HOLISTIC VISUAL DATA REPRESENTATION FOR BUILT ENVIRONMENT ASSESSMENT

JUNG MIN HAN & NAMJU LEE
Harvard Graduate School of Design, USA.

ABSTRACT

With an increasing interest of big data and its analysis, urban planners and architects use manifold data from different sources as an indicator of urban phenomena. To analyse them, machine learning models have been gotten fame to evaluate complex urban features with correlation matrix and cross validation. There are two major types of data such as top-down and bottom-up data. Interestingly, there is a third category of data that has not been widely deployed yet which we define as the 'Holistic visual data'. It is the general impression to your visual perception of urban environment when you at a specific spot that we believe can be captured by Google Street Views. This paper aims to ask how do top-down, bottom-up, and holistic visual data work individually or together to predict the built environment value in urban areas. Admittedly, housing price is a highly spatial metric that involves multifarious interests including macroeconomic policies, the development of the area, the local people, and specific houses' features. To investigate the quality of built environment, machine learning techniques are adopted for different groups of datasets. By comparing several regression and classification models for each groups of data such as top-down, bottom-up and holistic visual data with housing price, the Random Forest model could be proposed as a best model. The intricate urban matrix needs to be organized such an order, but there are multiple factors affecting urban issues including land values and housing prices. By utilizing relevant machine learning models, it can be provided the practical guideline to architects and designers for improving the quality of built environment.

Keywords: bottom-up data, classification, data science, feature selection, holistic visual data, housing price prediction, machine learning, random forest regressor, top-down data, urban analysis.

1 INTRODUCTION

The new and immerse location-aware urban datasets, so-called spatial big data, are emerging from diverse organization and private sectors and can be investigated to find meaningful patterns for use in urban design and planning. Urban patterns provide designers with an option to make data-driven decision making process for urban design and planning. Spatial big data includes top-down and bottom-up data as well as holistic visual data, which has accessibility from individuals. However, the multifarious interests on such datasets cause difficulties in developing plausible analysis between different features and characteristics. Therefore, among hundreds of datasets, critical features can be selected after applying given data process and visualization in urban context. Sensitive datasets across different types (top-down, bottom-up and holistic visual data), which have both positive and negative relationships to target values, could be visualized for further design or strategic interventions. By considering urban visual information such as Google Street View, designers deal with manifold datasets for their urban spatial analysis. With comparison between selected features and results of feature selection using Scikit learn (sk-learn, widely deployed machine learning package, [1]), the validity of using proposed components for urban big data analytics could be achieved. In this paper, different types of datasets and methodology for extracting such datasets will be introduced, and analysis techniques and results will be demonstrated.

ISSN: 1743-7601 (paper format), ISSN: 1743-761X (online), http://www.witpress.com/journals
DOI: 10.2495/SDP-V13-N4-516-527

2 OBJECTIVES

Land and housing price is a highly spatial metric that involves multiple interests including macro-economic policies, the development of the areas, the residents and the urban footprints. Defining city data such as top-down and bottom-up data is becoming feasible by disclosure of private data such as energy use, housing value, as well as building description with a manifold urban metrics such as land use and commercial information. The city infrastructural data including specific housing features, and urban use is classified as 'top-down data', and informal crowd-sourcing data such as, Twitter texts, Instagram tags, or Craigslist is categorized as 'bottom-up data'. There is a third category of data that has not been widely deployed yet which we define as the 'holistic visual data'. It is the general perception of your visual surroundings when you at a specific spot that we believe can be captured by Google Street Views. To date, visual information has potential to evaluate living environment having potential and additional descriptive narrations. There is a need to quantify potential implementation of such data widely. By demonstrating on the street views as dependent variables, missing information could be equipped as a significant feature in a housing value prediction analytical model. This paper aims to integrated holistic visual data with conventional datasets to predict housing value in the city of Boston by utilizing machine learning models, and demonstrate the best model among different estimators.

3 METHODOLOGY

3.1 Data exploration

Admittedly, spatial big data are increasingly becoming available to researchers and designers. Data processing and analysis are important in many different urban implementations such as transit management, mobility, social activity and affordability research. However, big data management reveals challenges on mining and analysing data due to its huge size and missing trajectories. Visual representation of spatial big data could show a great potential as they can intuitively present on the original geo-location allowing users to explore different feature sets both qualitative and quantitative ways.

