $\boldsymbol{\theta}$ is a regression coefficient vector of the feature vector $\boldsymbol{x}$, which includes energy relevant features such as recorded data, ambient condition, building geometry and so on. ϵ is an error term used for ridge regression.

Mathieu et al. (2011) further extend the regression model with time-of-week indicators, which divide a week from Monday to Friday into 15 minute intervals and different regression coefficients for each time-of-week, α_i, and thus obtain different predicted load. [10]. Python resource[1] based on this concept is also available for direct use and can serve as a foundation of the baseline model for this research.

4.4.2 Sequential Monte Carlo based forecasting model

Sequential Monte Carlo (SMC) or particle filters is an optimal estimation algorithm and is a sampled-based solution of the recursive Bayesian filter which is widely used in navigation system, computer vision and signal processing [8]. This research attempt to model the stochastic component of equation 1 based on the concept of Sequential Monte Carlo.

Future state of energy consumption is difficult to predict as many stochastic events can significantly affect estimation value. Namely, the future state of energy consumption is not available for direct measurement but could rely on the data-driven regression model mentioned in section 4.4.1. In such prediction model, good estimation of stochastic components is more likely to lead to more accurate prediction based on posterior probability derived from the future state observer. In this case, observation state can be modeled based on spatial features such as weather forecast or deterministic model results and iteratively update to the stochastic components in the prediction model (or say baseline model). The prediction and observation state should have the following form:

$$Prediction : s_t = f_{t-1}(s_{t-1}, w_{t-1}) \tag{3}$$

$$Observation : z_t = h_t(x_t, v_t) \tag{4}$$

Where f_{t-1} in equation 3 refers to state prediction function given previous state s_{t-1} and stochastic component w_{t-1} models prediction uncertainty; while h_t in equation 4 refers to observation function given future state measurements x_t and the associated observation noise v_t. The set of all observations up to time t is denoted by $Z_t = \{z_1, ..., z_t\}$. In general, we are interested in finding an estimate of state vector s_t provided with previous state vector s_{t-1} and using measurement z_t to update the state estimation.

In the case of particle filter, energy consumption state is not represented by a single object state s_t but by its probability distribution $p(s_t)$. *A prior* distribution $p(s_t|Z_{t-1})$ is calculated according to the state prediction function, refers to equation 3; while *a posteriori* distribution $p(s_t|Z_t)$ can be estimated given new measurement z_t from the state observer according to equation 4. See Figure 3 for the concept illustration.

5 Cross validation and evaluation

Since one scope of this research is trying to understand how "spatial feature enhancement" can increase forecasting model accuracy or even generalize it. Consequently, in the end of

[1] eetd-loadshape: https://bitbucket.org/berkeleylab/eetd-loadshape#markdown-header-baselines

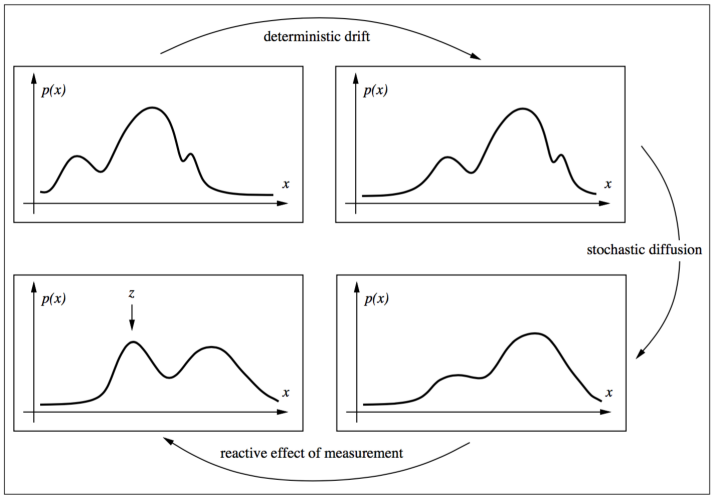

Figure 3: In the case of particle filter, state is modeled as probability and this diagram can be viewed as one discrete time-step. *deterministic drift* shown in the diagram is modeled according to prediction function while *stochastic* component diffuses the particle distribution. Different weights are given to the diffused particles based on observed state, and resampling results in new probability distribution of next time-step. Image source: [8]

the model building, performance index Mean Absolute Percentage Error (MAPE) introduced in section 2.5 will be used to cross validate the prediction performance as well as to cross validate with baseline prediction model manifested in section 4.4.1 without considering spatial features.

