8-2021

Housing Variables and Immigration: An Exploratory and Predictive Data Analysis in New York City

Jhonatan Medri Cobos
Utah State University

HOUSING VARIABLES AND IMMIGRATION: AN EXPLORATORY AND PREDICTIVE DATA ANALYSIS IN NEW YORK CITY

by

Jhonatan Medri Cobos

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

Approved:

Jürgen Symanzik, Ph.D. Major Professor	Daniel Coster, Ph.D. Committee Member
Lucy Delgadillo, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Interim Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2021

ABSTRACT

Housing Variables and Immigration: An Exploratory and Predictive Data Analysis in New York City

by

Jhonatan Medri Cobos, Master of Science

Utah State University, 2021

Major Professor: Jürgen Symanzik, Ph.D.
Department: Mathematics and Statistics

The relationship between housing and immigration has become relevant in the U.S., especially in a highly populated metropolis such as New York City (NYC). Determining whether immigration status is highly associated with housing variables such as home ownership percentage, home renting, or housing cost percentage could help understand the quality of life of NYC residents. Choropleth maps, box plots overlaid with dot plots, smoothed scatterplots, and linked micromap plots provide an exploratory data analysis across all 55 NYC sub-boroughs. Those graphical tools, besides of showing a high variability in the data, also suggest some spatial patterns in the housing data. This is confirmed by spatial autocorrelation tests that use both the Moran's I and Geary's C statistics. When using spatial autoregressive models to try to explain housing variables, we notice sociodemographic variables of the householder and other housing variables are significant; being an immigrant household is associated with a lower home ownership percentage, but a higher housing cost percentage. Immigration status didn't show a clear impact on household rent.

(106 pages)

PUBLIC ABSTRACT

Housing Variables and Immigration: An Exploratory and Predictive Data Analysis in New York City

Jhonatan Medri Cobos

The relationship between housing and immigration has become relevant in the U.S., especially in a highly populated metropolis such as New York City (NYC). Determining whether immigration status affects home ownership percentage, household rent, or housing cost percentage could help understand the quality of life of NYC residents. Graphical exploration, spatial dependence tests, and spatial autoregressive models of housing and immigration variables provide some insights about their relationships. Our exploration takes place at some geographic subareas of NYC.

Our results first indicate that the housing and immigration data reports spatial dependence; values of a geographic subarea are related to values of other nearby subareas. In addition, we notice that an immigrant householder is less likely to own a home and more likely to pay a higher rent as a proportion of their income. This result is more apparent in the Bronx and Manhattan. However, being an immigrant can't be associated with higher rent amounts since household rent depends more on other factors, such as the household income.

This thesis is dedicated to my parents, Anita and Enrique.

ACKNOWLEDGMENTS

I want to thank my advisor Dr. Jürgen Symanzik for giving me an opportunity in the field. His constant mentorship, appreciated patience, good humor, and passion for computational statistics motivated me to get more involved in the field. With a special mention to Dr. Lucy Delgadillo and Dr. Daniel Coster for the time they put in serving on my thesis committee and whose meaningful insights made this research possible.

I want to thank my brother Kike and my sister Cristy for encouraging me to be strong. I also want to thanks Utah State University's Department of Mathematics and Statistics, with a special mention to the Graduate Program Coordinator Gary Tanner, for not only motivating me to get such a challenging degree, but providing me with all the tools to get my Master's Degree.

I would finally like to thank the Sections on Statistical Computing, Statistical Graphics, and Government Statistics of the American Statistical Association (ASA) for providing the data used in these analyses. Data manipulations and visualizations made use of the R packages "rgdal" (Bivand et al., 2019), "rgeos" (Bivand and Rundel, 2014), "geojsonio" (Chamberlain and Teucher, 2019), "rmapshaper" (Teucher and Russell, 2018), "sp" (Bivand et al., 2013), "dplyr" (Wickham et al., 2019), "ggplot2" (Wickham, 2016), "micromap" (Payton and Olsen, 2015), "shiny" (Chang et al., 2021), "grid" (R Core Team, 2019), and "LMShapemaker" (Probst, 2020).

JM

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

Introduction

Immigration and housing are two popular topics broadly considered when discussing public policies. For example, owning a home is considered a symbol of welfare accumulation (Doling and Ronald, 2010). Unfortunately, sometimes having access to a place to live can involve some housing problems that could exacerbate gaps among low-income and minority groups (Herbert et al., 2005).

In the case of immigration, new difficulties could appear. The fact that immigrants are new to a place could imply they have less access to many resources. This could result in a lower household income, lower credit and mortgage access, higher household rent, or higher housing prices (McConnell and Akresh, 2010). Having an assesment of this housing and immigration relation is particularly important in the policy-making process which heavily bears on the integration of immigrants into the respective societies of receiving countries (Dell'Olio, 2004).

The 2019 Data Challenge Expo of the Sections on Statistical Computing, Statistical Graphics, and Government Statistics of the American Statistical Association (ASA) provided an opportunity to explore this relationship between housing and immigration variables in New York City (NYC). Our analyses focused on the following research question:

How are housing and immigration status variables related in NYC?

In addition, our analyses addressed the following specific questions in three housing variables:

- How are home ownership percentage and immigration status related in NYC?
- How are home renting and immigration status related in NYC?
- How are housing cost percentage and immigration status related in NYC?
- How do these relationships differ among NYC boroughs and sub-boroughs?

Previous work related to housing and immigration exists. Some authors have identified a positive relationship between immigration flows and housing prices, measured by rent or housing value

(Moos and Skaburskis, 2010; Mussa et al., 2017). In addition, immigration, among other sociodemographic attributes, could also influence home ownserhip rates, housing costs, and living conditions (DeSilva and Elmelech, 2012; McConnell and Akresh, 2010; Shier et al., 2016). Moreover, some studies have found a spatial correlation in this relationship (Liu et al., 2020; Pettit et al., 2017; Zou, 2014).

This thesis is organized as follows. In Chapter 2, we describe the data used and the variables selected in our analyses at the borough level. We then proceed in Chapter 3 to conduct an exploratory data analysis in home ownership percentage, home renting, and housing cost percentage using data visualization tools, all at the sub-borough level. Choropleth maps, box plots overlaid with dot plots, smoothed scatterplots, and linked micromap plots provide an exploratory data analysis across all 55 NYC sub-boroughs.

Then, in Chapter 4, we conduct spatial autocorrelation tests in the data and develop spatial autoregressive models in our three housing variables. All graphical tools and spatial autocorrelation tests have been included in an interactive Shiny R user application explained in Chapter 5. We finally provide some conclusions in Chapter 6.

Appendix A provides four-year comparison figures that complement the data visualization developed in Chapter 3. Appendix B depicts a graphical representation of the weighting methods used in Chapter 4. Appendix C gives more details about our tuning results in Chapter 4 and spatial autocorrelation results in variables not covered in this thesis. Finally, Appendix D summarizes the Variation Inflation Index (VIF) for the variables selected for our models.

This thesis is an extension of two previous articles that are based on the 2019 Data Challenge Expo of the Sections on Statistical Computing, Statistical Graphics, and Government Statistics of the American Statistical Association (ASA) (Medri et al., 2019, 2021). Although those articles contain similar data visualization and spatial statistics analysis, they also contain additional graphs not presented here. In addition, this thesis mainly analyses three housing variables in the 2017 data in Chapters 3 and Chapters 4. This data includes information from 13,266 NYC households. Suggestions to expand the scope of this research are covered in Chapter 6.

CHAPTER 2

Data and Variable Selection

The data set used for this thesis is the New York City Housing and Vacancy Survey 1991-2017 (NYCHVS), provided during the 2019 Data Expo Challenge (https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2019) organized by the American Statistical Association (ASA). The NYCHVS is a representative survey of the New York City housing stock and population, sponsored by the New York City Department of Housing Preservation and Development (HPD). This survey takes place every three years and collects information from both vacant and occupied housing units. The data set is available at the following URL:

https://www.census.gov/programs-surveys/nychvs.html

In this thesis, we mainly work with the data from 2017, but we also do some temporal assessments. The 2017 data includes observations from 13,266 NYC households.

2.1 Variable Selection

Table 2.1 indicates the variables considered in our analyses. We decided to focus our analyses on the variables from the survey relating to housing, immigration, and other sociodemographics.

From the housing variables perspective, we are including home ownership and renting variables. Home ownership variables, such as owner tenure and mortgage status in Table 2.1, provide information about how many tenants own a home and their mortgage rate, in case the home is not paid in full. The housing value is the respondent's estimate of how much the housing residential portion would sell for if it were for sale. In this thesis, when we talk about home ownership percentage we talk about the percent of households that own a home based on variable 01 of Table 2.1.

Table 2.1: List of housing and immigration variables included in our analyses. The sub-borough range indicates the minimum and maximum median value across all sub-boroughs and relates to the percentage of "1s" for variables 01, 04, 06, 09, 11, 12, 13.

Variable Name	Description	Sub-borough Range
Housing variables		
01. Owner Tenure	1 if tenant owns home	[4.3% - 73.6%]
02. Household Rent	Annual amount in current US$	[$10,800 - $37,800]
03. Housing Cost Percentage	Rent as a proportion of income	[25.1% - 47.9%]
Householder sociodemographic		
04. Householder's Birthplace	1 if outside the U.S.	[17.1% - 79.3%]
05. Householder's Age	Years	[40 - 63]
06. Householder's Sex	1 if female	[37.6% - 67.3%]
07. Household Income	Annual amount in current US$	[$20,636 - $145,000]
Housing features and value		
08. Housing Value	Amount in current thousands US$	[$110 - $1,925]
09. Mortgage Status	1 if home has a mortgage	[27.3% - 87.5%]
10. Mortgage Interest Rate	Percentage	[3.2% - 5.0%]
11. Walls Condition	1 if walls are in good condition	[94.4% - 100%]
12. Stairs Condition	1 if stairs are in good condition	[85.5% - 98.5%]
13. Floors Condition	1 if floors are in good condition	[89.9% - 99.8%]
14. Number of Rooms	Rooms in the house	[3, 6]

Variables 11, 12, 13, and 14 will primarily be used in the spatial autoregressive models (SAR) in Chapter 4.2

In addition, home renting variables, such as household rent in Table 2.1, indicate how much money people pay in rent, including utilities. In this thesis, when we refer to household rent, we consider its gross value in current US$.

In the case of housing cost, The United States Department of Housing and Urban Development (HUD) uses an index that calculates rent as a simple percentage of income in relation to housing costs, such as rent. Then, they categorize that ratio into housing affordability if the index is below 30 percent, housing cost burdent if it is between 30 and 50 percent, and severe housing cost burden if it is above 50 percent. Families that face severe housing cost may find difficulties affording basic necessities such as food, clothing, transportation, and medical care (Dacquisto and Rodda, 2006).

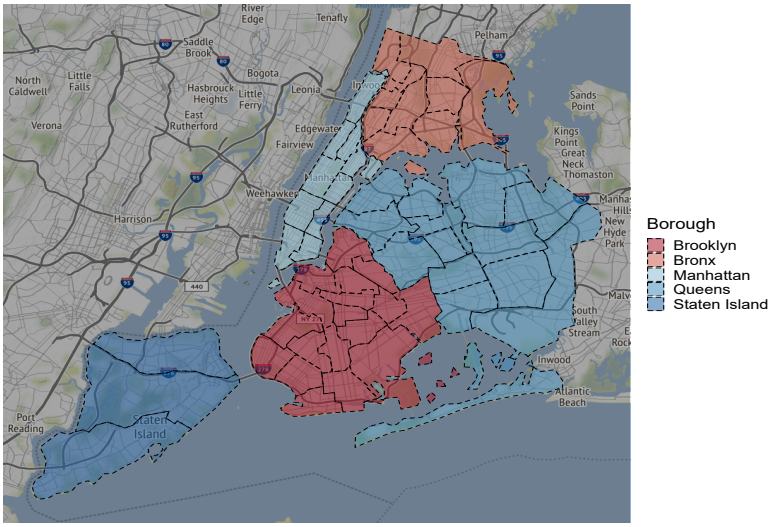

Fig. 2.1: NYC boroughs and sub-boroughs: There are 18 sub-boroughs in Brooklyn, 10 in the Bronx, 10 in Manhattan, 14 in Queens, and 3 in Staten Island

Nevertheless, this index fails to take into consideration other factors that influence housing costs, such as a cost of living variable (O'Dell et al., 2004) or actual financial constraints (Bogdon and Can, 1997). We chose to work with this index because of its ease of calculation and in a merely descriptive role. In this thesis, we define this index as the housing cost percentage.