As data preparation, general housing information and informal values are extracted using Zillow Inc [2], Craigslist [3] housing information, crime rates and school districts from the official website (Boston Open Data, [4]) that have geolocation information. In the respects of holistic visual data, this paper utilizes Google Application Programing Interface (API) [5] and collects street views. To capture the images of 360 degrees of panoramic views in Boston, discretized cell of grid is generated, and those images are used for semantic analysis with deep learning (Fig. 3).

Main datasets could be divided into three datasets: Baseline, Socio-economic, and Holistic visual dataset. For the bottom layer of the datasets, the (1) Baseline dataset: Room Type, Bathrooms size and so on, obtained from Geographic Information System (GIS) (Fig. 1), Zillow (Fig. 2), and Trulia API [6], representing quantitative perspective of the Boston area, is defined. The second layer of the main datasets is (2) Socio-economic information of the city, including walk-abilities of Massachusetts Bay Transportation Authority(MBTA), school, crime, retail, and so on. The last layer of the main datasets is (3) Holistic visual dataset: the percentage of tree, grass, sky, building, ground, river, or bridge in urban context.

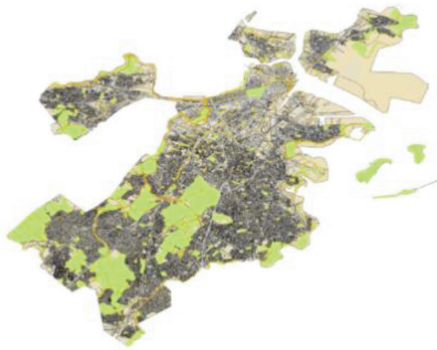

Figure 1: GIS visualized data of Boston downtown.

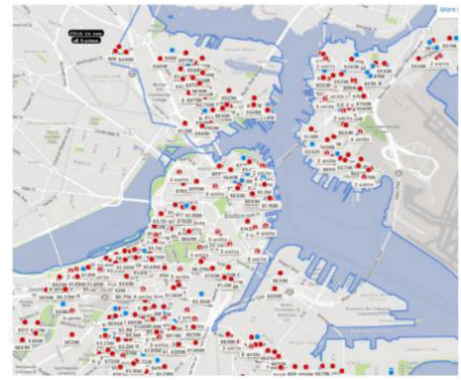

Figure 2: Boston downtown housing price data (Zillow Inc, [2]).

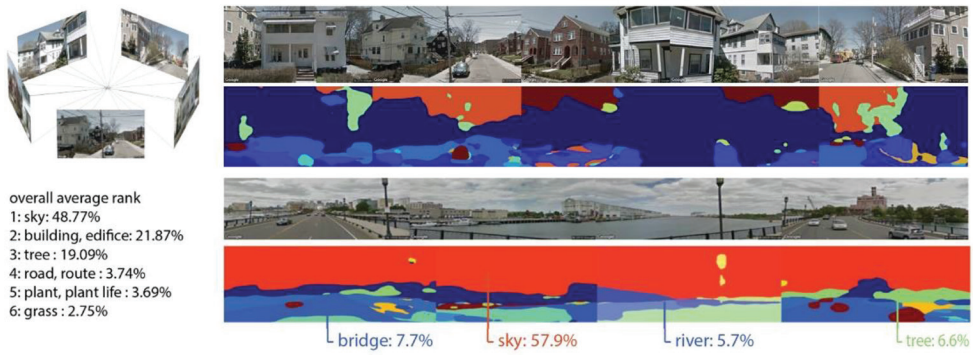

Figure 3: Semantic classification utilizing deep learning.

3.2 Data processing

To create comprehensive data process pipelines, and to prepare for machine learning with proper data, graph and pixel data structure are harnessed to produce the data since individual data sets need to possess explicit and implicit spatial relationship each data sets. Because certain data, such as transportation, need to be represented as a discrete point, while others such as crime data need to be organized as a continuous information.

Basically, pixel data structure, a two-dimensional matrix (Fig. 4), is utilized for the Baseline because the pixelated Boston area (Fig. 5(a)) makes it possible to compute the data which has a gradual impact on its neighborhood. For example, a point of crime has its effect around the point, and continuously disappears as they are further away from the point. On top on the pixel grid, graph data structure (Fig. 5(b)), a mathematical object consisting of set of discrete points and edges visualizing the topological aspect such as street network of urban, highway or the subway map, is overlapped for the data which require explicit relationships such as MBTA, retail, or school.

