House Quality Index Construction and Rent Prediction in New York City with Interactive Visulization and Product Design

Xiang Shen* Shunyan Luo* Mingze Zhang*

Abstract

Housing is of primary importance for immigrants in New York City. The study analyzed the house conditions and price changes of residents in New York City from NYCHVS survey during the past 30 years. Firstly, a house condition index is summarized by dimension reduction approach. In addition, we established a two-stage predictive model on the rent with spatial temporal information as well as house condition. Data visualization is utilized to show immigrant's preference and provide information for new residents to settle down in the city.

Key Words: 2019 Data Expo, Index Construction, Spatial Temporal Regression, Data Visualization

1. Introduction

The New York City Housing and Vacancy Survey (NYCHVS) dataset is provided by the New York City Department of Housing Preservation and Development (HPD) and the U.S. Census Bureau. It is a relatively-large, triennial survey of the New York City housing stock and population since 1965. It has been used to measure the rental vacancy rate originally in order to determine the continuation of current rent regulation laws. NYCHVS also collects many useful data on housing including rent regulatory and home ownership status, structural conditions, unit maintenance and neighborhood conditions; crowding, rents, utility costs, type of heating fuel, rent/income ratios; owner purchase price and estimated value, mortgage status and interest rate, etc. Such variety leads to many researches on exploring NYCHVS. For example, Freeman and Frank (2004) utilized NYCHVS dataset to analyze gentrification and displacement of NYC in 1990's, revealing the relationship between the neighborhood gentrification in New York City and a lower propensity of disadvantaged households to move. Klye et al. (2007) used this survey to obtain the built environment and conducted research on the relationship with alcohol consumption in urban neighborhoods. Edward et al. (2005) used NYCHVS to explore the reason of the extreme high house price in Manhattan.

Throughout explorations on NYCHVS, housing quality has been found getting higher dramatically since the 1970s. However, some sectors of the housing stock continue to face poor conditions. Meanwhile, certain specific maintenance deficiencies sustain showing higher prevalence. To unify measuring quality of houses or apartments in New York City, a house quality index is required to be created to summarize the information of a series of house quality related indicators. The potential composite indicator is much easier to interpret than discovering trends in a series of variables. As a data product, it provides more user-friendly insights and better visualization.

To construct the index, people usually assign different weights to a series of components for reflecting their importance to the index. Various statistical methods have been used in combining information to aggregate indicators while no agreed methodology exists yet. Nando et al. (2005) summarized a bucket of statistical methods for constructing index and two main general types of methods are widely used. One is based on factor analysis

*George Washington University, 1918 F Street, NW Washington, DC 20052

method such as principle component analysis (PCA) or multiple correspondence analysis (MCA) which is also popular in related studies. Gitelman et al.(2010) employed PCA to build a road safety index regarding observed traffic accidents and drivers behaviors. Asselin (2002) build a poverty index using MCA when the data source comes from survey and most variables are categorical. In our study, these methods does not work well on the data and the detailed explanation is shown in the following section. The other one is to employ regression to build the weights with a target variable as response. An example is from Porter and Stern (2001) in which they build national innovative capacity index by a linear regression model. The method is also employed in this study with an addition of $L2$ penalty to improve the performance under the sparsity of the data. There are also several other methods in index construction, such as unobserved components models, analytic hierarchy process and even some machine learning dimension reduction methods like autoencoder and tSNE. They have their advantages in different types of data and the performance of these approaches will not be discussed here.

Besides index, spatio-temporal data regarding prices should not be ignored. Many methods have been proposed to deal with it. For example, Gneiting and Guttorp (2010) proposed non-separable spatio-temporal covariance functions and Gamerman (2010) proposed dynamic model formulations. However, these methods all have some drawbacks. Some of them require relatively complete observation matrices, while others do not allow for sufficiently complex spatio-temporal dependencies.