From the immigration perspective, we are mainly considering the primary householder's origin. We first look at the place of the householder's birth country and classify this group as immigrants or US citizens. We also include other sociodemographic variables such as the householder's age, sex, and total household gross income in Table 2.1. Finally, we take into account the reported condition of the walls, stairs, and floors of the housing.

We run our analyses on the 55 sub-boroughs within the five main boroughs in NYC: Brooklyn (BK), the Bronx (BX), Manhattan (MN), Queens (QN), and Staten Island (SI). Boroughs are county-level administrative divisions of NYC, while sub-boroughs are groups of census tracts summing to at least 100,000 residents. Figure 2.1 shows the sub-boroughs for each NYC borough and more details about them can be found in https://www.census.gov/programs-surveys/nychvs/geographies/reference-maps.html.

Table 2.2: Home ownership descriptive statistics for 2017

Borough	% Owners	% Debtors	Median Mortgage Interest Rate	Median Housing Value
Brooklyn	27.7%	65.1%	4.2%	$758,500
Bronx	20.8%	58.1%	4.3%	$312,500
Manhattan	20.2%	56.2%	3.7%	$847,250
Queens	44.2%	61.7%	4.1%	$546,250
Staten Island	63.9%	68.1%	4.0%	$473,500
Total	31.3%	61.5%	4.1%	$623,900

2.2 Descriptive Statistics

We can make some general comments for each borough according to the descriptive statistics displayed in Tables 2.2, 2.3, and 2.4. New York City residents, compared to the rest of U.S. cities, tend to rent a place rather than own one. In addition, they report a higher household income and rent, but also a higher housing cost (U.S. Census Bureau, 2019a). Moreover, NYC has a relatively high immigration and first generation concentration in the United States (U.S. Census Bureau, 2019b).

As shown in Table 2.2, less than a third of the population owns a home in New York City. Staten Island is the borough with most owners (64%) and the Bronx is the one with least (21%). About 62% of those who own a home have a mortgage loan at an median interest rate of 4%. There are no major differences with respect to the median interest rates at the five boroughs. In contrast, the median housing value is different across all five boroughs. Manhattan hast the highest median housing value ($847,250) and the Bronx has the lowest ($312,500).

In contrast, as shown in Table 2.3, about 67% of New Yorkers tend to rent a place. Manhattan is the borough with the highest percentage (78%) and Staten Island with the lowest (33%). Similarly, the annual median household gross rent and income across all sub-boroughs are $18,100 and $61,500, respectively. Manhattan is the borough with the highest median household annual income and rent, $80,400 and $22,700, respectively, and the Bronx reported the lowest household median income and rent, $37,000 and $14,800, respectively. Nevertheless, the Bronx has the highest median housing cost percentage with about 40% and Staten Island has the lowest with about 20%.

The numbers in Table 2.4 provide more information about the sociodemographics of NYC. Table 2.4 shows that more than 40% of the population come from a foreign country, supporting the

Table 2.3: Home renting descriptive statistics for 2017

Borough	% Renters	Median Rent ($)*	Median Income ($)*	Median Housing Cost (%)
Brooklyn	70.0%	$17,600	$59,500	29.6%
Bronx	76.5%	$14,800	$37,000	39.9%
Manhattan	78.4%	$22,700	$80,400	28.3%
Queens	53.8%	$18,600	$64,500	28.9%
Staten Island	33.0%	$15,200	$77,300	19.6%
Total	66.6%	$18,100	$61,500	29.5%

* Annual value in current US$.

Table 2.4: Householder sociodemographic descriptive statistics for 2017

Borough	% Immigrants	% Female	Median Age (years)
Brooklyn	38.2%	56.1%	49
Bronx	36.0%	60.4%	51
Manhattan	22.7%	55.5%	50
Queens	50.8%	47.4%	51
Staten Island	24.5%	49.6%	54
Total	43.0%	54.2%	50

claim that NYC has a relatively high immigrant concentration. In addition, the householder sex is evenly split among most boroughs, with a slightly higher concentration of females. Finally, the householder median age is around 50 years in most boroughs, except Staten Island where it reaches a slightly higher value.

In the following chapter, Chapter 3, we will conduct an exploratory data analysis to identify some patterns using data visualization tools. In Chapter 4, we will use spatial statistics methods for our data to both conduct spatial autocorrelation tests and build spatial autoregressive models.

CHAPTER 3

Exploratory Data Analysis at the Sub-Borough Level

We will use graphical and statistical tools to explore in more detail any relationships between housing and immigration variables at the sub-borough level. The graphical tools we will use are choropleth maps, box plots overlaid with dot plots, smoothed scatterplots, and linked micromap (LM) plots. Choropleth maps use color for shading the NYC map to represent sub-borough values of a certain variable. Unfortunately, choropleth maps have some limitations.

Symanzik and Carr (2008, pp. 270-271) identified three main problems with choropleth maps (). The first problem relates to the region area because some map regions can be too small to effectively show color. A second key problem is that converting a continuous variable into a variable with a few ordered values can result in an immediate loss of information. The last key problem is that it is difficult to show more than one variable in a choropleth map. That is why, we are using LM plots to address these issues with choropleth maps.

In general, in LM plots, a column of small maps is linked via color to one or more columns with statistical information (Carr and Pierson, 1996; Gebreab et al., 2008; Symanzik and Carr, 2008). In our case, the maps show the sub-boroughs of NYC. The rows in LM plots are arranged according to an increasing (or decreasing) order of one of the statistical variables. Some correlation coefficients are calculated to provide supporting quantitative information. These correlation coefficients will be calculated based on the raw (all 13,266 observations from the survey) and aggregated data (median or percentage values in all 55 sub-boroughs). In the case of the correlations from the aggregated data, we are aware they represent an ecological correlation (Piantadosi et al., 1988), but they are still useful to explain some relations that can be seen in our LM plots. An ecological correlation is a correlation between two variables that represent group means, in this case sub-borough means. This type of correlation is, usually, artificially stronger compared to a correlation between two variables that describe individuals.

In the case of the scatterplots, we will use a LOESS (locally estimated scatterplot smoothing)

smoother (Cleveland et al., 1992), a non-parametric approach that fits multiple regressions in a local neighborhood to notice a possible association between two variables.

Sections 3.1, 3.2, and 3.3 will make an exploratory data analysis in the housing variables mentioned in Table 2.1: home ownership percentage, home renting, and housing cost percentage. We will use the statistical and graphical tools described above.

3.1 Home Ownership Exploration

Figure 3.1 shows the home ownership sub-borough percentages for NYC for 2017. Most of the Staten Island and eastern Queens sub-boroughs have the highest home ownership percentages (above 43.4%), which hasn't considerably changed in the last 26 years (see Figure A.1 in Appendix A for further details). Regarding the mortgage interest rates, Figure 3.2 suggests that most sub-boroughs have similar rates and spreads. The Manhattan borough has the lowest median interest rate, but also the highest household income. In previous literature, labor income has shown some

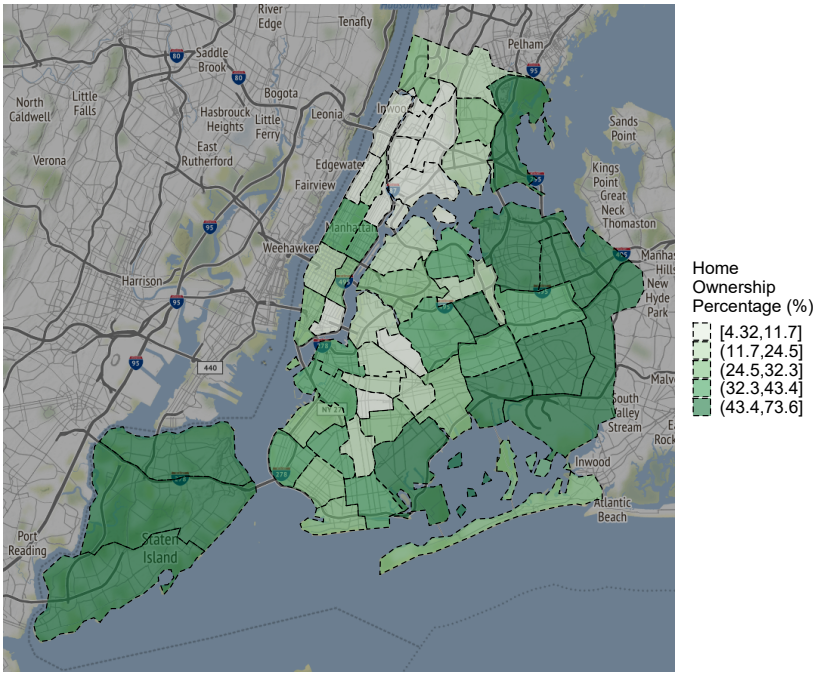

Fig. 3.1: NYC home ownership percentage choropleth map for 2017. The data are broken into quintiles, with about 11 sub-boroughs in each of the quintiles

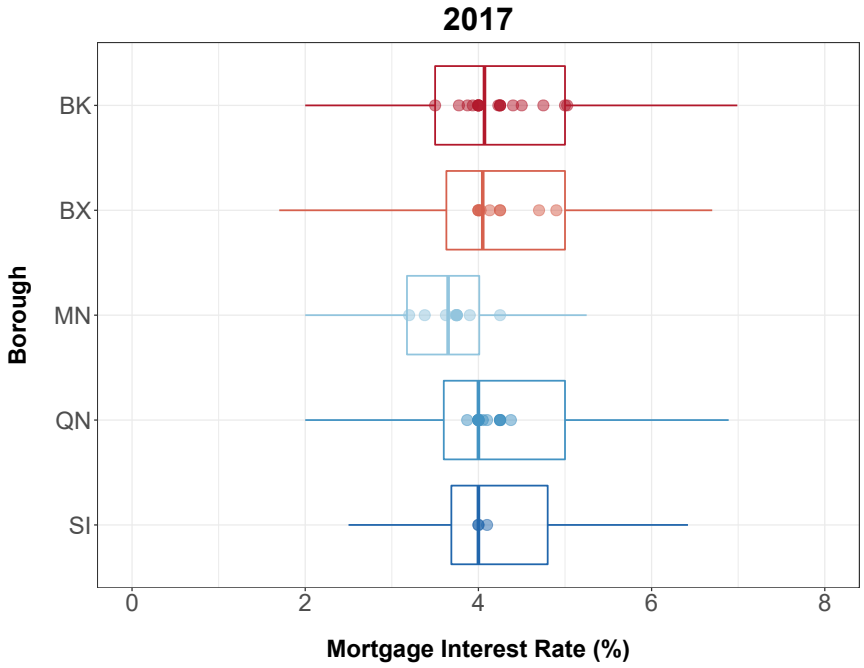

Fig. 3.2: NYC mortgage interest rates by borough for 2017. The dots represent the aggregated data of the 55 sub-boroughs, and the boxplot is constructed based on the raw survey data (1,685 observations)

correlation with the level of interest rates, but this correlation also varies according to the income risk, among other variables (Campbell and Cocco, 2015). When we look at the last 12 years, we notice an overall reduction in both the mortgage rates and spread across all NYC boroughs (see Figure A.6 in Appendix A for further details).

