Interactive Visualization of Housing Condition in NYC

Qi Qi

Background

Motivated by ASA Data Challenge Expo 2019, I proposed to address one research question that is to describe changes in housing conditions for the first and the second-generation immigrant householders in New York City.

The data set is from the New York City Housing and Vacancy Survey (NYCHVS). The NYCHVS is a representative survey of the New York City housing stock and population. The HVS is a triennial survey with data collected about every three years. Each decade, a representative sample of housing units is selected, which represents the core sample. Field representatives collect information about each sampled unit, including those that are vacant as well as those that are occupied. For occupied units, an interview is conducted that gathers information about the reference person, any each additional member of the household, the household overall, and the household unit and building. In each survey cycle, the HVS gathers information about the core sample of housing units as well as an updated set of additional units that are sampled for each cycle to ensure that a given year's data are representative of the citywide housing stock.

Linked interviews within a decade are available for the 1990s and 2000s. The current decade of data (2011, 2014, and 2017) are not able to be linked into a longitudinal file due to disclosure avoidance protections. Therefore, the variables in each year are different and the information about last year real estate tax is not available in 2017. In order to summarize and represent the change across years, I extracted the common variables in every year.

Method

A R shiny app is constructed to visually show the changes in housing conditions. The map of NYC is represented in the app. Given the input of Borough (Bronx, Brooklyn, Manhatann, Queens or Staten Island) and subBorough, the corresponding region on the map is represented and zoomed in. The housing conditions include external condition of buildings (such as external wall, windows), internal condition of the building (such as elevators, floors), condition of room facilities (such as heating, kitchen facilities). Given either one of the housing condition variable as input, the changes will be shown.

I summarized the change by summary statistics of each variable for the specified region and year. The trend plot of such input variable over all avaliable years is shown in the app as well. The birth place of household and birth place of parents of household are used to classify the first the generation and the second generation immigrant householders and the classification is based on following definition: the first generation imigrant householder is whose birth place is not US; the second imigrant householder is who was born in US but the birth places of parents are both outside of US. Plots and analysis results are presented in the shiny app for comparison.

Data Analysis

In the shiny app, the plots and analysis results from user specified borough and sub-borough are able to be presented. In this report, I only show the plots and analysis results for whole NYC as illustration.

Figure 1 contains the plots representing the proportion of condition of exterior walls for first generation and second generation. The condition includes Missing material OR sloping/bulging outside walls, Major cracks