Here we use the model described in Lindstrm (2014). The model uses temporal basis functions to account for the temporal variability in data. To account for spatial variability in the temporal structure, Sampson et al. (2011) and Szpiro et al. (2010) suggested the basis functions be modulated by spatially varying coefficients, and the coefficients be modeled using universal kriging. Having used spatially varying temporal basis functions to account for temporal variability, the residuals are assumed to consist of mean zero spatially dependent, but temporally uncorrelated fields.

This paper is organized as following: a detailed description of data used in the study will be presented in section 2. Section 3 describes the procedure of index construction and section 4 introduces the predictive model. In section 5, we illustrate how we design the data product and interactively visualize the data.

2. Data Description and Preprocessing

The data in this study mainly comes from two different sources. The main table *nycRent* is generated from NYCHVS. The study only focus on rent records rather than mortgage records which leads to the *nycRent* dataset. On the one hand, the rent records takes around 2 thirds of the total survey and the covariates are more complete. On the other hand, the study concentrates more on the comparison between natives and immigrants and the data product intends to provide an tool for future New Yorkers.

The *nycRent* data mainly contains two different types of features other than time and location. The first part contains personal information of the householder, including age, gender, marital status, number of people in the household, household income as well as the place of birth of the householder and his/her parents. The place of birth information is helpful to decide the immigrant status of the household. 3 different types of immigrants, Native, First Generation Immigrant and Second Generation Immigrant, are defined based on that.

The second part is house quality related information which are answers from the householder according to the survey. The followings are main aspects of the variables recorded:

– Condition of Exterior Walls

- Condition of Windows
- Condition of Stairways
- Condition of Floor
- Amenities (wheelchair access, elevator access)
- Kitchen and Plumbing Facilities
- Accident Reported (water leakage, heating system breakdowns, etc)
- Structure of the Building
- Other Variables

The study employs data from 10 surveys from 1991 to 2017 while these surveys are not exactly the same. 52 variables of house conditions are selected in the above categories while the most variables are kept with cleaned records. A few variables are recoded since the code of the record are not the same between different years. In addition, some variables are merged after a close look of detailed survey records for different years in order to keep as much useful information as we could.

In addition, The geographical information of these sub boroughs in New York City is manually input as complement of the main table. The *geo* data contains longitude and latitude of the center of these sub boroughs which are extracted from Google Map. In addition, for the purpose of interactively visualizing the data, a recoding match from the borough and sub borough index of NYCHVS to index of borough in map is employed within the shinyapp.

3. Housing Index Construction

3.1 Failure of PCA and MCA

The first thing comes into mind when referring to index construction is factor analysis. Principle Component Analysis (PCA) is the most commonly used statistical procedure in dimension reduction which converts a series of possible correlated variables to a smaller set of linearly uncorrelated variables. An natural idea of index construction is to employ the first principle component which accounts for the largest variability of the data.

Similiar to PCA, correspondence analysis also reduces dimension by creating orthogonal components. Multiple correspondence analysis can be also regarded as a generalization of PCA which works better for categorical variable, especially in survey data while responses are usually categorical.

However, both PCA and MCA do not work for this data by the result in table 1. The first component only explains 10.81% and 6.84% respectively which indicates that the potential index built by the first component does not cover most of the information from the data. There are two possible reasons for that. After a graphical display of correlation matrix of data in Figure 1, the correlation between variables are not strong. A few clusters can be identified from the figure. However, there is no strong correlation in general view. In this case, since the purpose of PCA or MCA is to summarize information from correlated variables, information from uncorrelated data can not be summarized through factor analysis without loss of large proportion of information. Also, as shown in Figure 2, most of the 52 variables have zero values with more than 80% and the sparsity of the house quality issue matches the practical situation but makes it reluctant to employ PCA or MCA to construct index.