Figure 3.3 shows that the immigration status has no major effect on mortgage interest rates. Both splines for US citizens and immigrants show a similar slightly decreasing LOESS smother. A similar pattern can be observed in previous years (see Figure A.9 in Appendix A for further details).

The LM plot in Figure 3.4 confirms observations from Table 2.2: the three sub-boroughs of Staten Island and five in Queens are among the ten sub-boroughs with the highest percentage of owned households. In contrast, four sub-boroughs in the Bronx have the lowest owned households percentage in all NYC, and more than 80% of the Bronx and Manhattan sub-boroughs are below the median. We don't notice in the LM plot any apparent correlation between owned households,

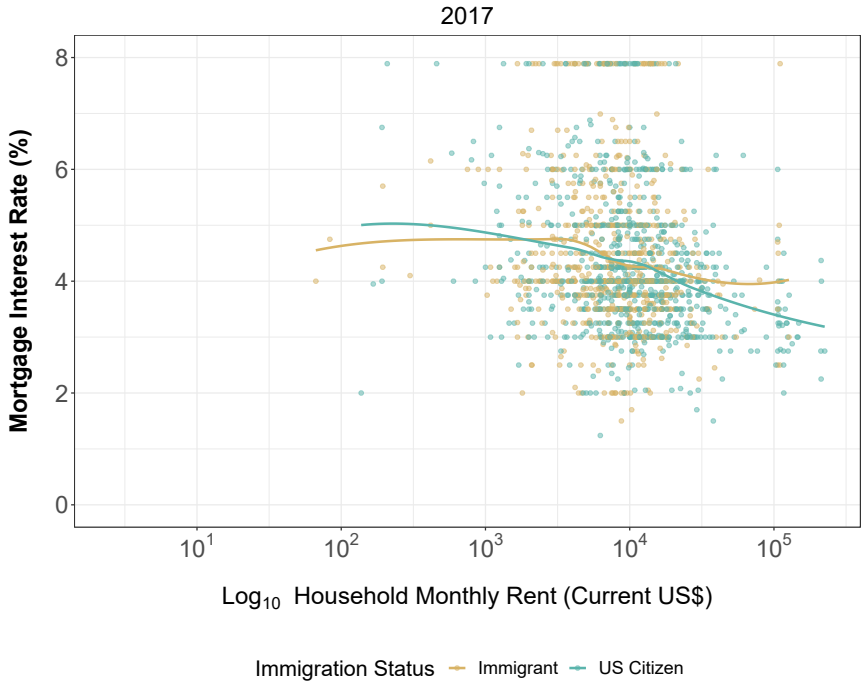

Fig. 3.3: Monthly income and mortgage interest rates scatterplot for 2017. We don't notice any significant difference between the immigrant and the US citizen group

households with a mortgage, and immigrant households, so we will look at more specific correlations.

Table 3.1 shows the correlation coefficients among owner tenure, mortgage status, and immigration status for the raw data (3,926 observations in 55 sub-boroughs) and the aggregated data (based on the aggregation of each sub-borough). Most of these correlation coefficients relate to very weak linear association or no associations at all. Immigration status has an almost zero correlation with owner tenure ($r = -0.02$ in the raw data and $r = 0.02$ in the aggregated data) and with mortgage status ($r = 0.10$ in the raw data and $r = 0.04$ in the aggregated data). Since householders that have a mortgage already have owner tenure, this correlation can't be calculated in the raw data.

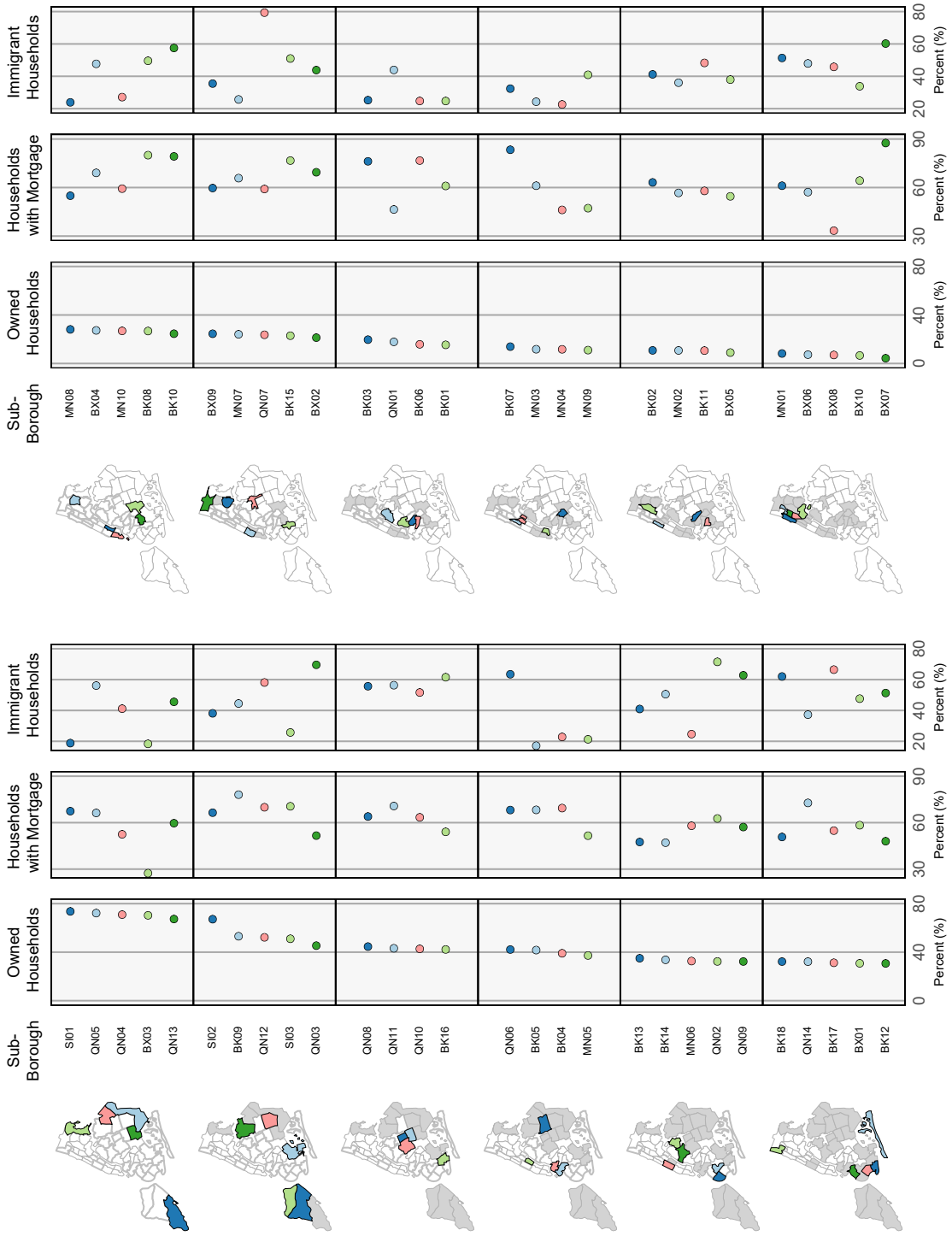

Fig. 3.4: NYC home ownership percentage LM plot for 2017. The 55 sub-boroughs are ranked in descending order by owned households percentage. The LM plot is meant to show the relationship between the primary panel, the percentage of owned households, with the secondary panels, the percentage of households with a mortgage and the percentage of immigrant households

Table 3.1: NYC home ownership percentage correlations for 2017

	Raw Data			Aggregated Data		
Borough	Owner Tenure	Mortgage Status	Immigrant Households	Owner Tenure	Mortgage Status	Immigrant Households
Brooklyn - 1,114 observations in 18 sub-boroughs						
Owner Tenure	1.00	–	-0.01	1.00	-0.20	0.17
Mortgage Status	–	1.00	0.03	-0.20	1.00	-0.39
Immigrant Households	-0.01	0.03	1.00	0.17	-0.39	1.00
Bronx - 373 observations in 10 sub-boroughs						
Owner Tenure	1.00	–	-0.05	1.00	-0.52	-0.68
Mortgage Status	–	1.00	0.27	-0.52	1.00	0.68
Immigrant Households	-0.05	0.27	1.00	-0.68	0.68	1.00
Manhattan - 687 observations in 10 sub-boroughs						
Owner Tenure	1.00	–	-0.10	1.00	0.09	-0.64
Mortgage Status	–	1.00	0.04	0.09	1.00	0.08
Immigrant Households	-0.10	0.04	1.00	-0.64	0.08	1.00
Queens - 1,317 observations in 14 sub-boroughs						
Owner Tenure	1.00	–	-0.06	1.00	0.13	-0.31
Mortgage Status	–	1.00	0.14	0.13	1.00	-0.01
Immigrant Households	-0.06	0.14	1.00	-0.31	-0.01	1.00
Staten Island - 435 observations in 3 sub-boroughs						
Owner Tenure	1.00	–	0.03	1.00	-0.87	-0.12
Mortgage Status	–	1.00	0.03	-0.87	1.00	0.04
Immigrant Households	0.03	0.03	1.00	-0.12	0.04	1.00
Total - 3,926 observations in 55 sub-boroughs						
Owner Tenure	1.00	–	-0.02	1.00	-0.04	0.02
Mortgage Status	–	1.00	0.10	-0.04	1.00	0.04
Immigrant Households	-0.02	0.10	1.00	0.02	0.04	1.00

Owner Tenure and Mortgage Status are yes/no answers in the survey.

A ”$-$” represents pairs of variables where correlation can’t be calculated.

The aggregated data considers the percentage of ”yes” answers in each sub-borough.

In the aggregated data, where each variable relates to the percentage of ”yes” in a sub-borough, the correlation is almost zero ($r = -0.04$). However, when looking at each individual borough, we

start noticing some relations. The Bronx ($r = -0.52$) and Staten Island ($r = -0.87$) reported a high, negative correlation between percentage of owned households and percentage of households with a mortgage. A plausible explanation for this relation could be that these boroughs either don't rely on mortgages to acquire a home, or may have already paid off the mortgage. For example when we take a look at Staten Island, the borough with the highest home ownership percentage and median householder age, the correlation between householder age and mortgage status in the raw data is $r = -0.45$; this could indicate that as the householder gets older, he or she is more likely to have already paid off his mortgage.

In addition, the Bronx ($r = -0.68$) and Manhattan ($r = -0.64$) also reported a negative correlation between percentage of owned households and percentage of immigrant households. In other words, being an immigrant household in those areas is associated, on average, with lower home ownership percentages. Finally, the Bronx ($r = 0.68$) reported a negative correlation between percentage of households with a mortgage and percentage of immigrant households. Being an immigrant household can be associated with a higher access to mortgage in that area.

Other variables, such as household income, may be playing a more important role when determining access to a mortgage. When looking at the aggregated data, some of the ecological correlations are considerably stronger than the correlations for the raw data. These could be due to a high variability or a small sample size (with only three sub-boroughs in Staten Island). We leave it to the reader to further interpret these. It is worthwhile mentioning that the correlations for the aggregated data are a numerical summary of the visual impression provided by the LM plot in Figure 3.4.

The consideration of other explanatory variables could help to explain a possible relationship between home ownership and immigration. Factors not considered in our analyses, but in previous research, include intergenerational transfers (Mulder and Smits, 1999) and ethnic variation on investment preferences (Owusu, 1998). Additionally, when looking at the rest of the United States, immigrants observed (e.g., education) and unobserved attributes (e.g., ambition), in combination with other socioeconomic and spatial variables, could also influence home ownership (DeSilva and Elmelech, 2012).