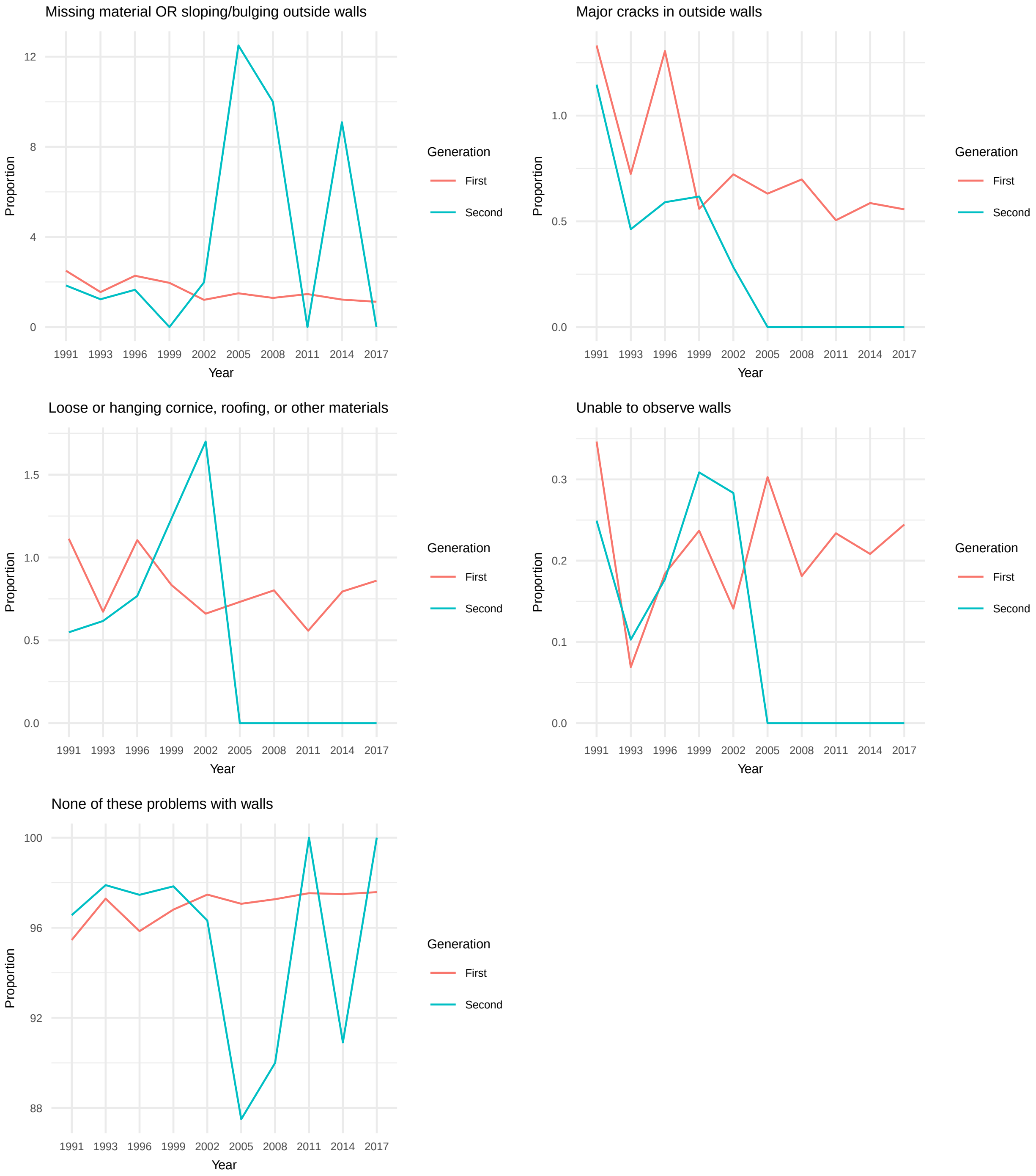

Figure 1: Proportion Plot of Condition of Exterior Walls

in outside walls, Loose or hanging cornice, roofing, or other materials, Unable to observe walls and None of these problems with walls.

According to the plots, we can notice that the trend of condition of walls for first generation is more flat than that for second generation. When we concentrate on the last plot, it can be noticed that the condition of walls for the first generation is improving steadily acorss years. The condition of walls for the second generation, however, has a dramatic change from 2012 to 2017.

In order to test whether significant difference exsits between first and second generation, results from chi-squared test are provided in the shiny app. Table 2, an example for condition of walls, shows p-values obtained from chi-squared test.

Table 1: p-values Obtained from Chi-squared Test

	1991	1993	1996	1999	2002	2005	2008	2011	2014	2017
Missing material OR sloping/bulging	0.114	0.367	0.142	0.019	0.290	0.27	0.302	1	0.317	1
Major cracks in outside walls	0.607	0.283	0.021	1.000	0.518	1.00	1.000	1	1.000	1
Loose or hanging cornice, roofing	0.037	0.919	0.282	0.640	0.046	1.00	1.000	1	1.000	1
Unable to observe walls	0.668	1.000	1.000	1.000	1.000	1.00	1.000	1	1.000	1
None of these problems	0.042	0.171	0.003	0.376	0.237	0.58	0.661	1	0.666	1

From the plot of non-problem on walls (last one in Figure 1), we can notice in 2005, the proportion for second generation is very different from the proportion for first generation. From Table 2, the chi-squared test shows no significant association between non-problem on walls and generation in 2005. Moreover, in 1991, the plot shows very close proportion of non-problem for first and second generation, which conflict with significant result shown in Table 2. These interesting contradictions may be caused by the huge difference between sample size for first generation (100,742) and second generation (6,368).

To investigate which factors alter the housing condition, I provide results from logistic regression on housing condition in the app. Table 3 and Table 4 are the examples from logistic regression on non-problem exsiting on exterior walls. There are some other variables included in the regression but not shown here.

Table 2: Logistic Regression on Condition of Exterior Walls for First Generation Imigrant Housholder

term	estimate	std.error	statistic	p.value
(Intercept)	3.457	0.132	26.282	0.0000
Householder's SexFemale	-0.006	0.044	-0.146	0.8842
Householder's Age Recode	0.002	0.001	1.472	0.1410
Total Household Income Recode	0.000	0.000	-0.202	0.8398
First Occupants of UnitNo	-0.284	0.094	-3.031	0.0024
Monthly cost (electric)	0.000	0.000	-5.786	0.0000
Monthly cost (gas)	0.000	0.000	6.658	0.0000
Tenure 1Rent	-0.461	0.057	-8.044	0.0000
Presence of mice or ratsNo	0.784	0.046	16.949	0.0000

