Abstract. Achieving energy efficient urban planning requires a multi-disciplinary planning approach. The huge increase in data from sensors and simulations does not help to reduce the burden of planners. On the contrary, unfamiliar multi-disciplinary data sets can bring planners into a hopeless tangle. This paper applies semi-supervised learning methods to address such planning data issues. A case study is used to demonstrate the proposed method with respect to three performance issues: solar heat gains, natural ventilation and daylight. The result shows that the method addressing both familiar and unfamiliar data has the ability to guide the planner during the planning process.

Keywords. Energy performance; S3VM; decision tree; familiar and unfamiliar.

1. Introduction

In response to broad concerns regarding the environment and climate change, energy performance is becoming much more prevalent and increasingly in demand by occupants. URA (2012) pointed at these current challenges for Singapore: a) catering for economic growth and a good quality of life, b) maintaining a clean and green environment, and c) making the best use of our resources. Singapore has a tropical rain forest climate with no distinctive seasons; the typical day in Singapore has temperatures ranging from 23°C to 34°C and high relative humidity of about 84% (BCA 2010). Such high temperatures and humidity lead to a need to cool and dehumidify spaces to create comfortable conditions for occupants. However, the need for such climatic modification of spaces within a building consumes much energy. This has posed a great challenge for the buildings in Singapore to balance the energy performance with the responsibility to the environment.

To create an energy efficient urban fabric in Singapore that can not only minimise the impact on the environment, but also remain practical, economical and comfortable for use, it is important to look into integrated urban planning processes, as well as consider each aspect of urban space in an integrative and holistic
manner. More than 80% of Singapore’s resident population live in public residential buildings. These buildings are developed and managed by the Housing and Development Board (HDB), and are all under temporary leaseholds. In order to reduce the energy demand of the residential estate, the initial plan establishes the orientation, massing and location of the buildings, all of which impact and set the parameters for energy efficient planning strategies. There are three primary issues to consider (BCA 2010, 2012; URA 2012):

1. Solar Heat Gain. Direct solar radiation increases the cooling load and hence energy use. In naturally ventilated spaces, solar heat gains heat up spaces such that they typically become thermally uncomfortable to occupants. The first step in minimizing solar heat gains is to optimize the orientation and massing of a project specific to its location. Certain orientations (east and west for example) provide more exposure to the sun and therefore greater heat gains. Also, the massing of a project could provide shade to itself or other blocks to further mitigate solar heat gains.

2. Natural Ventilation. Maximizing the amount of space to be naturally ventilated is another strategy towards reducing energy demand on a project since natural ventilation requires little energy use as compared to air-conditioning. Establishing and understanding prevailing wind directions and how they work on a specific site will affect massing and orientation decisions.

3. Daylight. It is also important to take advantage of and harness natural daylight for spaces. This reduces the need for artificial lighting which requires significant amounts of energy. Bringing in daylight via window openings at appropriate heights, skylights and/or atrium spaces are all effective strategies that will affect massing and orientation decisions too.

This research reconsiders multi-perspective planning data sets and classifies the data into familiar and unfamiliar groups with respect to the three primary issues outlined above. Semi-supervised support vector machines (S3VMs) are proposed in order to grapple with the familiar and unfamiliar data sets during the planning process. The proposed method includes the following steps:

1. Variables and parameters. The sufficient urban variables and parameters from the three issues are identified from literature review.

2. Similarity (Fréchet distance). The similarity between the familiar and other data variables (population, weather, economic, etc.) is checked.

3. S3VMs. Semi-supervised support vector machines are applied to these data.

4. Decision tree (C4.5). The planning decision is evaluated by the calculation of entropy (information gain ratio).

The method is applied to a simple case study and the result reveals that S3VMs have the ability to improve the data analysis to enhance the urban planning process. The remainder of the paper is organized as follows: first, the literature review is presented; next, the proposed method is described; subsequently, a case study for public housing is demonstrated by the method; finally, a discussion and future work are presented.
2. Literature Review

This section reviews the data methods for dealing with urban variables from an energy performance perspective.

