Rethinking the Urban Design Process from a Data Perspective

Yuezhong Liu¹, Rudi Stouffs², Abel Tablada³
¹,²,³ National University of Singapore
¹ liuyuezhong@u.nus.edu ²,³ {stouffs|abel}@nus.edu.sg

Urban design always requires the processing of large amounts of data from multi-disciplinary sources during the decision-making stages. However, unfamiliar multi-disciplinary data sets can only lead to confusion and uncertainty. This research proposes a data-driven approach for supporting the urban design process. A hybrid data mining method is used to cluster, classify and rank solution-instances according to geometrical properties and energy performance. An urban design case study is used to demonstrate the proposed method with respect to two performance issues: solar heat gains and natural ventilation. The result shows that the method addressing both familiar and unfamiliar data can effectively guide the designer during the design process.

Keywords: energy performance, S3VM, decision tree, familiar and unfamiliar

1. INTRODUCTION

Urban design is traditionally a heuristics problem solving process, involving different data acquisition and analysis techniques with the aim to establish a more evidence-informed design process. Schön (1983) described the design process as exhibiting “see-move-see” cycles, which involve preparing drafts to be evaluated, revised, and refined. Lawson (2005) further pointed out that urban design as a mechanism includes three parts: analysis, synthesis and evaluation. Meanwhile, the urban design process always includes some gratuitous proposals, which might bring a degree of randomness during the design stages. Though there has been substantial research on this iterative urban design process, the general conclusion has been that systematic design methods are not applicable to urban design (Cross 2006).

The rapid urbanization in the tropics has brought many challenges for urban designers to deal with the relationship between humans and the physical environment. Of growing concern to urban designers is the rapid change in urban climate associated with urbanization. At the same time, urban energy performance is highly impacted by the urban climate. And urban design has a significant influence on energy consumption for buildings and transportation, irrespective of the socio-economic context (Emmanuel 2005). Hence urban energy performance becomes an important aspect during the design process. Research of urban energy performance in the recent past has revealed a series of causal relationships between a wide range of urban variables and climate (Ratti et al. 2005; Stewart and Oke 2012). These range from urban geometry to human behaviour, from anthropogenic heat to thermal characteristics of urban
surfaces, and from obstruction of wind flow to lack of vegetation. As concluded by Emmanuel (2005), from the climate perspective, there are mainly two mitigation strategies for pursuing energy efficiency during the urban design process at the neighbourhood/district level: 1) radiation reduction during the day, and 2) ventilative cooling at night.

Furthermore, as in other hot-humid tropical climate contexts, urban design in Singapore needs to consider the impact of a number of variables to achieve energy efficient design, including: 1) location; 2) building/site orientation and dimensions; 3) size and height of individual buildings; 3) existence of high-rise buildings; 4) street orientation; 5) availability, size and distribution of open areas and green belts (Givoni 1998). However, the research problem remains how to manipulate urban geometry to achieve energy efficient design with respect to the “move” step in the urban design process.

Hence the objective of this research is going to develop a hybrid data mining method for design data to guide the design process on energy-related urban performance aspects. Furthermore, the design data also relate to many other planning data and policies, thus including both familiar and unfamiliar data. Hence, to achieve this aim, we will reconsider and improve energy performance assessment during the urban design process from a data perspective.

The proposed method is expected to guide and inform urban designers to achieve energy efficient designs through an intuitive process. Each of the iterations will output a recommendation list of the design parameters for reference, which is sorted by energy performance priority. Furthermore, the output will identify which performance aspect should be focused on at the next design step. As a proof of concept, this research adopts an urban design studio to demonstrate the proposed method. This paper also expands on the problem statement and gives a brief overview of the literature review on urban variables. The model sites which are used in this case study could be replaced or modified by the designers for different design contexts. It will also outline and discuss further developments necessary to fully address the problem statement identified and to become readily usable as an urban design support tool within education and, possibly, practice.