Table 1: Result of PCA and MCA

Method		1st Component	2nd Component	3rd Component
PCA	Proportion of Varinace	0.1081	0.07788	0.0585
	Cumulative Proportion	0.1081	0.18602	0.2445
MCA	Proportion of Varinace	0.06894	0.05184	0.0378
	Cumulative Proportion	0.06894	0.12078	0.15858

Figure 1: Display of Correlation Matrix of House Quality Variables

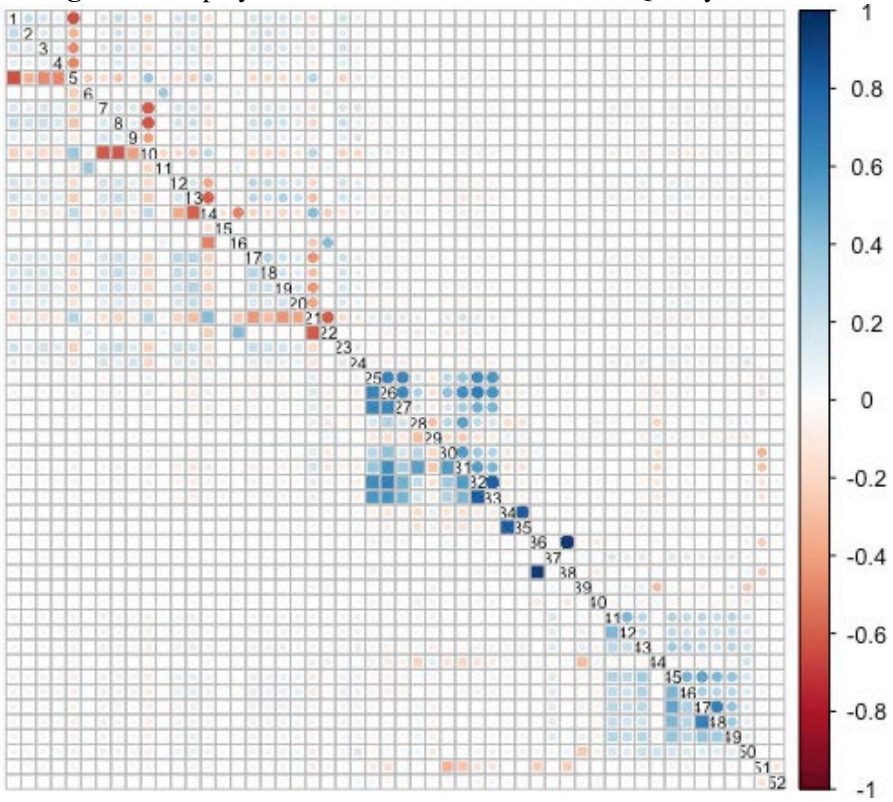
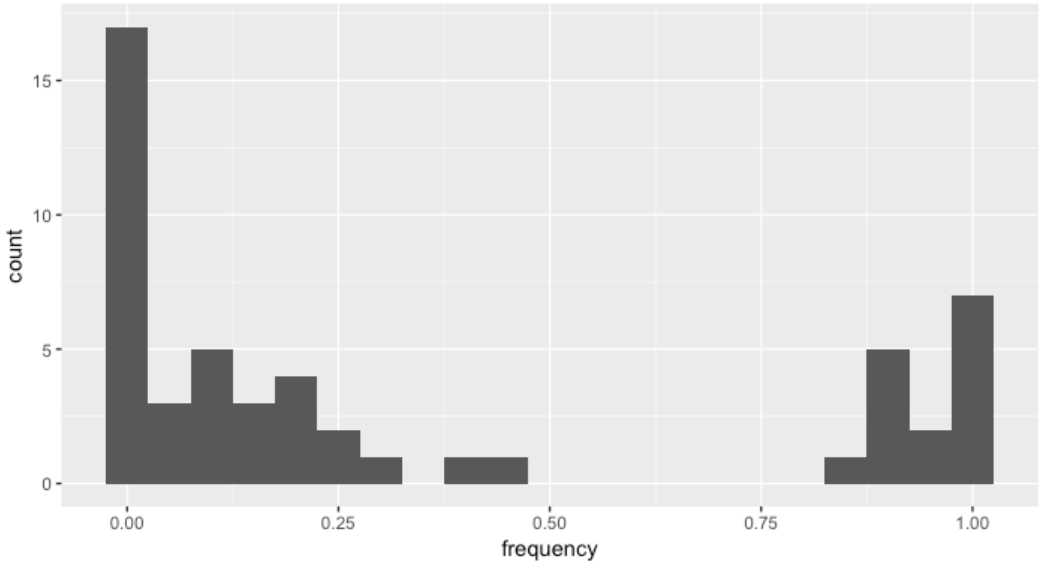