2.1. URBAN VARIABLES

Following on the energy crisis from the end of the seventies, many design projects and researches were carried out to understand energy consumption for urban design. Knowles & Ralph (1981) discussed architectural and urban design applications from the variables of solar envelopes. Subsequent researches started to investigate the energy performance of urban design considering climate variables (Akbari & Taha 1992; Owens & Susan 1986). From the mid-nineties, designers and researchers begin to consider energy performance not only from a single perspective such as climate, morphology, etc., but also from multiple perspectives. They realized that the multifaceted relationship between energy performance variables and urban environment variables is the key to promote energy efficiency within urban practice. The whole urban system comprises many systems which are too complex to be quantified and represented in numbers and models (Yeang 1995). However, it appears there is no limit to include numerous variables into the analysis to quantify the urban impact on energy performance and vice versa. With the development of information technologies, the regression models for understanding design variables became more complex. Givoni (1998) and Littlefair et al. (2000) conducted research in both building and urban design located in different climate regions. They refined the design methods from an environment perspective and generated regression models for human comfort and the effects of urban form on climatology. Kikegawa et al. (2003) used the observed data with regression models to address the significance of regional meteorological conditions and its interactions with buildings on evaluating impacts of urban-heat-island on buildings’ energy demand on a citywide scale. Ratti et al. (2005) pointed out that urban designers/researchers need to understand the building energy use as a comprehensive regression model of urban form, building design, energy system efficiency and occupant behaviour.

In order to address the abundancy of variables, new data analysis methods were proposed for the design process as well. Hanna (2007) implemented several techniques from machine learning and space syntax to define architectural archetypes. Liu & Seto (2008) built a method to simulate and predict urban growth from historical urban growth data. Gil et al. (2009) implemented a data mining method to extract descriptions of street and block typologies using attributes related to the morphology and density of urban blocks and street mobility. D’Oca & Hong (2014) developed a framework combining statistical analysis with two data mining techniques, clustering and association rules, to identify occupant behaviour patterns of window opening and closing in a naturally ventilated office building.

In conclusion, while supervised learning methods (regression models) are extensively used to deal with the abundance of design data/variables, and recent research explores the possibility of unsupervised (data mining) learning methods to handle with these kinds of data, there is lack of research on applying semi-supervised learning methods to grapple with both familiar and unfamiliar data sets.
2.2. DATA MINING

Data mining appears as a research discipline in the early 1990s (Piatetski & William 1991). It is the core stage of the knowledge discovery process that is aimed at the extraction of interesting and implicit information from large data (Fayyad et al. 1996). Data mining can be considered a superset of many different methods: statistical methods and machine learning (Jackson 2002). Due to this sense, data mining could be categorized into 1) supervised learning (predictive), 2) unsupervised learning (descriptive), and 3) semi-supervised learning (predictive & descriptive). The goal of supervised learning is to learn a relationship between input and output data (Kotsiantis 2007). Unsupervised learning aims to find interesting structures in the data (Ghahramani 2004). The target of semi-supervised learning is to enhance the accuracy of both predictive and descriptive processes (Chapelle et al. 2006). The comprehensive fundamental methods and algorithms have been given elsewhere (Chapelle et al. 2006; Zhou & Li 2010; Zhu 2008). They identified the popular methods such as expectation maximization with generative mixture models, self-training, co-training, transductive support vector machines, and graph-based methods.

This research addresses the question whether semi-supervised learning is meaningful to solve the large data problems during the urban planning process. Specifically, in comparison with a supervised algorithm that uses only labelled (familiar) data, can one hope to have a more accurate prediction by taking into account the unlabelled (unfamiliar) data points? In principle, the answer is “yes”. However, there is an important prerequisite: that the distribution of examples, which the unlabelled data will help elucidate, is relevant for the classification problem.