2. LITERATURE REVIEW
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 and Taha 1992; Owens and Susan 1986). From the mid-nineties, designers and researchers began 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, Baker and Steemers (2005) pointed out that urban design-
ers/researchers need to understand the building energy use as a comprehensive regression model of urban form, building design, energy system efficiency and occupant behaviour.

To address the abundance 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 and Seto (2008) built a method to simulate and predict urban growth from historical urban growth data. Gil, Montenegro, Beirao, and Duarte (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 and 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 deal with both labelled (familiar) and unlabelled (unfamiliar) data sets. Hence this research addresses the question of how semi-supervised learning methods can solve data issues during the urban design process. Specifically, in comparison with a supervised algorithm that uses only familiar data, will the semi-supervised learning methods have a more accurate prediction by considering the unfamiliar data points as well?

3. RESEARCH METHODOLOGY
The proposed method includes four iterative steps (Figure 1): 1) Design Variables & Parameters. This data includes the influencing data from different domain subjects. The variables and parameters will change between different design stages. 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 two issues: solar heat gains and natural ventilation. This result will help designers to identify which variable is the most important in the current design stages.

3.1 DESIGN VARIABLES AND PARAMETERS
The influence of urban variables on energy performance is obvious from previous researches (Bueno et al. 2012; Ignatius et al. 2016; Jusuf et al. 2007; Ratti et al. 2005; Wong et al. 2011), and the key urban design and planning strategies to reduce energy use and thermal discomfort variables are: site coverage ratio, façade to site ratio, sensible anthropogenic heat, albedo (the ratio of reflected radiation from the surface to incident radiation) and emissivity (relative ability of the surface to emit energy by radiation).

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, Buchin, and Wenk 2008; Eiter and 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 and Godau 1995): Given two curves, A, B in a metric space, the Frechet distance, $d_F(A, B)$ is defined as:

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

where $\alpha, \beta$ range over all monotone parameteriza-
tions and $d()$ represents the Euclidean distance, and inf 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 hyperplane that splits the given data into two different classes. SVMs are formulated into optimization problems to find a series of weights and a constant $b$, which together represent the separating plane. Such a decision boundary is defined as (Bennett and 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 labelled 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):

$$\text{GainRatio}(p, T) = \frac{\text{Gain}(p, T)}{\text{SplitInfo}(p, T)}$$

$$\text{SplitInfo}(p, T) = -\sum_{j=1}^{n} \left\{ p'(\frac{j}{p}) \cdot \log \left( p'\left(\frac{j}{p}\right) \right) \right\}$$

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. CASE STUDY - RETHINKING URBAN PRACTICES FOR JURONG VISION 2050

The case study for this research stems from an international, collaborative design studio (Winter School) involving over 170 students and 30 design tutors, organized by the International Forum on Urbanism (IFoU), in which the first author participated as designer from energy performance perspective. The brief was to develop proposals for the transformation of Jurong Industrial Estate (JIE), a 5000-ha industrial area in the west of Singapore from an almost mono-functional, segregated and fragmented, polluted industrial area into a major catchment area for future population growth that integrates clean industrial plants with green lungs, attractive housing and vibrant urbanity for one million people. During the IFoU winter school, designers were divided into
teams, who then worked intensively together to develop urban visions and proposals for a scenario of 100% renewable energy system by 2050.

The design case study focuses on sustainable energy production, consumption and distribution within the built environment. The design process is elaborated on two levels: On a macro scale (urban level), new approaches of energy exchange, circulation and balancing are investigated, while on a micro scale (precinct and block level), design explorations focuses on energy efficient building structures, natural ventilation, optimal daylight access and the integration of technologies for sustainable energy production. To identify the features of the proposed method, two design scenarios are described and compared in the following paragraph.