Figure 2: Frequency of Non Zero Responses



3.2 Ridge Regression Based Weighting

Linear regression models provides the linkage between a series of variables to one single response variable. If the target variable is denoted as Y and the candidate variables for constructing new index are denoted as $X_1, X_2, \cdot, X_k$. A linear model can be shown as:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon,$$

where βs are the coefficient of the model and ϵ are noise with mean 0 and constant variance.

After a collection of the data, least square can be used to estimate the coefficient βs which we minimize the loss function:

$$\beta = argmin_\beta (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta).$$

Then the fitted model is:

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k.$$

The index can be constructed by replacing the weights with coefficients as:

$$Index = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k.$$

Due to the sparsity of the data matrix, ridge regression (Hoerl and Kennard (1970)) is employed to improve the model. The motivation is to penalize the parameters but still keep all the weights non zero. Rent is a natural candidate for the response variable and the coefficient will be the weights for index construction. Similar to linear model but with a different optimization compared with equation 3.2, the solution of ridge regression is:

$$\beta = argmin_\beta (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) + \lambda \beta^T \beta$$

where λ is shrinkage parameter and a larger λ will lead to a smaller weights in index. The choice of $lambda$ is determined by a cross validation procedure with a selection of optimal

Mean Square Error (MSE). In the final step, we scale the index to a range of 0 to 1. A visualization work of the index in New York City can be shown in the later section of data product design.

4. Predictive Modeling

In the first stage, we fit a spatial-temporal model only using spatial and temporal information without any other covariates.

We first find the latitude and the longitude of the center of each sub-borough and use it as the coordinate. Then we take the mean value of the rents in each sub borough and use them to fit the following model:

$$y(s,t) = \mu(s,t) + \nu(s,t),$$

where $y(s,t)$ represents the spatio-temporal observations, i.e. rents, $\mu(s,t)$ is the structured mean field, and $\nu(s,t)$ is the space-time residual field. The mean field is modeled as

$$\mu(s,t) = \sum_{l=1}^{L} \gamma_l M_l(s,t) + \sum_{i=1}^{m}, \beta_i(s) f_i(t),$$

where the $M_l(s,t)$ are spatio-temporal covariates; γ_l are coefficients for the spatio-temporal covariates; $\{f_i(t)\}_{i=1}^{m}$ is a set of (smooth) temporal basis functions, with $f_1(t) \equiv 1$; and the $\beta_i(s)$ are spatially varying coefficients for the temporal functions. The $\beta_i(s)$ coefficients are treated as spatial fields with a universal kriging structure, allowing the temporal structure to vary between locations:

$$\beta_i(s) \sim N\left(X_i\alpha_i, \Sigma_{\beta_i}\left(\theta_i\right)\right) \quad \text{for } i = 1, \ldots, m,$$

where X_i are $n \times p_i$ design matrices, α_i are $p_i \times 1$ matrices of regression coefficients, and $\Sigma_{\beta_i}(\theta_i)$ are $n \times n$ covariance matrices. One of the advantages of this structure is that it allows us to make use of different covariates and covariance structures in each of the $\beta_i(s)$ fields, and the fields are assumed to be priori independent of each other.

The residual space-time field, $\gamma(s,t)$, is assumed to be independent in time with stationary, parametric spatial covariance

$$\nu(s,t) \sim N\left(0, \Sigma_\nu^t\left(\theta_\nu\right)\right) \quad \text{for } t = 1, \ldots, T \quad \text{and} \quad \nu\left(s_1,t_1\right) \perp \nu\left(s_2,t_2\right), t_1 \neq t_2.$$

Here we fit the above spatial-temporal model only using the spatial and temporal information without any other covariates, and follow the model fitting steps introduced in Lindstrm et al. (2014) with R package SpatioTemporal. During the model fitting process, the smooth temporal functions are relatively important. We run a leave one out cross-validation so that we can determine the number of smooth trends, and the results are shown in Figure 3.