3.2 Home Renting Exploration

The amount of people that do not own a home is around 70% in NYC. Figure 3.5 gives a general perspective of how much money NYC residents spend on rent. It indicates that the household median rent is relatively higher in middle and southern Manhattan compared to other places in NYC. In the last 26 years, these areas have also reported the highest median rent compared to the other sub-boroughs in NYC. Staten Island is an interesting case since it has reported a lower increase in rent compared to the other sub-boroughs (see Figures A.2 and A.3 in Appendix A for further details).

Figure 3.6 gives a general perspective of the rent spread within each borough. In the case of Manhattan, this borough has the largest sub-borough spread, a bimodal distribution, considering both the lowest- and highest-rent places are in northern and southern Manhattan, respectively. The Bronx borough, in contrast, reports both the smallest spread and the lowest median rent among the five boroughs. Even considering the increase in the Consumer Price Index (CPI) in the last 26 years, all boroughs have reported a similar pattern as in 2017 (see Figure A.7 in Appendix A for further

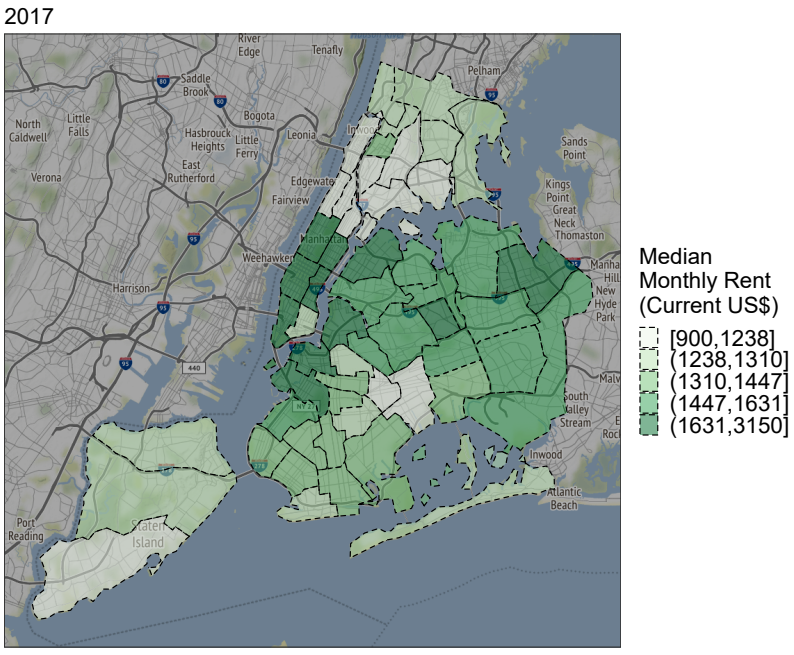

Fig. 3.5: NYC household median monthly rent choropleth map for 2017. The data are broken into quintiles, with about 11 sub-boroughs in each of the quintiles. Middle and southern Manhattan report the highest rent amounts and west Bronx the lowest

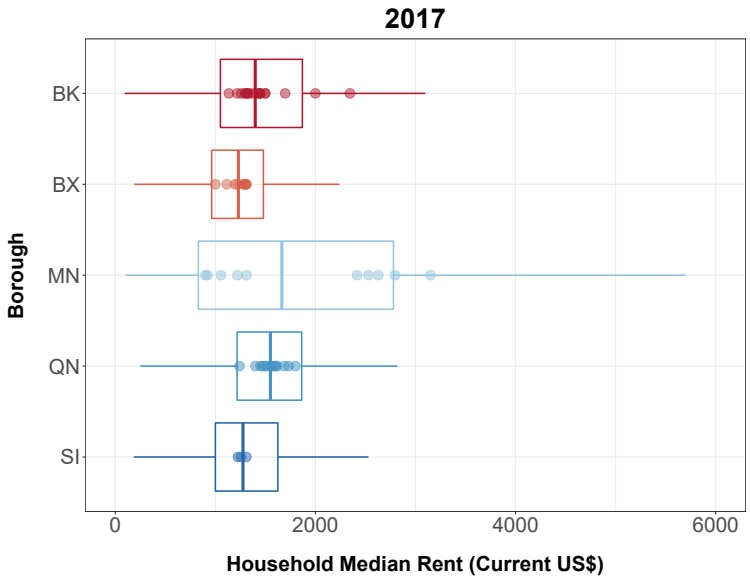

Fig. 3.6: NYC household median monthly rent by borough for 2017. The dots represent the aggregated data of the 55 sub-boroughs, and the boxplot is constructed based on the raw survey data (8,902 observations)

details).

Immigration status shows a weak or no noticeable relation with household rent or income. Figure 3.7 shows no major difference between the US citizen or the immigration group splines, a behavior that has become more apparent throughout the years (see Figure A.10 in Appendix A for further details).

The LM plot in Figure 3.8 confirms observations from Table 2.3: the five sub-boroughs with the highest monthly household rent are located in Manhattan, while eight of the ten sub-boroughs with the lowest are, also, located in Manhattan and Bronx. We also notice that immigration status is somewhat negatively associated with household income and rent, but the relationship with immigrant households is not that clear.

Table 3.2 shows the correlation coefficients among monthly rent, monthly income, and immigration status for the raw data (8,362 observations in 55 sub-boroughs) and the aggregated data (based on aggregations within each sub-borough). Immigration status has a very weak negative correlation with monthly rent ($r = -0.09$) and monthly income ($r = -0.08$) in the raw data. However, in the aggregated data, the ecological correlations are stronger ($r = -0.26$ and $r = -0.31$ respec-

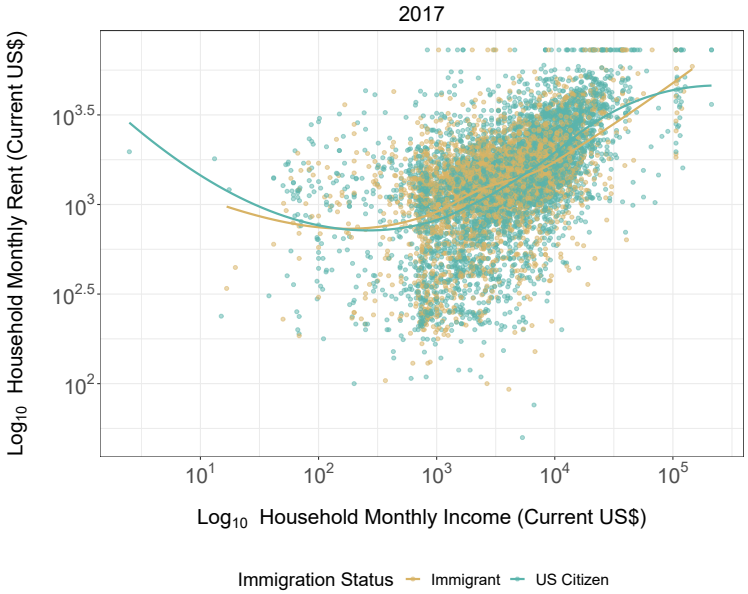

Fig. 3.7: NYC monthly income and rent scatterplot for 2017. The cases where monthly rent is greater than monthly income represent 12% of the data and can mostly be explained by households that live for free or receive external support. The 2017 NYCHVS survey had a maximum listed amount of $5,995 for monthly rent, and it assigned a value of $7,992 to households that listed a monthly rent above the listed maximum of $5,995 as $7,992 represented the mean of those manually reported higher monthly rents

tively). This could suggest the data has a high variability. When looking at each individual borough, the strongest correlation coefficients are located in Brooklyn ($r = -0.56$ and $r = -0.53$) and Manhattan ($r = -0.45$ and $r = -0.41$). Being an immigrant household in this areas is associated, in average, with a lower rent and income.

In contrast, the correlation between monthly income and monthly rent is stronger, both for the raw data and even more for the aggregated data. The LM plot in Figure 3.8 shows a strong positive association between median household income and median monthly rent ($r = 0.88$ for the aggregated data, shown in Table 3.2). This noticeable relationship makes sense when looking at the individual correlations in Table 3.2 for the raw data in the Brooklyn ($r = 0.49$), Manhattan ($r = 0.46$), Bronx ($r = 0.37$), Queens ($r = 0.37$), and Staten Island ($r = 0.30$) boroughs. However, the fact that the ecological correlations for the aggregated data are much stronger also suggests those variables have a high variability in the data. Overall, the correlations for the aggregated data in Table 3.2 confirm the visual impression provided by the LM plot in Figure 3.8.

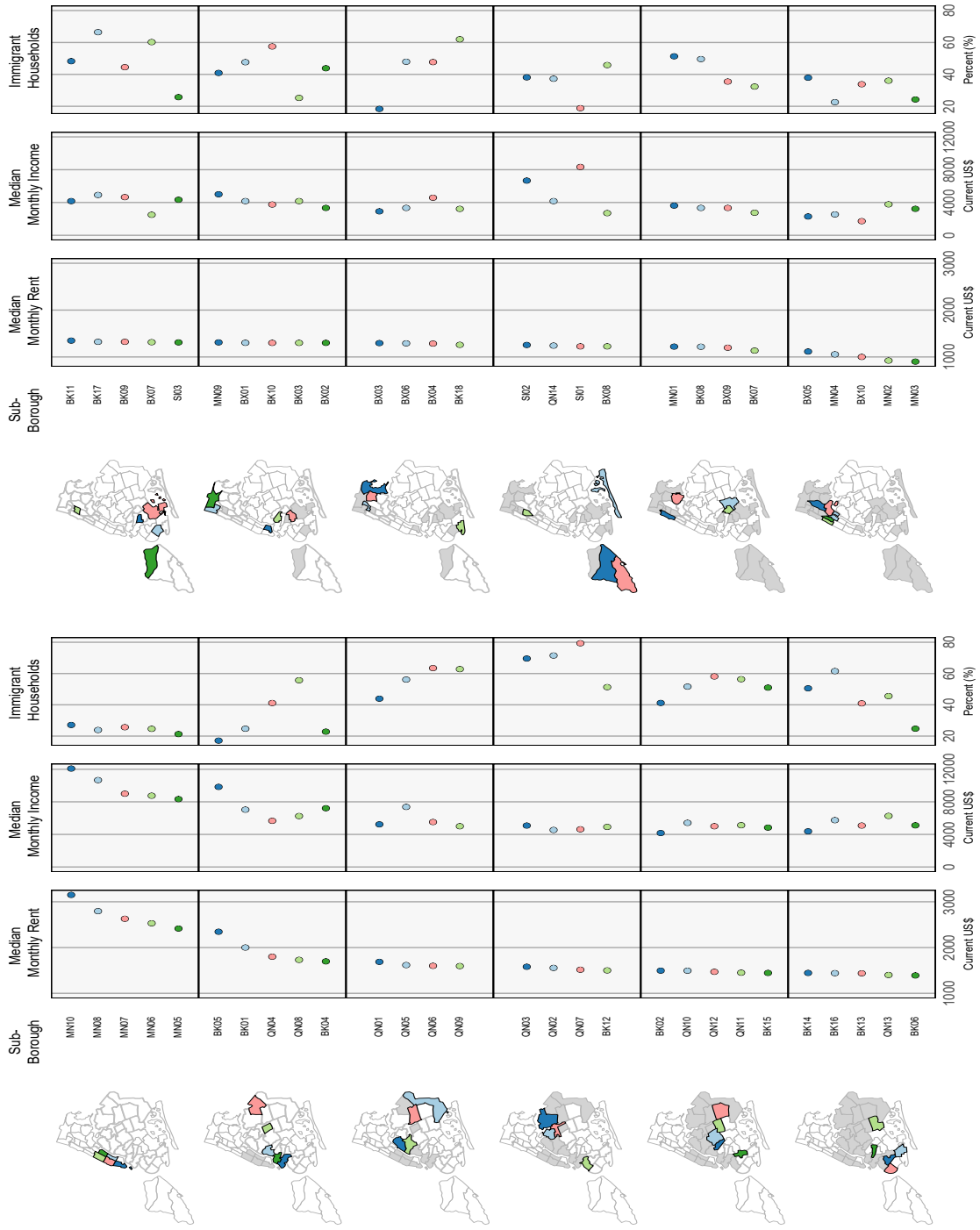

Fig. 3.8: NYC home renting LM plot for 2017. The 55 sub-boroughs are ranked in descending order by median monthly rent. The LM plot is meant to show the relationship between the primary panel, the median monthly rent, with the secondary panels, the median monthly income and the percentage of immigrant households