6 Tools and datasets used

6.1 Tools

The main tools to be used in this project include:

- Energy simulation: CitySim Pro
  The reason for choosing CitySim is because of direct hands-on experience and the set up workflow in collaboration with the researcher at Eidgenössische Technische Hochschule Zürich (ETH). It is also the few energy simulation software that models shared environment and targeting urban-scale energy simulation. Available simulation outputs from CitySim is presented in the following table.

Table 2: CitySim simulation output

Temporal scales	Simulation outputs
Hourly	Short-wave irradiation
Daily	Long-wave net irradiation
Monthly	Surface temperature
Yearly	PhotoVoltaic production
	Solar Thermal production
	Sky View Factor
	Heating demand
	Cooling demand
	Indoor temperature

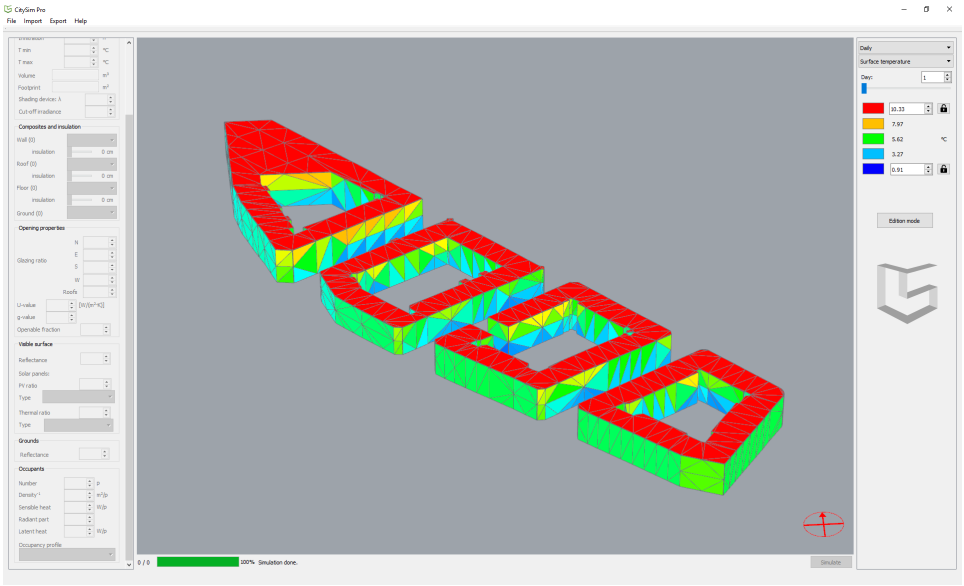

Figure 4: CitySim surface temperature simulation on the test site

- Data management: PostgreSQL, CityGML, FME
  Database PostgreSQL contains geo-component PostGIS, and is therefore an ideal tool to manage building statistics data. CityGML is a 3D data model combining geometry and the associated semantic information. To update heterogeneous spatial data to CityGML and perform energy simulation in CitySim, data integration software FME provides abundant spatial operation transformers.

- Programming languages: Python and Matlab
  Both Python and Matlab offer plenty machine learning libraries such as scikit, numpy or TensorFlow in Python, while Math, Statistics and Optimization toolbox in Matlab. Experience and familiarity with these languages is another reason.

- Visualization: Rhinoceros 3D with add-on Grasshopper
  Optional but if necessary, the result of building consumption pattern and demand forecasting at district scale can be visualized in Rhinoceros 3D environment or alternatively via JavaScript with WebGL API to allow visualization and interaction on web browser.