However these kinds of overall statistics may not be enough for us to make a decision. So we also examine the pairwise scatter plots of all the leave-one-site-out BIC statistics for prediction of each site based on different numbers of trends in the smooth SVD model, which is shown in Figure 4.

Note that as we increase the number of trends, the sites don't behave equally. Some sites require more trends but some fewer. Here we use five smooth trends.

After we complete fitting the spatio-temporal regression, we can get the prediction for the mean of the rents in each sub borough, say R_1. Note that R_1 is the prediction using only the spatial and temporal information, which may not be very accurate. Besides, since

Figure 3: Result of Cross-validation

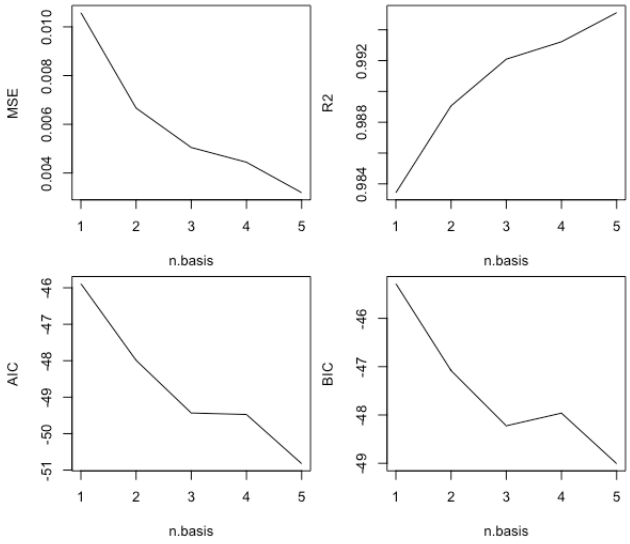
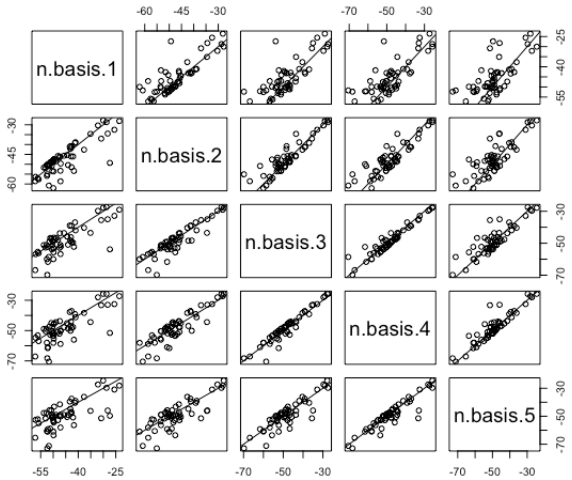


Figure 4: pairwise scatter plots of all the leave-one-site-out BIC statistics



the growth of the house rent is not quite uniform, there may even be some negative values in R_1 for the first year, which requires some modification and more information.

So in the second stage, in order to improve the prediction, for each observation, we fit a linear regression model using the following formula, including R_1 as one of the covariates:

$$Rent = \alpha R_1 + \beta \mathbf{X} + \epsilon,$$

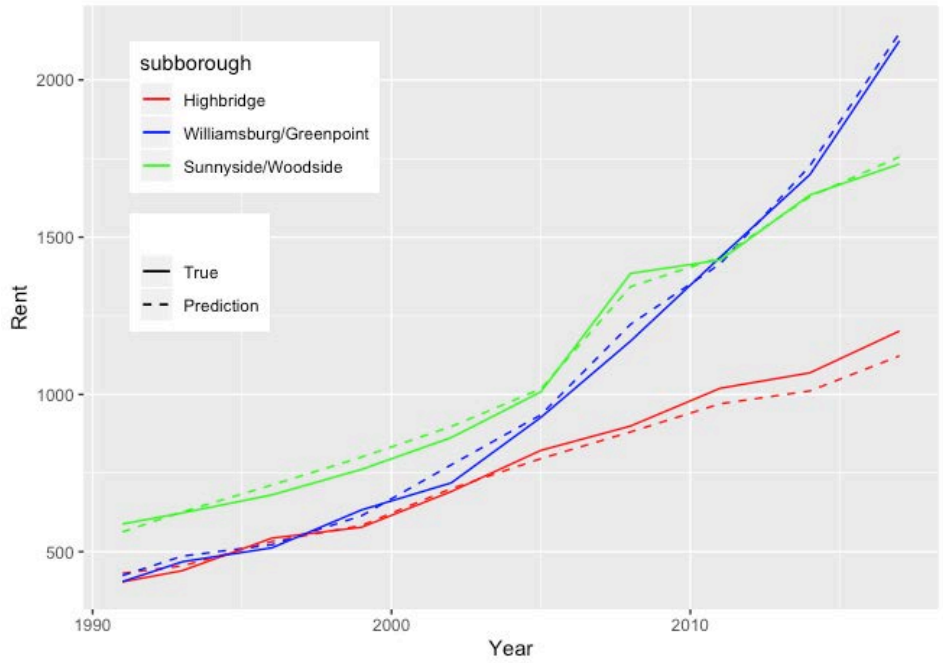
where X is all the other covariates, including number of persons in the house, the length of lease and the index that we summarized in Section 3. The fitted coefficients are quite significant as we show in Table 2.

Table 2: Regression Coefficients

	Estimate	Std. Error	t value	p-value
Intercept	-355.72	16.47	-21.60	<2e-16
Number of Persons	73.41	1.29	56.87	<2e-16
Length of Lease	27.19	2.67	10.18	<2e-16
Index	1187.27	34.12	34.80	<2e-16
Rent1	0.55	0.002	228.16	<2e-16

We select 3 sub-boroughs and show the curve of the house rents in Figure 5. The solid line is the true rent while the dashed line is our fitted rent. It is obvious that our model can fit the data quite well and the curve shows an increasing trend just as we expected.

Figure 5: True and Fitted Curves

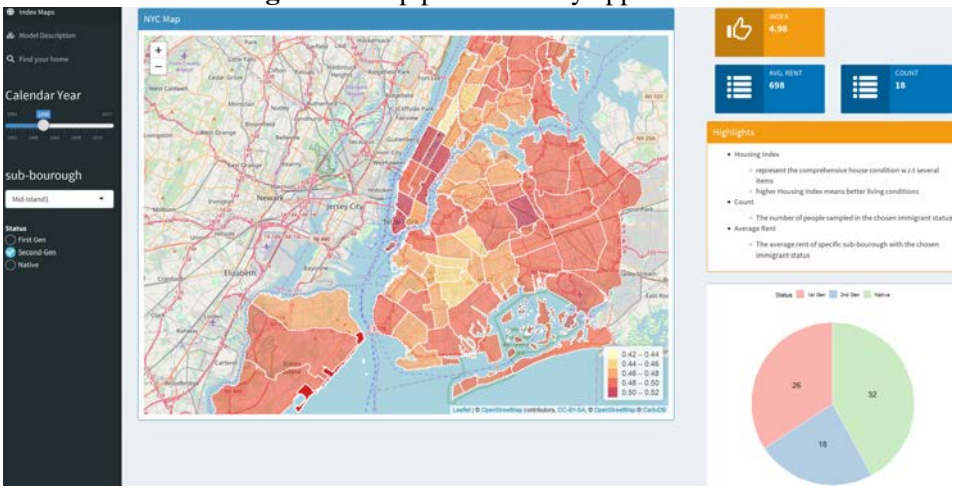


Note that we did not fit the linear regression model directly, since the house rent is changing over year and sub-borough. If we do that, we will lose the spatial and temporal information. This kind of two-stage model fitting method takes advantage of both the spatio-temporal information and other necessary covariates, which allows us to employ as much information as possible and make good prediction.