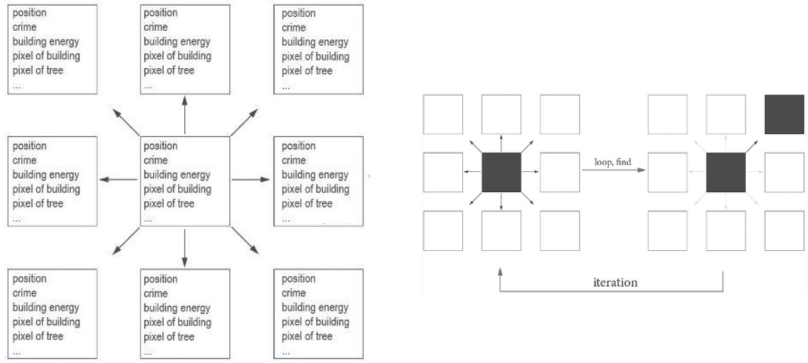

Figure 4: Diagram of pixel structure.

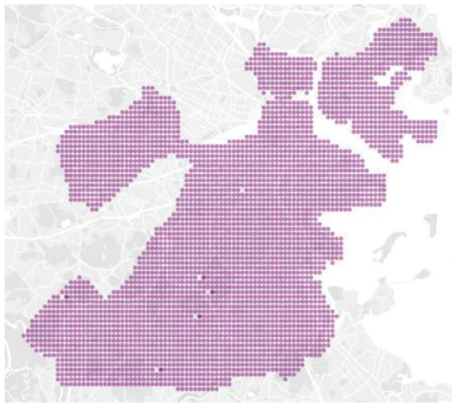

Figure 5 (a): Pixelated Boston areas.

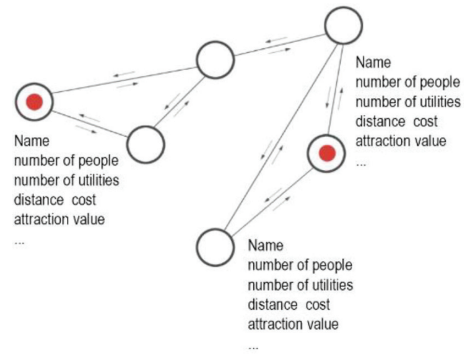

Figure 5 (b): Diagram of graph structure.

Finally, the both data structures provide a framework for holistic computation of the spatial data based on different relationships. Each point of pixel and graph has data and its connectivity, for instance (1) a price of house, a number of room and crime (2) distances to schools, parks, MBTA, etc. and (3) a ratio of visual urban perception from images. Therefore, this process promotes to recreate summarized and normalized data feature sets, integrating and representing areas as numerical values for training machine in the next chapter.

4 MACHINE LEARNING

In this chapter, the overall data process pipeline is introduced which consists four different parts: the feature selection, the regression analysis, the classification analysis, and the result comparison. The Boston housing and rent price are estimated by conducting different types of machine learning analysis and best analytical model is going to be demonstrated.

4.1 Feature Selection with Boston housing data

Boston's average housing price is around \$58/m^2 and the median is about \$60/m^2 while it has a range from \$9.4/m^2 to \$447/m^2, which ranks higher than national standards at \$13.2/m^2.

Baseline features from top-down datasets include longitude, latitude, zipcode, bedrooms, bathrooms, house type and land size. After conducting correlation analysis of baseline data, the result of the baseline model yields 0.32 in R-squared of the test set. To improve given model with Baseline, we added bottom-up and holistic visual data, which encompass walking distance to MBTA, walking distance to school (k-12 education), walking distance to university, walking distance to park, crime rates, land energy use, craigslist house posting, craigslist room posting as well as Google Street View data. These factors appear to be crucial prediction features, especially in a cultural city with strong academic resources like Boston, school districts appeared to be a major factor (Fig. 6).