4.1 CONVENTIONAL DESIGN METHOD
4.1.1. Macro-scale design stage. The designers follow the idea of urban acupuncture to deal with the energy issues. Just as the practice of acupuncture is aimed at relieving stress in the human body, the goal of urban acupuncture is to relieve stress in the built environment. Due to the five-element theory of Chinese philosophy, the designers chose five sites from this area which also was figurative as “material energy” system. Each site is improved from the concept of the definition of the element. In this design, the definitions of the elements are: “earth” represents underground space, “metal” means reinforced frame pier, “wood” refers to green space and residential area, “water” implies hydropower and “fire” indicates the energy plants. In the meantime, designers carefully consider energy consumption flow through analysis of aggregate urban environmental, social, economic and ecological factors. The final decision of the selected sites is shown in Figure 2.

4.1.2. Micro-scale design stage. The designers selected one node “wood” for the integration of industrial and residential areas. Two main aspects were considered for the design: 1) radiation reduction during the day, and 2) ventilative cooling at pedestrian level and indoors. Designers compared the site with Punggol New Town which is developed as an energy...
efficient residential area in Singapore, and completed
the masterplan and concept design as shown in Fig-
ure 3a. Furthermore, a building typology is proposed
based on their energy performance as shown in Fig-
ure 3b.

4.2 HYBRID DATA MINING METHOD

Before applying the hybrid data mining method to
the design data, the designers need to specify what
is the familiar data and unfamiliar data. In this case
study, the familiar data will always refer to the spatial
variables from macro-scale to micro-scale. The unfa-
miliar data will include the population structure, eco-
nomic structure and industrial structure from Pung-
gol New Town (Table 1).

Hence, based on the different design solutions,
the final estimated values of the three categories will
always be compared and classified during the S3VM
step. The detailed design scenarios are described be-
low.

4.2.1. Macro-scale stage. Considering the preview
dataset, we denote the designer not as an expert but
as a novice who only is familiar with the urban en-
environmental data which contains fundamental infor-
mation of geography and transportation. The nodes
selection will go through the proposed method: 1) Fa-
miliar design variables: crossroads points (more
than four boundaries), bus stop points. 2) Similarity:
Because designers want to identify the isolated sites,
the number of bus stops inside the circle should be
as few as possible. And the curvature of the roads will
also be compared with an empty crossroad. The ra-
dius of the circle is 200 m. 3) S3VM: compare the se-
lected nodes for their prominence with the social and
economic data sets. 4) C4.5: Ranking the select sites
based on the information gain ratio. The detailed cal-
culation steps are shown in Figure 4. Population den-
sity is analysed in Figure 5. The final important nodes
are also shown in Figure 5.
Table 1
The urban design data from energy perspectives in Punggol New Town.