Table 3.2: NYC home renting correlations for 2017

	Raw Data			Aggregated Data		
	Monthly Rent	Monthly Income	Immigrant Households	Monthly Rent	Monthly Income	Immigrant Households
Brooklyn - 2,596 observations in 18 sub-boroughs						
Monthly Rent	1.00	0.49	-0.14	1.00	0.94	-0.56
Monthly Income	0.49	1.00	-0.12	0.94	1.00	-0.53
Immigrant Households	-0.14	-0.12	1.00	-0.56	-0.53	1.00
Bronx - 1,464 observations in 10 sub-boroughs						
Monthly Rent	1.00	0.37	0.09	1.00	0.67	0.36
Monthly Income	0.37	1.00	0.06	0.67	1.00	0.24
Immigrant Households	0.09	0.06	1.00	0.36	0.24	1.00
Manhattan - 2,454 observations in 10 sub-boroughs						
Monthly Rent	1.00	0.46	-0.08	1.00	0.98	-0.45
Monthly Income	0.46	1.00	-0.05	0.98	1.00	-0.41
Immigrant Households	-0.08	-0.05	1.00	-0.45	-0.41	1.00
Queens - 1,630 observations in 14 sub-boroughs						
Monthly Rent	1.00	0.37	0.01	1.00	0.42	0.10
Monthly Income	0.37	1.00	-0.05	0.42	1.00	-0.2
Immigrant Households	0.01	-0.05	1.00	0.10	-0.2	1.00
Staten Island - 218 observations in 3 sub-boroughs						
Monthly Rent	1.00	0.30	-0.04	1.00	-1.00*	0.19
Monthly Income	0.30	1.00	-0.04	-1.00*	1.00	-0.26
Immigrant Households	-0.04	-0.04	1.00	0.19	-0.26	1.00
Total - 8,362 observations in 55 sub-boroughs						
Monthly Rent	1.00	0.49	-0.09	1.00	0.88	-0.26
Monthly Income	0.49	1.00	-0.08	0.88	1.00	-0.31
Immigrant Households	-0.09	-0.08	1.00	-0.26	-0.31	1.00

In the aggregated data:

The monthly rent and income values are calculated based on the medians of the sub-boroughs.

The immigrant households considers the percentage of ”yes” answers in each sub-borough.

*This correlation was rounded from -0.99738.

Previous research supported the fact that, even for immigrants, being employed and having

sustainable income secured not only stable, but positive housing situations (Shier et al., 2016). Other variables such as social and human capital, adequate social network support, and foreign-earned credentials and education are also associated with household income and rent (George and Chaze, 2009; Nuesch-Olver, 2002).

3.3 Housing Cost Exploration

How much of your income should go to pay rent? The United States Department of Housing and Urban Development (HUD) proposed a percent rule to determine which families are cost-burdened. If a family spends more than 30% of their income in rent, such a family reports a moderate shelter cost-burden. This means the family may have difficulties having access to basic needs such as food, clothing, etc. If this number reaches 50%, the family reports a severe shelter cost-burden, which represents a serious housing affordability problem (Schwartz and Wilson, 2008). Some authors (Elmelech, 2004) define housing cost as shelter cost-burden; however, we will only refer to this variable as housing cost percentage.

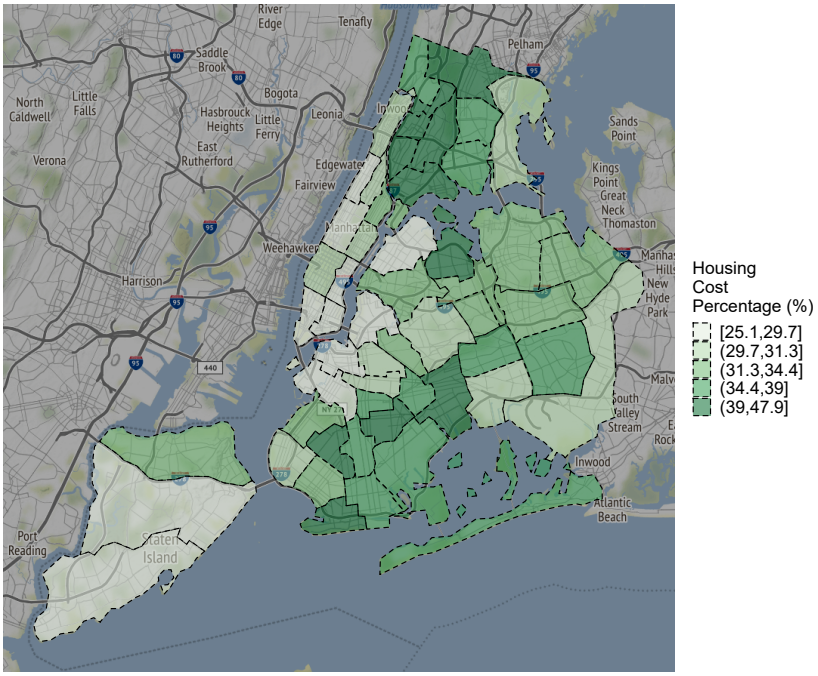

Fig. 3.9: NYC housing cost percentage choropleth map for 2017. The data are broken into quintiles, with about 11 sub-boroughs in each of the quintiles

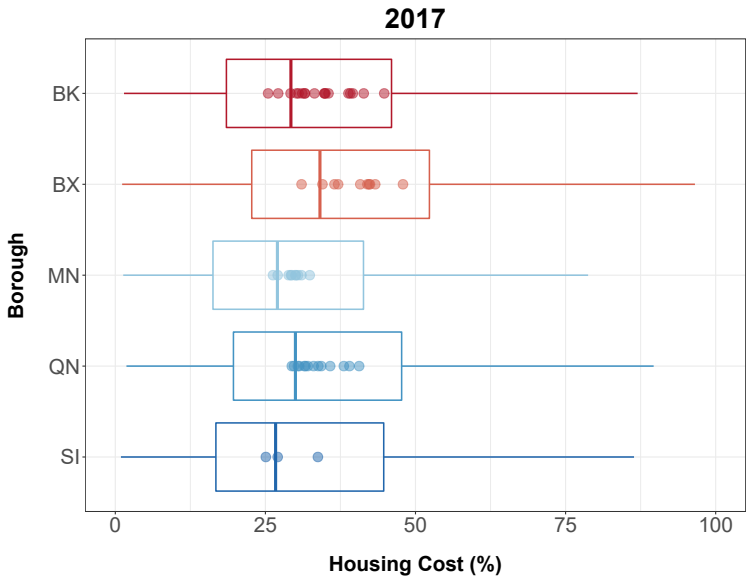

Fig. 3.10: NYC household housing cost percentage by borough for 2017. The dots represent the aggregated data of the 55 sub-boroughs, and the boxplot is constructed based on the raw survey data (8,652 observations)

In New York City, we can draw a general perspective from the data. Figure 3.9 shows that many sub-boroughs in the Bronx are the most cost-burdened, with a median housing cost percentage of over 39%. In contrast, most Staten Island, some western Queens, and most lower Manhattan sub-boroughs report median housing cost percentage below 30%. Compared to 1991, the housing cost percentage in NYC has increased in most sub-boroughs (see Figure A.4 in Appendix A for further details).

Figure 3.10 shows the housing cost percentage spread within each main borough. The Bronx borough, in addition to having the highest housing median cost, also has the highest spread with a noticeable skewness to the right. The Manhattan borough, in contrast, displays both a relatively low median cost and spread compared to the other boroughs, explained most likely by a higher household income. In the last 26 years we can see housing cost percentage has increased and reports more variability across the five boroughs (see Figure A.8 in Appendix A for further details).

Figure 3.11 depicts the relationship between housing cost percentage and rent by immigration status. The high variability in the data doesn't allow us to notice any major distinction between US citizens or immigrants. However, the smoothed lines suggest that the housing cost percentage may

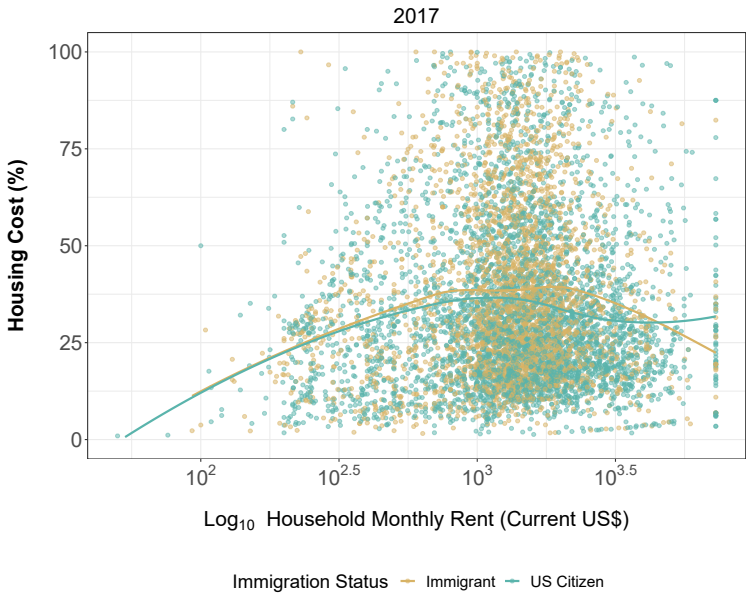

Fig. 3.11: NYC housing cost percentage and monthly rent in scatterrplot for 2017. About 13% of the data represent cases where the housing cost percentage is greater than 100%. These cases are not shown in the scatterplot. The 2017 NYCHVS survey had a maximum listed amount of $5,995 for monthly rent, and it assigned a value of $7,992 to households that listed a monthly rent above the listed maximum of $5,995 as $7,992 represented the mean of those manually reported higher monthly rents

be a few percentage points higher for immigrants than for US citizens. This difference seems to be lower compared to previous years (see Figure A.11 in Appendix A for further details).

The LM plot in Figure 3.12 confirms observations from Table 2.3 that the sub-boroughs of the Bronx are among the ones with the highest median housing cost percentage. We can see that eight of the ten sub-boroughs with the highest housing cost percentage are located in Bronx. In contrast, two sub-boroughs in southern Staten Island, five in southern and middle Manhattan, and three in southwestern Brooklyn are among the ten with the lowest median housing cost percentage. The correlation between housing cost percentage and immigrant households percentage is not very clear in the LM plot, so we will look at specific correlations across all five boroughs.