Interestingly, additional features of holistic visual data reveal a correlation with housing price at 'pixelWater', 'pixelVan', 'pixeCar', 'pixelRoad', 'pixelSideWalk', 'pixelSky'. As a result of Forward and Backward selection of comprehensive features; 'pixelWall' 'pixelCeiling' 'pixelPath' 'walkSchool' 'walkMBTA' 'energySiteEUI' 'walkPark' 'pixelBridge' 'pixelWindow' 'pixelGrandstand' 'latitude' 'bathrooms' 'bedrooms' were selected while 'pixel Window' is the only different feature from the two selections. The combination of distinctive features shows higher R-squared at 0.55, which reveals meaningful usage of combined data.

Lasso and ridge model have similar R-squared compared to the linear regression model. This is because both model pick up more features than the backward and forward selection. After tuning alpha parameter, better r squared at above 0.6 was achieved. By using regression method, we found that the linear regression with feature selection can predict the housing price with an R-squared around 0.55 while random forest regression method with depth of 5 have an R-squared of 0.59. These have much improved from the baseline model (housing features only). Adding the urban infrastructure data and holistic visual data, the model accuracy has improved. Among those features, walking distance to MBTA and walking distance to park appeared important in all regression model. Urban environment features such as public stands, public paths, denser construction are associated with higher housing value. More covered ceiling and more bridges are associated with lower housing value. Figure 7 shows the similarity and the discrepancy between original values from the test set and the predicted values after the cross validation.

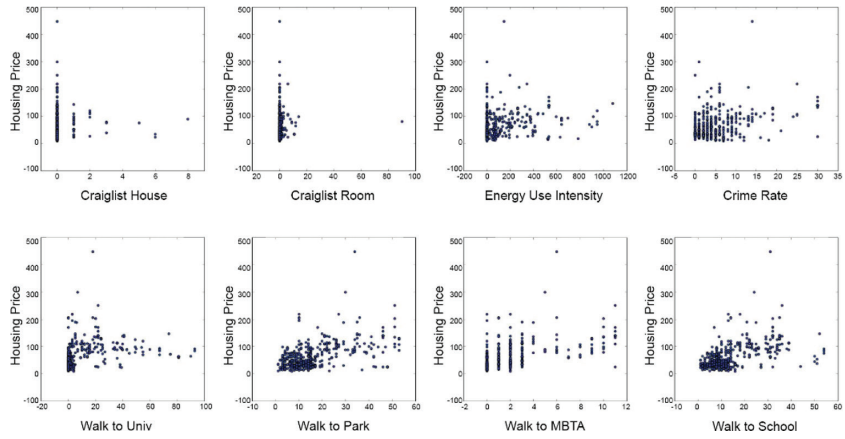

Figure 6: Correlation Analysis among features and housing price.

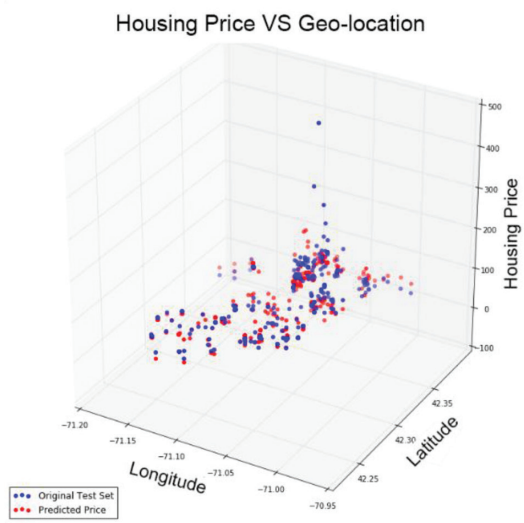

Figure 7: Predicted values and values from the original test sets.

In next section, the large number of datasets including rent data from Boston is utilized to find optimized regression model for the comprehensive analysis. Following analysis could become an important factor when urban designers apply housing values as well as rent values into their design decision making.

4.2 Advanced regression analysis with Boston rent data

For more comprehensive data analysis, 13,049 datasets from the Trulia was added on top of existing feature selection. Among increased datasets, features consist of three main class: web scraping from Trulia, google place information, socio-economic data, and google street view data (holistic visual data). After post processing of collected datasets, 11,569 datasets with 49 features without any missing properties are obtained. Unlike a housing data, rent datasets do not have a correlation between features and predictions. This means that it is difficult to predict rent price than the housing price with given features obtained from diverse sources. With processed rent datasets, initial R-squared from linear regression reflects at -2.09. With forward and backward search, R-squared have increased up to 0.34 and 0.34 respectively. As you can see the Table 1, 'pixelWall' and 'pixelWater' shows a positive relationship while 'pixelField', 'pixelSky', 'pixelBus', 'pixelCeiling' reveals a negative relationship with rent values for Boston housing. This represents that the amount of exposed walls and the water are important to predict higher rent price, while the number of bus and exposed sky and field reflect lower rent price. Furthermore, when it comes to R-squared, Lasso and Ridge regression with cross validation still shows lower values at 0.34, which is a similar value for feature selection. In this section, more advanced machine learning algorithm called Random Forest Regressor (RFR) is proposed for better prediction.