5. Interactive Visualization and Design of Data Product

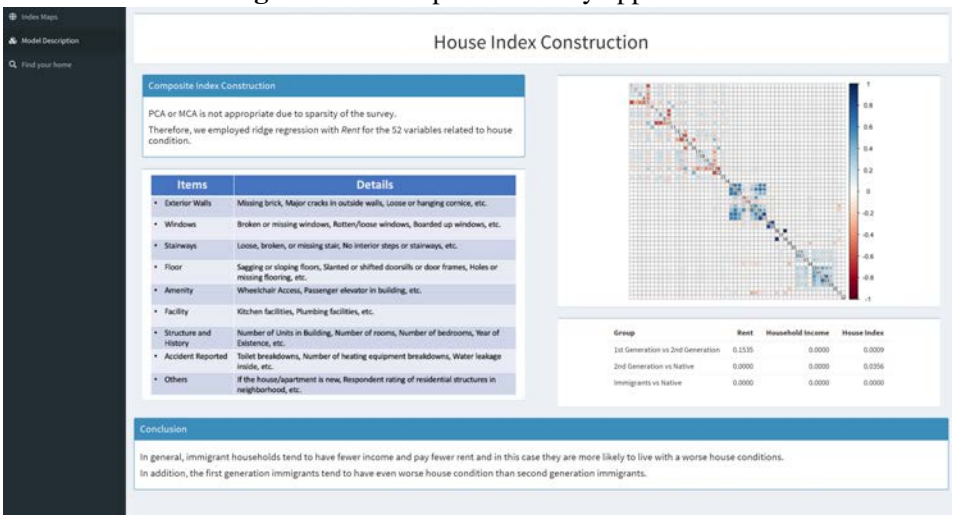
In order to integrate our results in this competition, we developed a R-shiny app on the Internet. It consists of three panels and each of them has specific objective.

Figure 6: Map panel of shiny application



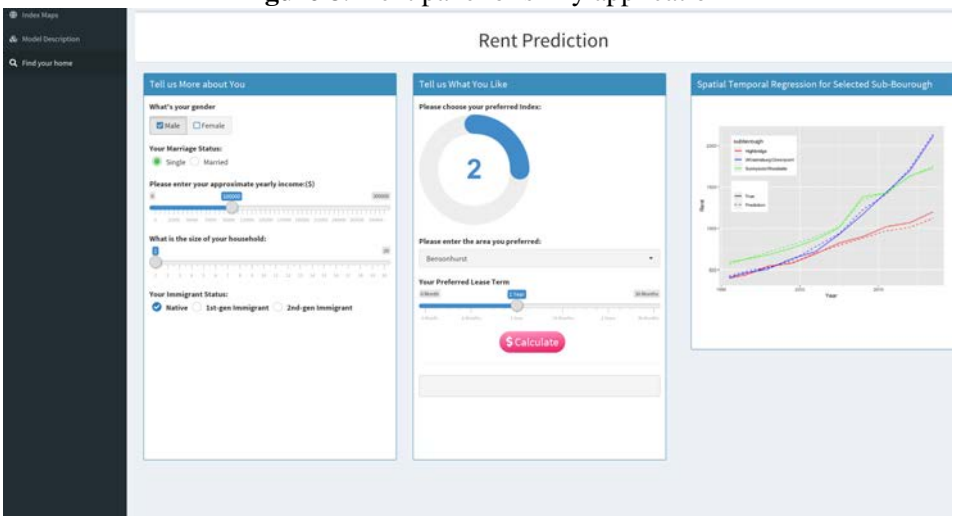
In the map panel(Figure 6), we provide a powerful interactive map of the New York City. When users pointed at some sub-borough, the corresponding area would be highlighted and display its name and house index which makes user to overview the information of New York City intuitively. Besides it, we also present some tuning parameters like calendar year, sub-borough and immigrant status for advanced use. Results are displayed in the information boxes on right hand side. A pie chart also created in order to exhibit the composition of immigrant status.

Figure 7: Model panel of shiny application



In the model panel(Figure 7), we demonstrate the data proceeding procedure and list some housing conditions and their descriptions. For the purpose of better understanding these decisions, we create co-variance matrix for these variables. A table of the p-values are presented to emphasize our conclusion on the impact of different immigrant status on the housing index.