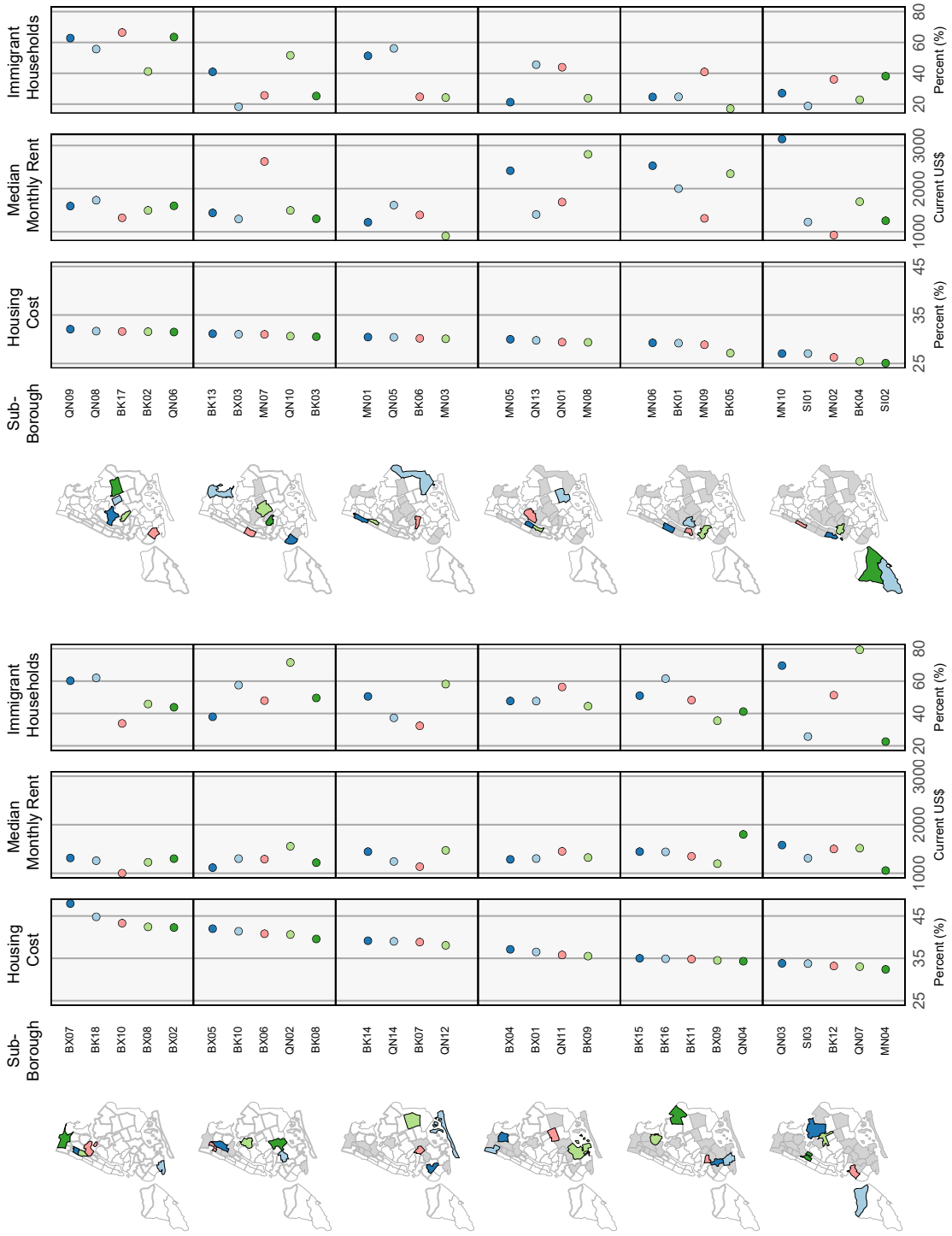

Fig. 3.12: NYC housing cost percentage LM plot for 2017. The 55 sub-boroughs are ranked in descending order by housing cost percentage. The LM plot is meant to show the relationship between the primary panel, the housing cost percentage, with the secondary panels, the median monthly rent and the percentage of immigrant households

Table 3.3: NYC housing cost percentage correlations for 2017

	Raw Data			Aggregated Data		
	Housing Cost %	Monthly Rent	Immigrant Households	Housing Cost %	Monthly Rent	Immigrant Households
Brooklyn - 2,596 observations in 18 sub-boroughs						
Housing Cost %	1.00	0.01	-0.02	1.00	-0.64	0.67
Monthly Rent	0.01	1.00	-0.14	-0.64	1.00	-0.56
Immigrant Households	-0.02	-0.14	1.00	0.67	-0.56	1.00
Bronx - 1,464 observations in 10 sub-boroughs						
Housing Cost %	1.00	0.05	-0.05	1.00	-0.20	0.68
Monthly Rent	0.05	1.00	0.09	-0.20	1.00	0.36
Immigrant Households	-0.05	0.09	1.00	0.68	0.36	1.00
Manhattan - 2,454 observations in 10 sub-boroughs						
Housing Cost %	1.00	0.03	0.00	1.00	-0.15	-0.21
Monthly Rent	0.03	1.00	-0.08	-0.15	1.00	-0.45
Immigrant Households	0.00	-0.08	1.00	-0.21	-0.45	1.00
Queens - 1,630 observations in 14 sub-boroughs						
Housing Cost %	1.00	0.02	-0.03	1.00	-0.39	0.13
Monthly Rent	0.02	1.00	0.01	-0.39	1.00	0.10
Immigrant Households	-0.03	0.01	1.00	0.13	0.10	1.00
Staten Island - 218 observations in 3 sub-boroughs						
Housing Cost %	1.00	0.24	-0.03	1.00	0.84	-0.37
Monthly Rent	0.24	1.00	-0.04	0.84	1.00	0.19
Immigrant Households	-0.03	-0.04	1.00	-0.37	0.19	1.00
Total - 8,362 observations in 55 sub-boroughs						
Housing Cost %	1.00	0.01	-0.02	1.00	-0.39	0.43
Monthly Rent	0.01	1.00	-0.09	-0.39	1.00	-0.26
Immigrant Households	-0.02	-0.09	1.00	0.43	-0.26	1.00

In the aggregated data:

The monthly rent is calculated based on the median of the sub-boroughs.

The immigrant households considers the percentage of ”yes” answers in each sub-borough.

Table 3.3 shows the correlation coefficients among housing cost percentage, monthly rent, and immigration status for the raw data (8,362 observations in 55 sub-boroughs) and the aggregated data

(based on aggregations within each sub-borough). When looking at the raw data, Table 3.3 reveals that there is basically no association among the three variables. We notice no linear association when looking at the individual correlation coefficients between immigration status and housing cost percentage for the Bronx ($r = -0.05$), Queens ($r = -0.03$), Staten Island ($r = -0.03$) , Brooklyn ($r = -0.02$), and Manhattan ($r = 0.00$) boroughs. However, there is a very weak negative correlation ($r = -0.09$) between median monthly rent and immigration, as discussed in Section 3.2. Finally, we notice a correlation of almost zero between housing cost percentage and monthly rent in four of the five boroughs, with Staten Island being the exception ($r = 0.24$).

In contrast, we notice some relationships when looking at the aggregated data of the 55 sub-boroughs. The correlation between percent of immigrant households and housing cost percentage is positive and moderate with $r = 0.43$, specially strong in the Brooklyn ($r = 0.67$) and Bronx ($r = 0.68$) boroughs. In other words, being a immigrant household in those areas is associated with having a higher housing cost percentage. However, it's noteworthy mentioning that the reason behind the difference between the correlations found in the raw and the aggregated data is the high variability within the data, specially in the housing cost percentage data as depicted in Figures 3.10, 3.11, and 3.12.

We also get stronger results when looking at the correlation between housing cost percentage and median monthly rent with $r = -0.39$, specially in Brooklyn ($r = -0.64$) and Staten Island ($r = 0.84$). The correlations between monthly rent and immigrant households percentage were discussed in Section 3.2.

Previous research on this topic indicated that racial/ethnic differentials are too complex to be analyzed generally because of a high variability within those groups. However, during times of large-scale migration and a shortage of affordable housing, immigrants could experience a higher shelter cost-burden (Elmelech, 2004). In addition, human capital characteristics, stage in life, traditional assimilation, and contextual variables are also associated with a high immigrant housing cost percentages in the United States (McConnell and Akresh, 2010).

CHAPTER 4

Spatial Dependence Tests and Models

4.1 Spatial Dependence Tests

In this section, we will assess spatial autocorrelation in our three housing variables, as well as in the immigration data. The illustration of the methods part of this section is based on the work of Getis (1991, 2010).

4.1.1 Methods

We define spatial autocorrelation as the relationship between our observed variables in each of the localities n, in this case the 55 NYC sub-boroughs, and a measure of geographical proximity defined for all $n(n-1)$ pairs of sub-boroughs. A positive spatial autocorrelation takes place when near locations report similar values for a variable of interest; for example, high values of one location are associated with high values of nearby locations. In contrast, in a negative spatial autocorrelation, those places report a negative relation in those values; for example, high values in one location are associated with low values in nearby locations. Instead of positive and negative spatial autocorrelation, one can also speak of clustering and dispersion, respectively.

Equation 4.1 shows a simple, introductory representation of this definition that could help us choose the appropriate statistical test for assessing spatial autocorrelation (Hubert and Golledge, 1981; Hubert et al., 1981):

$$\Gamma = \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} Y_{ij}, \tag{4.1}$$

where Γ is a measure of spatial autocorrelation for n georeferenced observations. While W, a matrix of weights, represents the values of these spatial relationships of each location i to all other locations j, matrix Y represents the non-spatial relationship, such as a covariance matrix. W and Y are $n \times n$ matrices. This measure Γ is also known as the cross-product statistic.

In our analyses, we use three different measures of proximity, schemes to structure the W matrix (Getis and Aldstadt, 2004): spatially contiguos neighbors (such as "Queen"), k-nearest neighbors (equal weighting of matrix entries), and centroids within a distance d (density dependent).

In addition, we use Moran's I and Geary's C statistics as our spatial autocorrelation measures. Both, as global statistics, indicate the existence of a positive or negative spatial autocorrelation, and the degree of it. A global measure means we take all the elements of the W and Y matrices and assess the global spatial autocorrelation. In a local measure, instead of using all elements, we would consider particular spatial units, which result in so-called local indicators of spatial association (LISA) statistics (Anselin, 1995) which are not considered here.

Moran's I statistic (Moran, 1948) is based on a covariance structure, and its estimation is depicted in equation 4.2. We assume $W_{ii} = 0$, which means there is no self association. This statistic can take a value approximately between -1 and $+1$; a positive value means there is a positive spatial autocorrelation and a negative value means there is a negative autocorrelation. The closer this statistic is to -1 or $+1$, the stronger the spatial autocorrelation is.

$$I = \frac{n}{\left(\sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} W_{ij}\right)} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)} ; W_{ii} = 0. \tag{4.2}$$

Geary's C is another statistic (Geary, 1954) used to measure spatial autocorrelation, defined in equation 4.3. While Geary's C also is a measure of global spatial autocorrelation, it is more sensitive than Moran's I to absolute differences between neighboring variables. Geary's C can take a value between 0 and 2 in most of the cases. A small value ($0 < C < 1$) represents a positive spatial autocorrelation, and values close to 0 are an indicator of a strong positive spatial autocorrelation. Values of $C > 1$ represent a negative autocorrelation, and values $>> 1$ are an indicator of a strong negative spatial autocorrelation. Geary's C is negatively related to Moran's I, since higher values of the second one ($I > 0$) represent a positive spatial autocorrelation.

$$C = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(y_i - y_j)^2}{2\left(\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} W_{ij}\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}; W_{ii} = 0. \tag{4.3}$$

Both Moran's I and Geary's C are approximately Normal distributed, with means $\mu_I = -\frac{1}{n-1}$ and $\mu_C = 1$; ranges r_I and r_C, and variances σ_I^2 and σ_C^2 are functions of the W_{ij}'s; it always follows $C \geq 0$ (De Jong et al., 1984).

We can perform hypothesis tests for spatial randomness using Moran's I and Geary's C statistics. We can use z-tests for the null hypothesis H_0: "data does not show spatial autocorrelation" as follows: calculate $\frac{I-\mu_I}{\sigma_I}$ or $\frac{C-\mu_C}{\sigma_C}$ and see whether this is significantly different from 0. Possible alternative hypotheses are the one-sided alternatives H_a : "data is spatially clustered" and H_a : "data is spatially dispersed", and the two-sided alternative H_a : "data shows spatial autocorrelation". Similar to the work of Leung et al. (2000), we consider the one-sided alternative H_a : "data is spatially clustered" and a result significant when it reaches a p-value of at most 5%. Alternatively, instead of using z-tests, we could also use permutation tests (see Bailey and Gatrell (1995), page 281).