The RFR is advanced machine learning algorithm developed from simple decision tree method. The RFR allows us to find important features by stepped down depth analysis. By utilizing RFR, R-squared at 0.64 is achieved with analysis on important feature sets. In

Table 1: Positive and negative coefficient of selected holistic visual data.

	Linear Regression	
	Forward selection	Backward selection
Positive Coefficient	'pixelWall' 'propertiesAsses' 'pixelWater' 'crime' 'walkPark' 'walkUniversity' 'pixelBridge' 'Longitude' 'Bathrooms'	'pixelWall' 'propertiesAsses' 'pixelWater' 'crime' 'walkPark' 'walkUniversity' 'pixelBridge' 'Longitude''Bathrooms'
Negative Coefficient	'pixelBus' 'pixelCeiling' 'walkSchool' 'walkMbta' 'pixelField' 'pixelSky' 'pixelBuilding' 'Zip' 'RoomType'	'pixelBus' 'pixelBus' 'pixelCeiling' 'pixelBuilding' 'walkSchool' 'walkMbta' 'pixelField' 'Zip' 'RoomType'

Figure 8, 'pixel Sky', 'pixel Building', 'pixel Park', 'pixel Bridge', 'pixel Ceiling, 'pixel Bus' and 'pixel Water' still show high correlations for the rent price prediction.

Figure 9 shows rent prediction histogram of given datasets (left) and predicted values (right) by using RFR. Both distribution shows highly well-predicted distribution at the same level of accuracy of the regression analysis in Boston housing data.

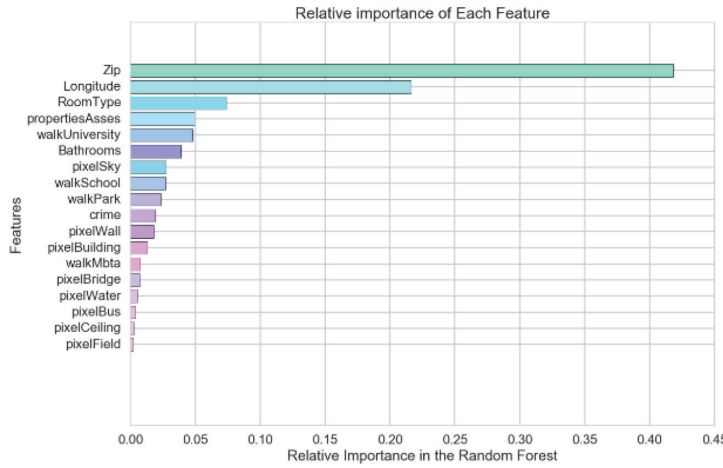

Figure 8: Relative importance of features from random forest regressor.

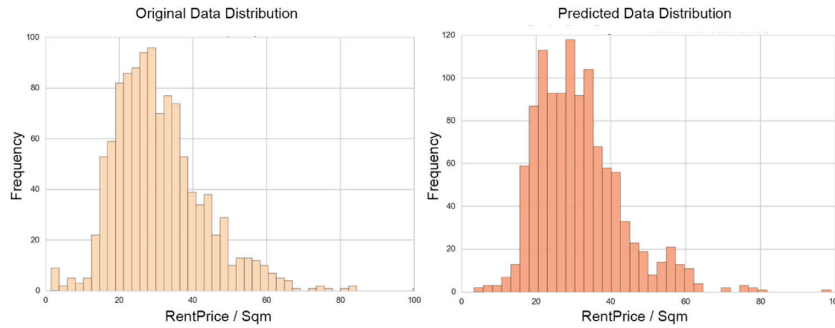

Figure 9: Comparison distribution between original (left) and predicted (right) datasets.