<table>
<thead>
<tr>
<th>Ethnic Group</th>
<th>Chinese</th>
<th>Malays</th>
<th>Indians</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>75.35%</td>
<td>15.83%</td>
<td>6.88%</td>
<td>1.94%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marital Status</th>
<th>Single</th>
<th>Married</th>
<th>Widowed</th>
<th>Divorced</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.36%</td>
<td>75.91%</td>
<td>3.50%</td>
<td>4.23%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Religion</th>
<th>No Religion</th>
<th>Islam</th>
<th>Hinduism</th>
<th>Buddhism</th>
<th>Taoism</th>
<th>Sikhism</th>
<th>Catholic</th>
<th>Christens</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>21.47%</td>
<td>10.25%</td>
<td>4.95%</td>
<td>34.50%</td>
<td>9.41%</td>
<td>0.72%</td>
<td>6.63%</td>
<td>11.58%</td>
<td>0.48%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Highest Qualification Attained</th>
<th>No qualification</th>
<th>Primary</th>
<th>Lower Secondary</th>
<th>Secondary</th>
<th>Post-secondary</th>
<th>Polytechnic</th>
<th>University</th>
<th>Professional</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.90%</td>
<td>4.26%</td>
<td>6.19%</td>
<td>15.10%</td>
<td>8.39%</td>
<td>13.42%</td>
<td>34.45%</td>
<td>9.29%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age Group</th>
<th>0~4</th>
<th>5~9</th>
<th>10~14</th>
<th>15~19</th>
<th>20~24</th>
<th>25~29</th>
<th>30~34</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.68%</td>
<td>8.95%</td>
<td>5.00%</td>
<td>3.38%</td>
<td>3.13%</td>
<td>6.27%</td>
<td>16.38%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Monthly Income from Work</th>
<th>&lt;1k</th>
<th>1k~1.5k</th>
<th>1.5k~2k</th>
<th>2k~2.5k</th>
<th>2.5k~3k</th>
<th>3k~4k</th>
<th>4k~5k</th>
<th>5k~6k</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.97%</td>
<td>5.31%</td>
<td>4.64%</td>
<td>4.98%</td>
<td>5.80%</td>
<td>11.77%</td>
<td>10.28%</td>
<td>10.12%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Industry</th>
<th>Manufacturing</th>
<th>Construction</th>
<th>Wholesale &amp; Retail Trade</th>
<th>Transportation &amp; Storage</th>
<th>Accommodation &amp; Food Service</th>
<th>Information &amp; Communication</th>
<th>Financial &amp; Insurance Service</th>
<th>Real Estate Service</th>
<th>Professional Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>6500</td>
<td>2700</td>
<td>10300</td>
<td>6100</td>
<td>2800</td>
<td>3100</td>
<td>5100</td>
<td>1400</td>
<td>5400</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Administrative &amp; support Service</th>
<th>Public Administration &amp; Education</th>
<th>Health &amp; Social Services</th>
<th>Arts, Entertainment &amp; Recreation</th>
<th>Other Community, Social &amp; personal Services</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>2200</td>
<td>8600</td>
<td>2500</td>
<td>1100</td>
<td>2300</td>
<td>700</td>
</tr>
</tbody>
</table>
ilar sites in Punggol New Town as a model, two sites Treelodge@Punggol and Punggol Water Park are selected to compare for similarity. The curvature is generated by linking all the centre points of the residential areas. The drawing with the highest similarity value is selected as site coverage proposal. S3VM will help to classify whether the proposed site coverage is sufficient for the future increased population and economic. Finally, the decision tree will rank the design parameters for designers to consider for the next step. Figure 6 lists a single initiative progress of the site coverage decision-making stage. After the decision of site coverage is confirmed, the decision tree suggested to consider massing layout in a single

---

### Table 2

<table>
<thead>
<tr>
<th>Location</th>
<th>Tree coverage</th>
<th>Average building height</th>
<th>Site coverage ratio</th>
<th>Facade-to-site ratio</th>
<th>Sensible anthropogenic heat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Punggol New Town</td>
<td>0.19</td>
<td>26</td>
<td>0.38</td>
<td>1.55</td>
<td>commercial area: 10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>residential area: 4</td>
</tr>
</tbody>
</table>

---

Figure 5
The five important nodes identified by proposed method.

Figure 6
Decision of site coverage process by proposed method.
plot. Also from Table 2, there is a limitation for the façade-to-site ratio. Because the single plot area is 6 ha, and the perimeter and area of a typical residential building in Singapore is 90 m and 700 m², it means that if the height is 26 m, the maximum number of buildings in a single plot is 39. Hence here the single plot also is separated into 600 sub slots (100 m² per slot). Then the decision of massing layout also becomes a mathematic problem which is how to insert the massing to achieve the maximum energy performance and comfortable again. The plot models are also selected from the Punggol New Town area, and the calculation process is shown in Figure 7. Hence the same calculation process is also applied on the decision of tree coverage. And the final micro-scale design proposal is shown in Figure 8.