4.1.2 Results and Discussion

After tuning parameters for our three measures of proximity schemes to structure the W matrix, we decided to use the "Queen" approach for contiguous neighbors, the three nearest neighbors, and a maximum proximity distance of 8 km. A geographical representation of these proximity schemes is provided in Appendices B.1 - B.5 for each of the five boroughs. In the case of the maximum proximity distance, the largest minimum distance between two sub-boroughs centroids is 7.89km. These measures reported the strongest Moran's I and Geary's C statistics for home ownership percentage, median household rent, housing cost percentage, and immigrant household percentage in the year 2017. Therefore, we used the same settings for the other years.

Appendices C.1 - C.8 provide more details about the results of the tuning of these measures of proximity. Appendices C.9 - C.12 provide results not discussed in this section about spatial auto-

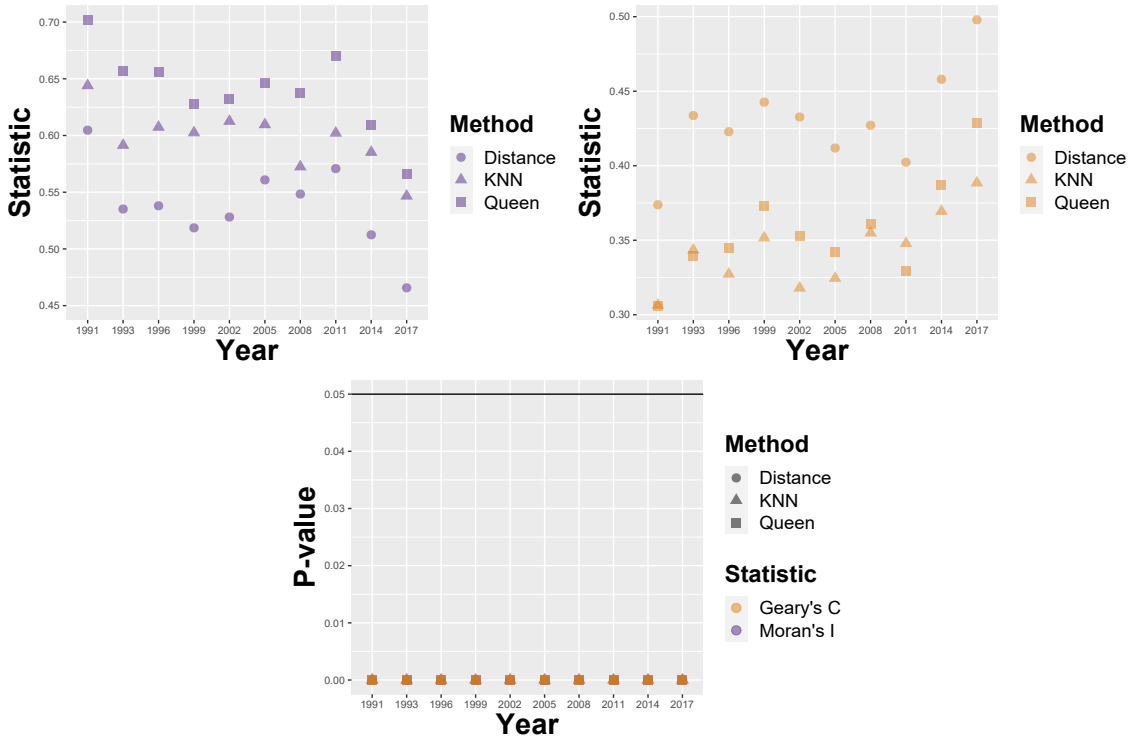

Fig. 4.1: NYC home ownership percentage spatial autocorrelation 1991 - 2017. The top left graph shows the values of Moran's I statistic. The top right graph shows the values of Geary's C statistic. The bottom graph shows the p-values obtained from the hypothesis tests for spatial randomness using both statistics

correlation tests in other housing and sociodemographic variables such as housing value, household gross income, female householder percentage, and householder age. Spatial autocorrelation can also be found in these variables for most years. In addition, the hypothesis tests are normally-based and their results were considered significant if they resulted in a p-value of at most 5%. We did not adjust the p-values for multiple tests and would leave it to the reader to make necessary adjustments of the p-values as outlined in Wright (1992).

Figures 4.1 (home ownership percentage), 4.2 (median household rent), 4.3 (housing cost percentage), and 4.4 (immigrant household percentage) show our results for all four variables. In every figure, the graph in the top left shows the values of Moran's I statistic, while the graph in the top right shows the values of Geary's C statistic. These statistics are useful to assess the strenght of the spatial autocorrelation on a certain variable, and also whether it is positive or negative. Finally, the graph in the bottom depicts whether this relationship is significant by showing the obtained p-value

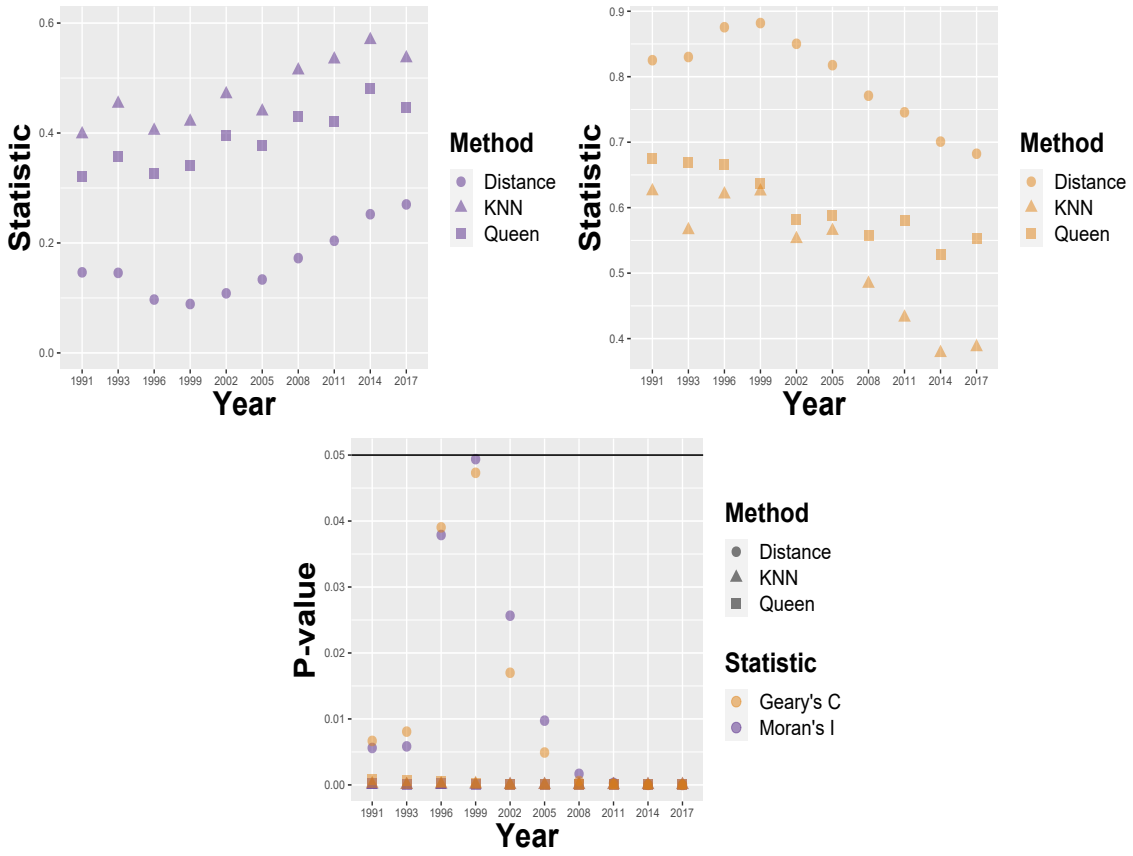

Fig. 4.2: NYC median household rent spatial autocorrelation 1991 - 2017. The top left graph shows the values of Moran's I statistic. The top right graph shows the values of Geary's C statistic. The bottom graph shows the p-values obtained from the hypothesis tests for spatial randomness using both statistics

for both tests. Overall, the results indicate there is a strong positive spatial autocorrelation for most of the variables in all years.

Figure 4.1 shows that in every year, using every proximity method, spatial autocorrelation results for both tests for home ownership percentage were the most significant among the four variables analyzed, with p-values less than 1%. The "Queen" contiguous neighbors method reported the strongest autocorrelation in Moran's I every year, while the three nearest neighbors in Geary's C in seven out of ten years. Moreover, this strong spatial autocorrelation tells us that there likely is a strong spatial pattern in the data. This confirms our preliminary observations in Figure 3.1 where contiguous sub-boroughs in Staten Island and Eastern Queens show the highest percentages, while contiguous sub-boroughs in Northern Manhattan and Western Bronx show the lowest home

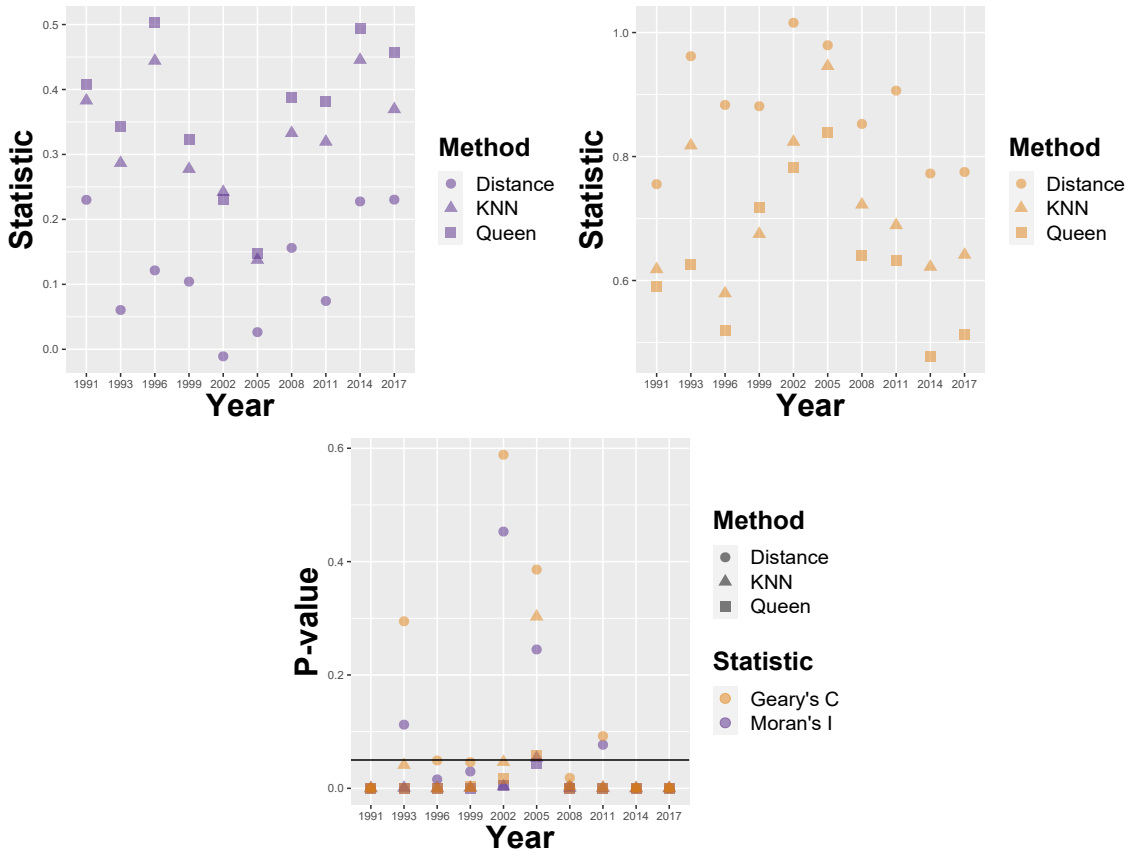

Fig. 4.3: NYC housing cost percentage spatial autocorrelation 1991 - 2017. The top left graph shows the values of Moran's I statistic. The top right graph shows the values of Geary's C statistic. The bottom graph shows the p-values obtained from the hypothesis tests for spatial randomness using both statistics

ownership percentages.