4.3 Classification analysis

Housing values is sometimes meaningful when we know estimable values based on several classes. In this section, this paper is going to focus on three classes of housing values: high (75 percentile), middle (50 percentile), and low (25 percentile), which is shown in Figure 10. Unlike the previous regression analysis, the classification method aims to predict rent values based on the given three classes. By knowing the class of the rent price, urban designers can actively adopt this information for infrastructure and amenities' design.

Firstly, the Principal Component Analysis (PCA) is implemented for three different classes. For PCA, the logistic regression model is utilized, and the results of analysis represent a model accuracy at 0.38. To increase the accuracy of the classifier, a cross validation method for decision tree is conducted. The prediction accuracy with decision tree model reveals at 0.59 for training datasets, and at 0.58 for testing datasets respectively. Additionally, retrieved accuracy with the max depth at 4 reaches accuracy at 0.62 for training and at 0.61 for testing datasets. This means that by tuning the parameters of PCA model more accurate prediction values could be achieved.

Lastly, the Random Forest Classifier (RFC) and combination of PCA are introduced to get more accurate data predictions. To achieve higher accuracy, we increased number of estimators of RFC from 2 to 7, thus results from RFC in higher accuracy on test data at between 0.74 and 0.77. Figure 11 shows important pixel data such as 'pixel Plant', 'pixel Pole', 'pixel Person', 'pixel Tree', 'pixel Water', 'pixel Road, 'pixel Path', 'pixel Bus', 'pixel Building', 'pixel Grass', 'pixel Wall' and 'pixel Car'.

4.4 Result analysis

Based on the regression analysis in sections 4.1 and 4.2, we found that certain features among entire datasets have impact on housing value prediction. Those features were categorized into three different categories: Baseline component, Socio-economic component, and Holistic-visual component. To create comprehensive datasets after feature selection, exhaustive search, forward selection, backward selection, and cross validation was used. Finalized feature sets for this study are listed in the Table 2 below.

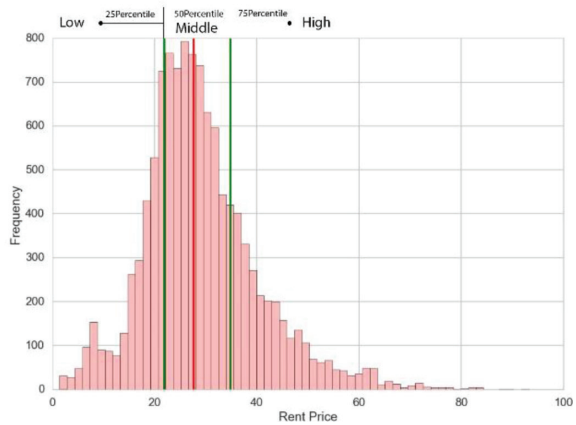

Figure 10: Percentile data for classification in rent data distribution.

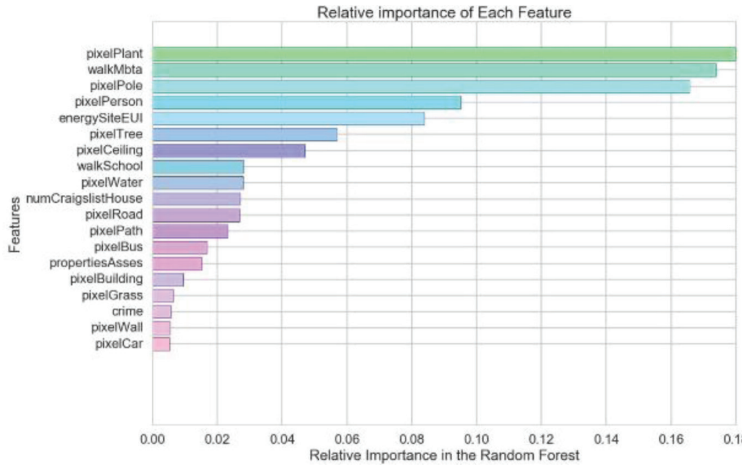

Figure 11: Relative importance of features from random forest classifier.