Figure 8: Rent panel of shiny application



In the rent panel(Figure 8), we develop an application program interface(API) to help renters approximate their rents in 2020 based on their conditions and expected house index. The conditions include gender, age, household size, yearly income, etc. Although such information may be sensitive to some users, we do not collect any of these information for future use since our model is optimized by the NYCHVS dataset only. Therefore, the privacy and confidentiality of the users are secured. On the right side, we also present a plot which indicates how spatial information influence the rent. It tells us that the spatial information do have a big impact on the rent and can not be ignored in modeling.

6. Discussion

In this study, we proposed a new method in building composite indicator when the correlation of the covariates are not too strong. The construction of house quality index also works as an important part to the next two step rent prediction model. In the final, the study build an interactive visualization and data product to help new immigrant with decision making in housing.

While presenting in the data exposition, we focused more on the visualization and interactive product display with a shiny application. However, there are still several extensions and validations worth further studies. The study did not validate the method of index of construction in a statistical view. The new construction method has certain advantages. It overcomes the sparsity of the data which leads to the nullity of traditional dimension reduction method. In addition, the method also assigns non zero weight to each category of house deficiency. However, like all other response related dimension reduction methods, the index can be related to the rent directly which fail to eliminate the effect of time and location.

The two step spatial temporal model may capture the correlation between locations as well as the trend over time. In addition, the model gives accurate estimate of the rent in different areas. However, the model failed to provide a more flexible framework to accommodate using the data directly without centralizing the spatial temporal effect.

The data product as an interactive application for Researchers and new immigrants who are interested in moving to New York City provides recommendations and information in different perspectives. The shiny app not only come up with the historical house quality and immigrants preference, but also calculates the personalized estimation of expected rent

in the future to help users with better financial assignment.

REFERENCES

Aebischer, N. J., Robertson, P. A., and Kenward, R. E. (1993), "Compositional Analysis of Habitat Use From Animal Radio-Tracking Data," *Ecology*, 74, 1313–1325.

Asselin, L. M. (2002). Composite indicator of multidimensional poverty. Multidimensional Poverty Theory.

Bernstein, K. T., Galea, S., Ahern, J., Tracy, M., & Vlahov, D. (2007). The built environment and alcohol consumption in urban neighborhoods. Drug and alcohol dependence, 91(2-3), 244-252.

Freeman, L., & Braconi, F. (2004). Gentrification and displacement New York City in the 1990s. Journal of the American Planning Association, 70(1), 39-52.

Gamerman, D. (2010). Dynamic spatial models including spatial time series. Handbook of Spatial Statistics, 437-448.

Gitelman, V., Doveh, E., & Hakkert, S. (2010). Designing a composite indicator for road safety. Safety science, 48(9), 1212-1224.

Glaeser, E. L., Gyourko, J., & Saks, R. (2005). Why is Manhattan so expensive? Regulation and the rise in housing prices. The Journal of Law and Economics, 48(2), 331-369.

Gneiting, T., & Guttorp, P. (2010). Continuous parameter spatio-temporal processes. Handbook of Spatial Statistics, 97, 427-436.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

Lindstrm, J., Szpiro, A. A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., & Sheppard, L. (2014). A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. Environmental and ecological statistics, 21(3), 411-433.

Nardo, M., Saisana, M., Saltelli, A., & Tarantola, S. (2005). Tools for composite indicators building. European Comission, Ispra, 15, 19-20.

Porter, M. E., & Stern, S. (2001). National innovative capacity. The global competitiveness report, 2002, 102-118.

Sampson, P. D., Szpiro, A. A., Sheppard, L., Lindstrm, J., & Kaufman, J. D. (2011). Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. Atmospheric Environment, 45(36), 6593-6606.

Szpiro, A. A., Sampson, P. D., Sheppard, L., Lumley, T., Adar, S. D., & Kaufman, J. D. (2010). Predicting intraurban variation in air pollution concentrations with complex spatiotemporal dependencies. Environmetrics, 21(6), 606-631.