3. Research Methodology

The proposed method includes four iterative steps (figure 1): 1) Planning data sets. This data includes the influencing data from different domain subjects. 2) Similarity (Fréchet distance). The similarity between the familiar and other data variables (population, weather, economic, etc.) is checked. 3) Semi-Supervised Support Vector Machines (S3VMs). This classification method is carried out to separate the result of similarity as positive or negative energy performance. 4) Decision Tree (C4.5). All variables are recalculated by their information gain ratio with respect to the three issues: solar heat gains, natural ventilation and daylight. This result will help planners to identify which variable is the most important in the current planning stages.

3.1. DESIGN AND PLANNING DATA SETS

In order to demonstrate the proposed method clearly, the data sets are narrowed down to massing geometry variables and environmental variables, in relation to the three issues identified: solar heat gains, natural ventilation and daylight: 1) physical variables: massing orientation, massing height, massing location, site coverage; and 2) climate variables: solar, wind.
3.2. SIMILARITY (FRECHET DISTANCE)

The Frechet Distance was first defined by Maurice Frechet in 1906 as a measure of similarity between two parametric curves (Buchin et al. 2008; Eiter & Mannila 1994). Subsequently, it has become a standard measure between parametric curves used in many areas. The Frechet distance is typically explained as the relationship between a person and a dog connected by a leash walking along two curves and hoping to keep the leash as short as possible. The maximum length the leash reaches is the value of the Frechet distance. The standard definition for the Frechet distance (Alt & Godau 1995): Given two curves, $A, B$ in a metric space, the Frechet distance, $d_F(A, B)$ is defined as

$$d_F = \max_{t \in [0, 1]} \{d(A(t)), B(t)\}$$

where $\alpha, \beta$ range over all monotone parameterizations and $d(.,.)$ represents the Euclidean distance, and $f$ is the infimum.

3.3. SEMI-SUPERVISED SUPPORT VECTOR MACHINES (S3VMS)

Semi-Supervised Support Vector Machines (S3VMs) are developed from Support Vector Machines (SVMs). SVMs rely on training data to generate a separating hyper plane that splits the given data into two different classes. SVMs are formulated into optimization problems in order to find a series of weights and a constant $b$, which together represent the separating plane. Such a decision boundary is defined as (Bennett & Demiriz 1999):

$$f(x) = w^T x + b$$

where $w^T$ is the parameter vector that specifies the orientation and scale of the decision boundary, and $b$ is an offset parameter.

However, traditional SVMs require the data to be labeled before classification analysis. Considering the limitation of the SVMs, this research adopts the S3VM to deal with the data during the design process, specifically, the software SVM-light (Joachims 1999).

3.4. DECISION TREE (C4.5)

The decision tree method C4.5 is proposed for the classification issues, because it adopts a simple hierarchical structure that aids user understanding and decision making. The C4.5 algorithm includes the information gain ratio concept, which is defined as follows (Quinlan 1993):
\[ GAIN_{Ratio}(p, T) = \frac{GAIN(p, T)}{SplitInfo(p, T)} \]  
\[ SplitInfo(p, T) = -\sum_{j=1}^{n} p'(\frac{j}{p}) \cdot \log(p'(\frac{j}{p})) \]  

where \( p'(\frac{j}{p}) \) is the proportion of elements present at the position \( p \), taking the value of the \( j \)-th test. The information gain ratio is independent of the distribution of examples inside the different classes.

4. Public Housing Estate Planning Case Study

We propose an estate planning case study at an empty site in the Punggol district of Singapore. From the URA master plan, the area has a maximum 3.0 stipulated gross plot ratio (GPR).