Table 2: Selected features by different categories

Exhaustive Search/Forward Selection/Backward Selection/CV	
Baseline	'Latitude', 'Longitude', 'RoomType', 'Bathrooms', 'SQFT', 'Zip'
Socio-economic	'walkMbta', 'walkUniversity', 'crime', 'numCraigslistRoom', 'numCraigslistHouse', 'propertiesAsses',
Holistic visual	'pixelCar', 'pixelRoad', 'pixelGrass', 'pixelWall', 'pixelHouse', 'pixelLake'

Comparison study between all types of regression model is basically done by adding and replacing selected features between different components. For instance, the Baseline features were run, and Socio-economic data were added. To avoid the biased result from Baseline datasets, only Socio-economic data and Holistic visual data was separately evaluated. Finally, total proposed features out of entire datasets were selected by adding and removing each component.

As described in Table 3, the combined Baseline and Socio-economic data gives a better prediction compared to Baseline. To make more precise model, Holistic visual data was

Table 3: Comparison analysis between different regressors.

Selected Features	Lin_Reg()	Lasso_Reg (alpha = 0.01)	Ridge_Reg (alpha = 1000, cv = 9)	Random Forest Regressor (n_estomators = 22, max_depth = 30, random_state = 2)
Baseline	0.367	0.371	0.376	0.371
Socio_added	0.417	0.451	0.456	0.096
Socio_Holistic	0.272	0.295	0.302	0.434
Total	0.414	0.466	0.456	0.615

added on top of both thresholds. When only added visual data, the performance of the model is not improved, but combined visual data into merged model (baseline and socio-economic), the best performance was observed in both linear and random forest regression. By tuning parameters with CV optimization, the finalized model with all three datasets shows best performance. Overall, when we combined all possible components, the performance of regressors reveals higher accuracy from 0.414 to 0.615. Among different regression models, Random Forest regressor performs best at 0.615.

As discussed classification in section 4.3, the comparison analysis is conducted with more classifiers. Compared to PCA with logistic analysis, other classifiers can predict housing value better at around 0.65 accuracy. With a manipulation of parameters, Random Forest classifiers with 7 estimators perform well at 0.78 of accuracy on test sets (Table 4), which is almost double that of logistic regression with PCA (3 components). Besides random forest classifiers, Gradient boosting classifiers and Ada Boosting classifiers have relatively higher accuracy than logistic regression, but the computation time of cross validation is heavy than other classifiers. Therefore, for the classification of housing value prediction, Random Forest classifier is highly recommended in terms of both performance and computational efficiency.

In conclusion, Random Forest Model of both regression and classification was chosen as the best estimators. It can be assumed that the tree models work better for predicting housing values with visualized data of built environment. To use multiple features including holistic visual data and environment information for value prediction (can be categorized by geo location), Random Forest model is strongly recommended.

5 DATA VISUALIZATION

In design practice, data visualization is an imperative part enhancing intuitive data analysis with spatial information. In this chapter, visual analysis and spatial mapping by utilizing Grasshopper and Rhinoceros (parametric and architectural modelling tool) is going to be introduced. By visualized urban data, urban designers and architects could easily access and interpret enormous number of datasets.

Figure 12 is a visualized map for rental data in Boston areas. In this map, designers could assume that the highly rated rent values are concentrated on downtown areas, and the low rental values are sparsely distributed across the neighborhood. Initially, it can be assumed that the accessibility to schools, university, MBTA and park affects housing value. However, the walkability to MBTA shows negative relationships for value prediction. To validate assumptions and results of machine learning, the visualized data in different location could be an estimator for value prediction for designers.

Additionally, two of distinctive features are Energy (Fig. 13) and Crime data (Fig. 14). Both are positively correlated with the housing values. Therefore, among multiple features of socio-economic data, the site Energy Use Intensity (EUI) data and the crime rate data could be a positive estimator for the value prediction.

Table 4: Comparison analysis between different tuned classifiers.

	Logistic Regression (PCA)	Decision Tree Classifier _Tuned	Ada Boost Classifier _Tuned	Random forest Classifier _Tuned
Train	0.388	0.69	0.89	0.95
Test	0.4	0.67	0.73	0.78