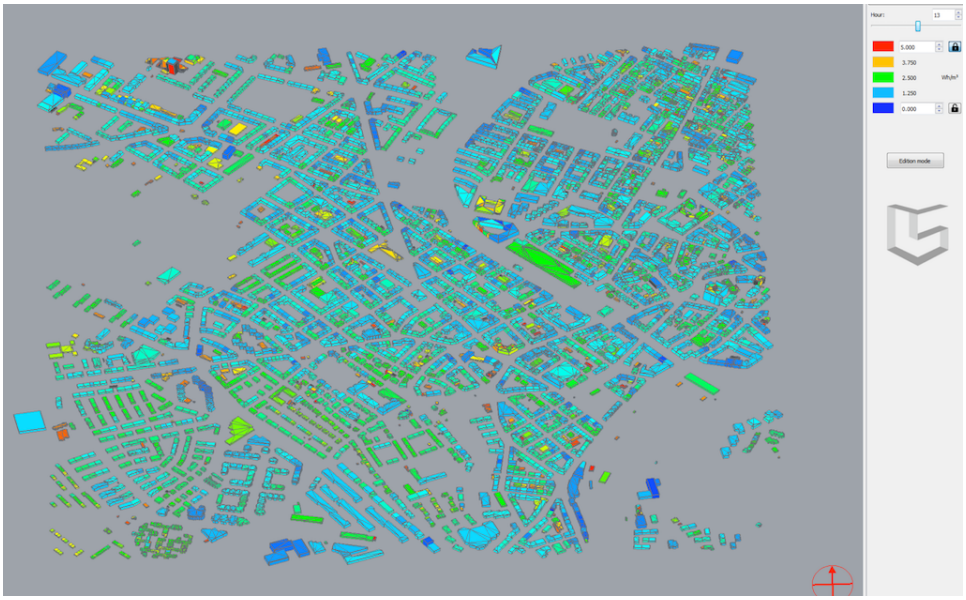

Figure 5: CitySim simulation of partial Zürich. Credit: Danielle Griego

6.2 Datasets

Initial data collection phase has searched and assessed many candidates from Singapore, Zürich, ETH Hönggerberg campus, Delft, Rotterdam, to London. However, it is still challenging to find out the complete datasets contain meter data and spatial data of the same area. Table 3 shows that Zürich is the most promising candidate by far. City of Rotterdam has many open source spatial data available online, but contacting utilities might be needed to access to smart meter data.

Table 3: Currently available datasets

City	Meter data	Building and spatial features	Simulation
Zürich	53 households 13 buildings (residential) 1 minute frequency 1 year period	3D Geometry LOD1(2) Building usage type Built year Refurbishment year Thermal system Number of occupants Occupants profile Thermal properties [a] Temperature Humidity Solar gain	LOD1 simulation
Rotterdam	-[b]	3D Geometry LOD1(2) Building usage type Built year	-[c]

[a] Inferred from archetype
[b] Need to contact utilities
[c] The setup workflow can generate simulation for any area of interest if dataset is available