The proposed method assumes a distinction between familiar and unfamiliar data. To simplify the matter, the planners are assumed to be only familiar with parameters such as massing height, orientation, and locations. Instead, weather data (temperature, humidity, solar, etc.) and sociodemographic data (monthly income from work, marital status, age group, tenancy, household type, etc.) are grouped with the unfamiliar data sets. The design requirements reflect on solar heat gain, natural ventilation and daylighting. Therefore, planners need to understand the sun path and wind speed. In Singapore, the sun is almost directly overhead throughout the year since Singapore is located only 1º north of the equator. East and west orientations receive the most solar exposure and therefore have the most potential for solar heat gains. The Sun Path diagram for Singapore (figure 3a) indicates that both north and south orientations also receive solar exposure for a portion of the year. In Singapore, wind directions are predominantly N-NNE and S-SSE throughout the year depending on the monsoon season. Although Singapore generally has low wind speeds, the velocities achieved are enough to provide comfort to spaces with the help of optimised planning (figure 3b). In order to calculate the similarity, two vector curves are created: 1). Sun path curve as a horizontal line; 2). Wind direction curve line (N-NNE and S-SSE) (figure 4). Note that these two curves are only for illustration of the planning case; actual planning requirements may prescribe different curves.

Figure 2. Singapore Climate. a) Sun Path diagram b) Annual wind rose (m/s) (1982-2015).
Considering the formulated curves, the planners begin to sketch the configuration with respect to either orientation, height or location. For instance, from a single massing sketch, the massing orientation can be identified. In the second step of the proposed method the similarity is calculated between the orientation curve and the two formulated curves (figure 2). The similarity is specified as a value between 0 and 1, with a smaller value indicating higher similarity. For the sun path, a higher similarity means the sketch plan has a low possibility to achieve energy efficiency at this stage. But for the wind direction, more similar curves represent a higher possibility to achieve energy efficiency.

The following step uses the software SVM-light (Joachims 1999) to determine which option is positive (high possibility of energy efficiency) or negative (low possibility of energy efficiency) clustering. The similarity of orientation is provided as familiar/labelled input data. At the same time, another similarity calculation between orientation and monthly energy consumption data is provided as unfamiliar/unlabelled input data. The monthly energy consumption data is based on HDB (Treelodge) which was awarded the Green Mark Platinum Award. The similarity results of the unfamiliar/unlabelled data help the S3VM to set the boundaries of positive energy performance clustering. Options 3 and 4 are clustered positively (figure 3). In the last step, the decision tree algorithm recalculates the similarity deviation between energy consumption and the solar heat gain, natural ventilation and daylighting variables. The planning parameters (orientation, height, location) are ranked by the information gain ratio of the similarity results. For both the third and fourth options, the decision tree lists the parameters in the order of location, height and orientation (figure 3).

Location and height can be determined in a similar way. The sketch of location is the sum of similarity of every single massing (figure 4); the curve of height is created by every massing’s highest point (figure 5).
The sequence of sketching and analysis illustrated in figures 3-5 does not have priority and can be altered by the planner. Considering the decision tree results, figure 3 could be linked to either 4 or 5 and vice versa. The process is also iterative; for the first option in figure 6, the decision tree also suggests to recheck the orientation. The trade-offs between different aspects are represented by the similarity values derived from the calculation step. This method reduces the burden of decision-making related to energy performance calculation for urban planners.
Comparing the final massing at this stage with the existing high energy performance HDB development in Singapore, we can conclude that the patterns of the massing design variables (orientation, location, height) are highly similar. This demonstrates that the proposed method distinguishing familiar and unfamiliar data has the ability to support planners/designers.

5. Conclusion
The objective of this research is to create a data analysis method to help the planner solve the multi-perspectives problems during urban planning. The proposed method mainly includes three parts: similarity calculation, classification (S3VM) and decision tree (C4.5). The important point is that the original data is separated into two groups: familiar and unfamiliar. The case result study reveals a typical residential planning pattern in Singapore. Hence it also proves the feasibility and possibility to apply this method into current planning processes. This demonstration is limited to the energy domain. Future research will extend the method to other perspectives: accessibility, traffic, etc. Additionally, the proposed method will be further developed to become readily usable as an urban planning support tool within research, education and, possibly, practice.

Acknowledgements
The authors would like to acknowledge URA and NEA for providing us with residential design guidelines and the raw weather data.

References
27th eCAADe Conference, 269–278.