4.3. COMPARISON AND DISCUSSION

From the above mentioned urban design case study, the conventional design method and the proposed hybrid data mining method are compared in Figure 89. There are two important points which should be discussed further. First, both design processes used the same available datasets which are collected from

---

**Figure 7**  
Decision of massing layout by proposed method.

**Figure 8**  
The final micro-scale site design proposal for massing structure and layout.
public websites. The only difference is the design method. The conventional design method could be extended to any other regression models (like Gravity Model, Urban Energy System), which require the designers to pass all the assumptions as input data to evaluate the design solutions. But the proposed data mining method will allow the designers to improve the design process by skipping the data which is unfamiliar by them. The proposed method allows designers to simply focus on the design variables which are familiar to them, while keeping track of the design quality through the unfamiliar data. Hence back to the design “see-move-see” cycles, the proposed method removes the burden of data from multi-disciplinary sources. Secondly, considering the design results from the two methods, both design proposals were evaluated with respect to the total PV cell area and total energy consumption at the site. The present total energy consumption of the Jurong Industrial Estate is 2567.84 Ktoe (EMA 2016). And the annual average solar radiation on tilted panels (shadings not included) is 1663 kwh/m2. Then the conventional design method achieves the reduction of energy consumption by 75%, and the energy production of PV could cover 70% of the total energy consumption in the year 2050 with a total panel area of 3352140.379 m2. The PV area is estimated from built-up (roof, façade, etc) and open space (8% of natural space). In the meantime, the final design by the data mining method reduces present energy consumption by 79.1% and increases the energy production of PV (3686820.70 m2) to 85% of the final estimated energy consumption in 2050. Hence the results demonstrate the possibility of applying this hybrid data mining method as a complementary decision-making approach for energy performance aspects in the urban design process. The proposed method is able to prioritize the design variables to guide design-
ers at the micro-scale stage. Furthermore, the conventional design result needs an expert to explain the design proposals and the audience also needs the relevant knowledge to understand the evaluation progress. On the contrary, the design proposals from the second method could be explained with data analysis results as evidenced by novices. In summary, this proposed data mining method is able to make the urban design process from an energy perspective more trackable and transparent.

5. CONCLUSION

The objective of this research is to create a hybrid data mining method to help designers solving decision-making problems during the urban design process. The proposed method mainly includes three parts: similarity calculation, classification (S3VM) and decision tree (C4.5). The main difference with the conventional regression model is that the design data is separated into two groups: familiar and unfamiliar. The proposed data mining method could allow designers to improve their design without fully understanding all the knowledge. The case study also shows the possibility to apply this data mining method as a complement considering energy performance into urban design processes.

In the future, we will evaluate this method with other design perspectives: accessibility, mobility, greenery etc. Additionally, the proposed method will also be further developed to become readily usable as an urban design support tool.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Jurong Town Corporation (JTC) and International Forum on Urbanism (IFoU) for founding us the studio’s design research in January 2015. We thank all the other professors and students who were involved the energy team: Prof Thorsten Schutze, Prof Kang Min-Jay, Dr Adam Rysanek, Gemma Salinas Escayola, Ana Marie Piedra Leon, Chang Yushi, Loh Sze Sian, Tulika, Chen Qianwei, Penny Ng Ru Yi, Guan Li, Sara Mamou, Andhi Prasetyo, Choi Jong Won, Calvian Oscar, Yujin Lee, Chen Qieyi, Zhang Siqi, Leena Sahu for providing the design ideas and images.

REFERENCES


D’Oca, S and Hong, T 2014, ‘A data-mining approach to discover patterns of window opening and closing behavior in offices,’ *Building and Environment*, 82, pp. 726-739

Eiter, T and Mannila, H 1994, ‘Computing Discrete Fréchet Distance,’ *Notes*, 94, p. 64


Gil, J, Montenegro, N, Beirão, J and Duarte, J 2009
‘On the Discovery of Urban Typologies. Data mining the multi-dimensional character of neighbourhoods,’ *Proceedings of the 27th eCAADe Conference*, pp. 269-278


Kikegawa, Y, Genchi, Y, Yoshikado, H and Kondo, H 2003, 'Development of a numerical simulation system toward comprehensive assessments of urban warming countermeasures including their impacts upon the urban buildings' energy-demands', Applied Energy, 76(4), pp. 449-466


Lawson, B 2005, How designers thinks: the design process demystified, Elsevier