Table 3: Logistic Regression on Condition of Exterior Walls for Second Generation Imigrant Householder

term	estimate	std.error	statistic	p.value
(Intercept)	3.906	1.113	3.510	0.0004
Householder's SexFemale	0.029	0.235	0.122	0.9029

term	estimate	std.error	statistic	p.value
Householder's Age Recode	0.013	0.006	2.142	0.0322
Total Household Income Recode	0.000	0.000	2.376	0.0175
First Occupants of UnitNo	-2.081	1.014	-2.052	0.0402
Monthly cost (electric)	0.000	0.001	-0.682	0.4949
Monthly cost (gas)	0.000	0.001	0.478	0.6329
Tenure 1Rent	-0.206	0.289	-0.712	0.4765
Presence of mice or ratsNo	0.856	0.250	3.427	0.0006

Based on these two tables, we can find that for both first and second generation, first occupant of unit and no presence of mice or rate provide larger odds while gender of householder does not have significantly different influence on wall condition. For first generation, higher monthly electric cost, higher monthly gas cost and owner of the house (compared with renter) provide higher odds. For second generation, older householder and higher total household income can provide higher odds. The factors affecting wall condition are different for first and second generation imigrant householder.

I considered principle component analysis to generate an index to represent the housing condition and provided further comparison of first and second generation. With all variables related to housing condition, I performed principle component analysis and generated the index for housing condition by the first principle component. With such index, I provided the p-values obtained from t test in Table 4 to represent the difference between first and second generation.

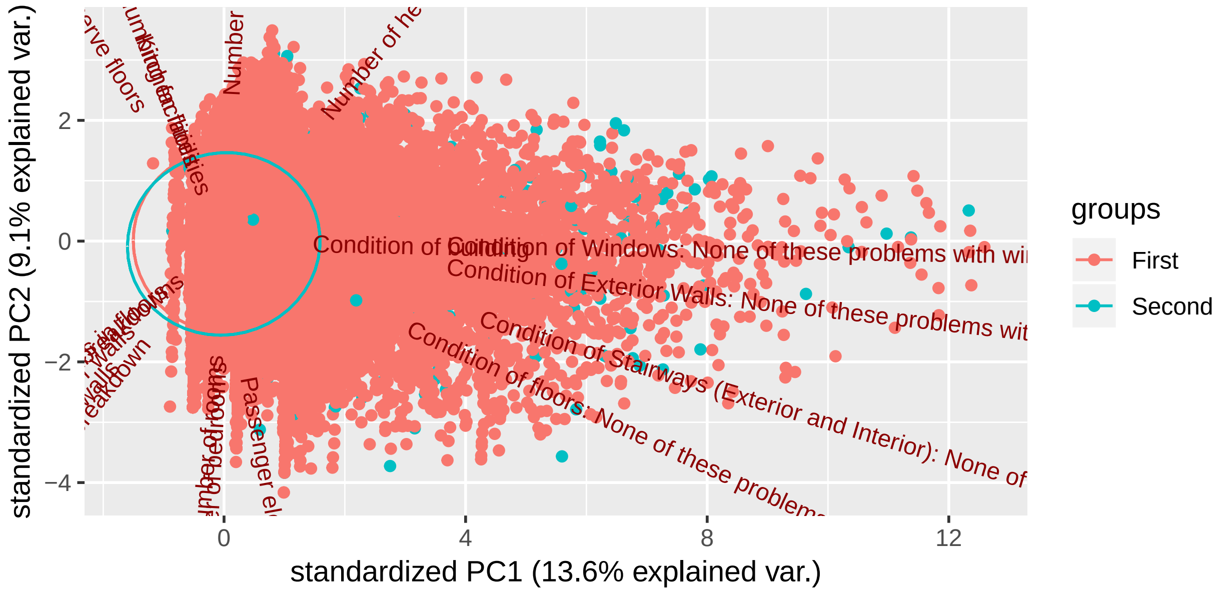

Table 4: p-values Obtained from t Test

	Any	1991	1993	1996	1999	2002	2005	2008	2011	2014	2017
p-value	0	2e-04	0	0.0474	0.7616	0.1411	0.2815	0.0269	0.2977	0.5339	0

The first generation and second generation form two very close distinct clusters to the right. However, there is significant difference between the index of housing condition for first generation and index for second generation not regrading years. The difference is not significant in 1999, 2002, 2005, 2011 and 2014.

Discussion

One issue is that the first principle component only explains 13.65% of variance. I may consider using more principle components to represent housing condition but I may need at least 10 components so as to explain more than 60% variance, which makes principle component analysis useless rather than using original variables. Another issue is the interpretational difficulty of principle component. It is hard to show the housing condition is good or bad by larger or lower index. I will explore other method to construct index for housing condition. There is missing data issue as well (around 12% missing rate). I only performed analysis on complete data. Since most variables are dummy variables and categorical variables, I considered imputation by mode but I will explore whether there is any method making more sense.

Moreover, so far I analyze the difference of first and second generation at each year separately. I am working on time series analysis and functional data analysis to achieve showing the difference of change in housing condition accross years for two generations.