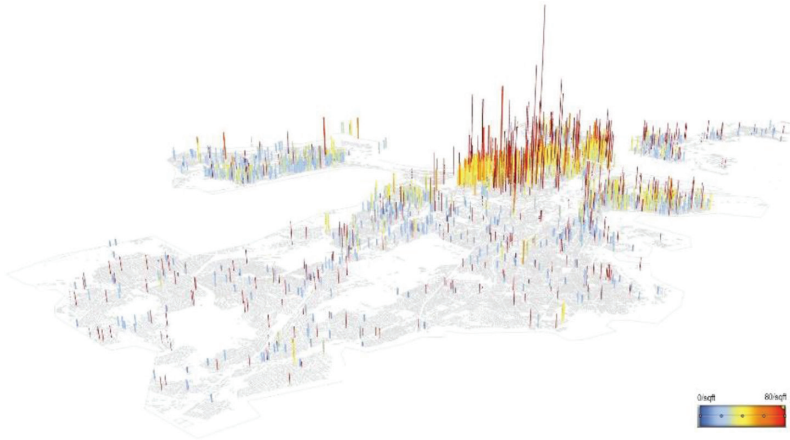

Figure 12: Visualized Rent data.

Figure13: Site EUI of building sectors.

Figure14: Crime rate within boundary areas.

There are several critical factors which have a positive influence on housing value predictions. Based on the results of feature selections (in chapter 4), it can be concluded that the river (represented as water feature) and bridges near the center of the downtown shows positive relationships (Fig. 15). While, a large amount of cluster of trees in low housing value areas from Figure 16 shows unclear boundaries for defining rent price, the information of grass and parks could be a great indicator to predict housing values with its accessibility.

Unlike the previous analysis, features which have negative relationship of housing values could be another indicator. Especially, the resolution of visualized buildings and path (Fig. 17) shows the areas where more affordable housing units and rent units are located. Besides buildings and city infrastructures, the visualized number of cars captured by Google street views represents lower price for housings (Fig. 18).

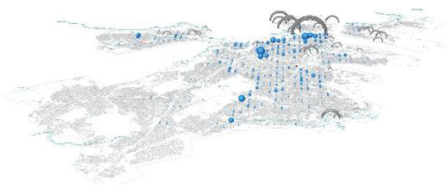

Figure 15: Visualization for ‘pixel Water’ and ‘pixel Bridge’.

Figure 16: Visualization for ‘pixel Trees’ and ‘pixel Grass’.

Figure 17: Visualization for 'pixel Building' and 'pixel Path'.

Figure 18: Visualization for 'pixel Car'.

While chapter 4 shows machine learning analysis techniques to have a sense of urban features and their relationship to the housing values, this chapter demonstrates visualized techniques to find the positive and negative relationship between numerical datasets in urban context. When designers obtained enormous urban datasets with limited data analytic background, they can utilize urban visualization techniques to figure out correlations between selected data and a target value.

6 CONCLUSION

In conclusion, an intricated urban matrix needs to be organized such an order, but there are multiple factors affecting urban issues including land values and housing prices. Traditionally, statisticians or data scientists have explored top-down and bottom-up data on various instances. However, architects and urban planners should incorporate the visual environment data in extensive ways. By demonstrating the use of holistic visual data, missing information could be equipped as a significant feature to predict built environment.

From this paper, it can be said that Random Forest model predict better to access built environment with holistic visual data. Even portion of selected features for each class shows that the importance of urban condition to estimate property values. Therefore, both visual analysis and contextual analysis are crucial to evaluate monetary values for residential units. The number of buildings, people, and cars could be the negative estimators, while the accessibility to greens, water, and bridges, the use of energy and the rate of crime can possibly indicate positive relation for Boston housing value prediction. Densely packed urban areas sometimes can be categorized with different portion between walls and sky with a simple visual analysis. All in all, it can be argued that the potential possibility to analyse certain values based on visually represented urban metrics is significant to estimate and evaluate the urban housing values.

REFERENCES

[1] Scikit learn. Available at: http://scikit-learn.org/stable/ (Accessed 14 September 2016).
[2] Zillow, Inc. "Real Estate, Apartments, Mortgages & Home Values." Zillow. Available at: https://www.zillow.com/ (Accessed 14 September 2016).
[3] Craig list data in Boston, Available at: https://boston.craigslist.org/ (Accessed 14 September 2016).
[4] Boston Open Data, Inc. "Boston Open Data". Available at: https://data.cityofboston.gov/ (Accessed 14 September 2016).
[5] Google Street View, Inc, google place information and Google street view API. Available at: https://developers.google.com/maps/documentation/streetview/ (Accessed 14 September 2016).
[6] Trulia housing rent data in Boston. Available at: https://www.trulia.com/ (Accessed 14 September 2016).