7 Time planning

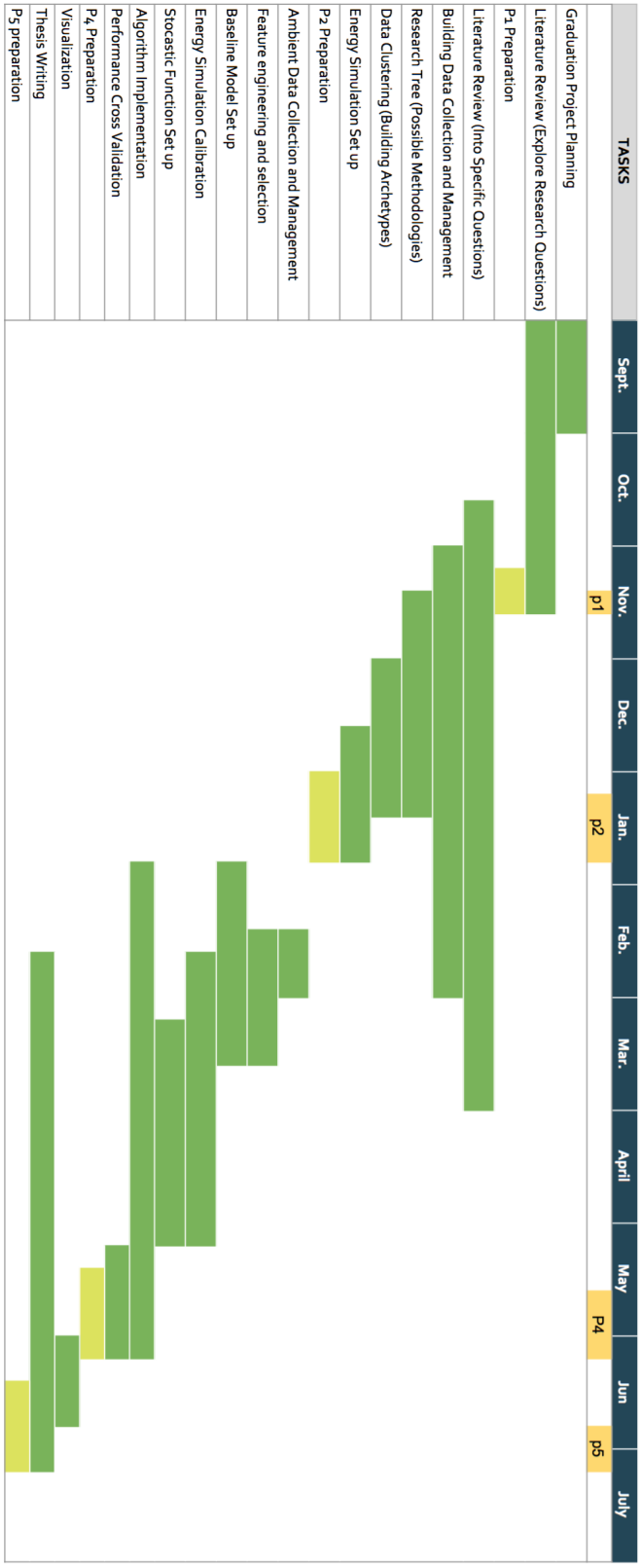

Figure 6: Time planning

References

[1] Alahakoon, D. and Yu, X. (2016). Smart Electricity Meter Data Intelligence for Future Energy Systems: A Survey. *IEEE Transactions on Industrial Informatics*, 12(1):425–436.

[2] Carrie Armel, K., Gupta, A., Shrimali, G., and Albert, A. (2013). Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy Policy*, 52:213–234.

[3] Deb, C., Zhang, F., Yang, J., Lee, S. E., and Shah, K. W. (2017). A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74(February):902–924.

[4] Fonseca, J. A., Nguyen, T. A., Schlueter, A., and Marechal, F. (2016). City Energy Analyst (CEA): Integrated framework for analysis and optimization of building energy systems in neighborhoods and city districts. *Energy and Buildings*, 113:202–226.

[5] Frankel, D., Heck, S., and Tai, H. (2013). Sizing the potential of behavioral energy-efficiency initiatives in the US residential market. *McKinsey & Company*, (November):6.

[6] Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. 22:679–688.

[7] International Energy Agency (2010). *Energy Technology Perspectives: Scenarios & Strategies To 2050*.

[8] Isard, M. and Blake, A. (1998). Condensation—conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28.

[9] Koponen, P., Mutanen, A., and Niska, H. (2014). Assessment of some methods for short-term load forecasting. *IEEE PES Innovative Smart Grid Technologies, Europe*, pages 1–6.

[10] Mathieu, J. L., Price, P. N., Kiliccote, S., and Piette, M. A. (2011). Quantifying changes in building electricity use, with application to demand response. *IEEE Transactions on Smart Grid*, 2(3):507–518.

[11] Miller, C. (2016). Screening Meter Data: Characterization of Temporal Energy Data from Large Groups of Non-Residential Buildings.

[12] Miller, C., Nagy, Z., and Schlueter, A. (2016). A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. *Renewable and Sustainable Energy Reviews*, (June):1–13.

[13] Molina-Solana, M., Ros, M., Ruiz, M. D., Gómez-Romero, J., and Martin-Bautista, M. J. (2017). Data science for building energy management: A review. *Renewable and Sustainable Energy Reviews*, 70(May 2016):598–609.

[14] Price, P. (2010). Methods for Analyzing Electric Load Shape and its Variability. *California Energy Commission*, (May):1–63.

[15] Reinhart, C. F. and Cerezo Davila, C. (2016). Urban building energy modeling - A review of a nascent field. *Building and Environment*, 97:196–202.

[16] Rubel, F. and Kottek, M. (2010). Observed and projected climate shifts 1901 – 2100 depicted by world maps of the Köoppen-Geiger climate classification World Map of Köppen − Geiger Climate Classification. 19(2):135–141.

[17] Schofield, J., Carmichael, R., Tindemans, S., Woolf, M., Bilton, M., and Strbac, G. (2014). Residential consumer responsiveness to time-varying pricing. *Report A3 for the "Low Carbon London" LCNF project: Imperial College London*.

[18] Wate, P. and Coors, V. (2015). 3D data models for urban energy simulation. *Energy Procedia*, 78(0):3372–3377.

[19] Wijaya, T. K., Humeau, S. F. R. J., Vasirani, M., and Aberer, K. (2014). Residential Electricity Load Forecasting: Evaluation of Individual and Aggregate Forecasts. pages 1–22.

[20] World Business Council for Sustainable Development (2015). Transforming the Market: Energy Efficiency in Buildings.