In the case of median household rent, Figure 4.2 shows significant results in every year with every proximity method and both statistical tests. For this variable, in contrast to home ownership percentage, the three nearest neighbors method was the proximity method that reported most significance. In contrast, a distance approach, although significant at 5%, reported the weakest spatial autocorrelation. When looking at Figure 3.5, this makes sense as we can see that the sub-boroughs have either a gradual increase or decrease when looking at some, but not all, contiguous neighbors. An exception would be a cluster in the Southern Manhattan area where rent is high among most sub-boroughs.

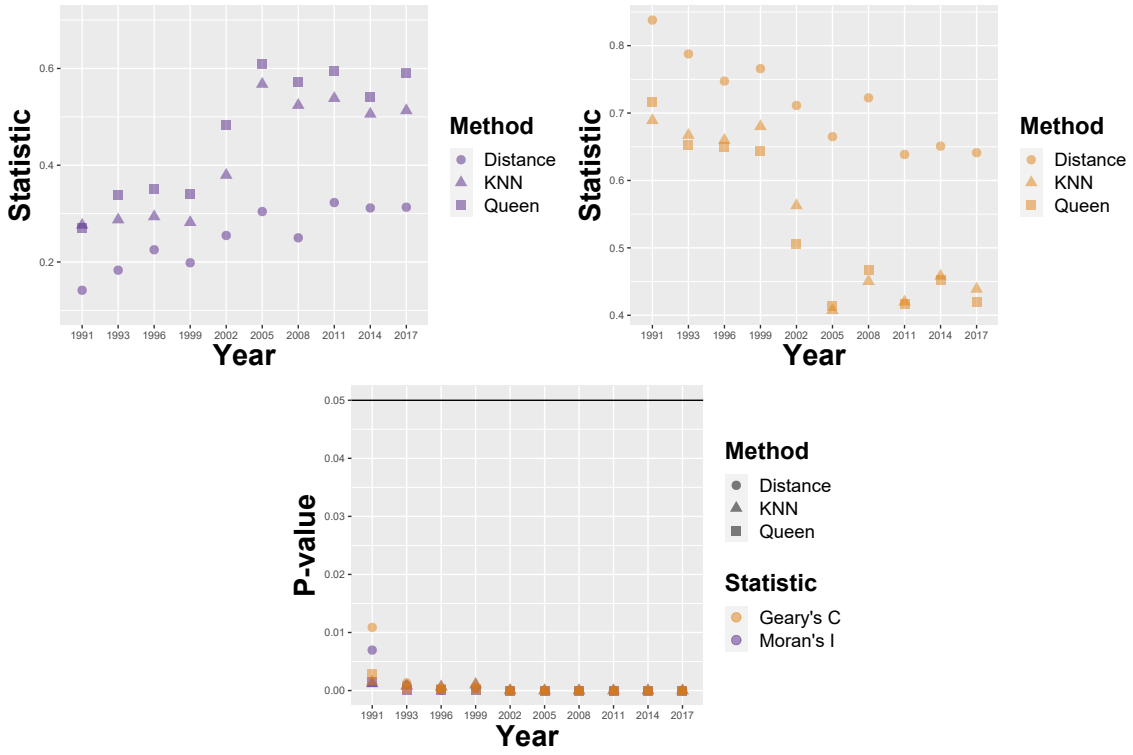

Fig. 4.4: NYC immigrant household percentage spatial autocorrelation 1991 - 2017. The top left graph shows the values of Moran's I statistic. The top right graph shows the values of Geary's C statistic. The bottom graph shows the p-values obtained from the hypothesis tests for spatial randomness using both statistics

Similarly, Figure 4.3 shows significant positive spatial autocorrelation in housing cost percentage, overall. However, using distance as the proximity method, there was no significant spatial autocorrelation for 1993, 2002, 2005, and 2011 at the 5% level. In contrast, contiguous and nearest-neighbors reported both a strong positive spatial autocorrelation and small p-values except in 2005 when only the "Queen" method in combination with Moran's I resulted in a significant p-value. As reviewed in Chapter 3.3, the fact that the housing cost data has a high variability may explain why these results change more compared to the other variables.

Finally, in the case of the immigrant household percentage, we can see in Figure 4.4 that the spatial autocorrelation got stronger over time for all proximity methods. Although the contiguous "Queen" neighbor method had stronger statistics with lower p-values associated, the three nearest neighbors and distance based proximity approaches also reported strong significance. The fact that the spatial autocorrelation increased over time could be providing information about an immigration

pattern. A plausible explanation could be that NYC immigrants in the 1990s did not have a specific preference about where to settle, while in later years, based on previous immigrant experiences, they were more likely to settle in specific sub-boroughs.

Previous research has found spatial autocorrelation in housing and sociodemograhic variables, such as immigration status, race, or ethnicity in other cities in the U.S. (Zou, 2014), China (Liu et al., 2020), and Australia (Pettit et al., 2017). These studies noticed some sociodemographic groups tend to settle in specific places, showing a spatial pattern. In addition, housing variables also displayed a spatial pattern among the population, but its conditions could show some disparity among several sociodemographic groups. Although in this thesis we didn't analyze disparities among groups, we noticed strong spatial patterns in both immigration and housing variables.

4.2 Spatial Autoregressive Models

In Section 4.1, we showed that the NYC data is spatially autocorrelated. In this section, we build Spatial Autoregressive (SAR) models to try to predict our three housing variables using immigration status, among other housing and sociodemographic variables. For comparison purposes, we also build Ordinary Least Squares (OLS) models. The OLS methods section of this text is based on the work of Hayes and Cai (2007) and the SAR methods section on the work of Getis (1991, 2010).

4.2.1 Methods

OLS models are our starting point and can be defined by the following equation:

$$y = X\beta + \epsilon \tag{4.4}$$

where $y_{n\times 1}$ is a vector of observations that represents any of our housing variables, $X_{n\times k}$ is a matrix of the predictor variables, $\beta_{k\times 1}$ is a vector of coefficients associated with the predictors, and $\epsilon_{n\times 1}$ is a vector that represents the residual term of the model. Notice n is the sample size and k is the number of independent variables to be used in the model.

OLS models include six assumptions: (1) Linearity in parameters, (2) Random sampling of observations, (3) Expected value of the mean of the error equal to zero, (4) No multi-collinearity

between variables, (5) Homoscedasticity and no autocorrelation, and (6) Error terms normally distributed. However, as discussed in Section 4.1, our data reports spatial autocorrelation. This could violate some OLS assumptions, especially assumption (5). That is why we will use SAR models to address this issue.

Generally attributed to Whittle (1954) with later developments in econometrics (Anselin, 2013; LeSage and Pace, 2009), SAR models can be defined by the following equation:

$$y = X\beta + \rho W y + \varepsilon \tag{4.5}$$

$$\varepsilon \sim N(0_{nx1}, \sigma^2 I_n) \tag{4.6}$$

where, similar to Equation 4.4, $y_{n\times 1}$ is a vector of observations that represents any of our housing variables, $X_{n\times k}$ is a matrix of the predictor variables, $\beta_{k\times 1}$ is a vector of coefficients associated with the predictors, and $\epsilon_{n\times 1}$ is a vector that represents the residual term of the model. In addition, $W_{n\times n}$ is a spatial weighting matrix of known constants, ρ a scalar autoregressive parameter, and the vector $Wy_{n\times 1}$ reflects a spatial lag of $y_{n\times 1}$.

Based on our previous results, we will consider three different spatial weighting matrices: one for the "Queen" contiguous neighbor method, one for the k-neighbor method (with the three closest neighbors), and one for the distance proximity method (with a maximum distance of 8 km). In addition, we will select variables that have a Variance Inflation Factor (VIF) of less than 10 to avoid multicollinearity issues (Mansfield and Helms, 1982). The VIF is computed (Neter et al. (1983), p. 391) as

$$VIF_j = \frac{1}{1 - R_j^2}, j = 1, \ldots, k \tag{4.7}$$

where R_j^2 is the multiple correlation coefficient which gives the proportion of variance in the j^{th} predictor associated with the remaining $k-1$ predictors. When the VIF_j is closer to zero, we can imply that predictor is independent from all other predictors.

The variables used in this section have been described in Table 2.1 and we include the VIF

results in Appendix D. All models take into account either the median value or the percentage of "1s" of all variables across all 55 sub-boroughs during the year 2017. Thus, the number of observations n is equal to 55.

To estimate the models we make use of the *lagsarlm* function from the "sp" R package (Bivand et al., 2013). The *lagsarlm* function provides maximum likelihood estimation of SAR lag where the parameter ρ is found through an optimization process (Griffith, 1988).

4.2.2 Results and Discussion

Analyzing the results, we notice housing variables depend on many variables related to the householder sociodemographics, housing value and features, and location. Tables E.1, E.2, and E.3 show the estimates for each of the OLS and SAR models. Figures 4.5, 4.6, and 4.7 show choropleth maps that illustrate the residuals for each SAR model.

Overall, SAR models were more accurate, they had higher R-squared values, and they were more significant, compared to OLS models. Including a proximity matrix to weight the coefficients improved the significance of some explanatory variables in the model. Regarding the weighting matrix, the distance proximity method reported a higher R-squared and more significance for the home ownership percentage model (see Table E.1). In the case of home renting and housing cost percentage, the knn (see Table E.2) and the "Queen" (see Table E.3) method for contiguous neighbors were more accurate, respectively.

Table E.1 shows the estimates for the home ownership percentage model. Although all models have similar R-square, we notice that the distance proximity method has most significant variables. Householder sociodemographic variables, such as immigration status, age, and income, reported significance. Other housing variables, such as the housing value and the conditions of the walls also showed a relation with the model. Living in the Bronx, Brooklyn, or Manhattan has an effect on the model; as we saw in Figure 3.1, these boroughs report lower home ownership percentages than the two other boroughs, Staten Island and Queens.

Table E.2 shows the results for the home renting model. We notice that all SAR methods had a very similar R-squared and significant variables. Householder sociodemographic variables, besides

Table 4.1: OLS and SAR model estimates for home ownership percentage model

Variable	**OLS**	**SAR: Queen**	**SAR: knn**	**SAR: max. dist**
Intercept	-4.232***	-4.220***	-4.339***	-4.366***
Householder Sociodemographics				
Householder's Birthplace	-0.352**	-0.336***	-0.344***	-0.345***
Householder's Age	0.029***	0.029***	0.030***	0.030***
Householder's Sex	-0.080	-0.066	-0.064	-0.071
Log Household Gross Income	0.139*	0.132**	0.143**	0.139**
Housing Value and Features				
Log Housing Gross Value	0.102	0.102*	0.101*	0.106**
Mortgage Status	-0.093	-0.110	-0.105	-0.117
Mortgage Interest Rate	-0.236	-0.158	-0.153	-0.160
Walls Condition	0.830*	0.799**	0.732*	0.813**
Stairs Condition	-0.007	0.029	0.071	0.027
Number of Rooms	-0.027	-0.031	-0.035	-0.032
Location Variables				
Brooklyn Location	-0.190**	-0.186***	-0.175***	-0.204***
Bronx Location	-0.156*	-0.145*	-0.129*	-0.153**
Manhattan Location	-0.455***	-0.445***	-0.432***	-0.465***
Queens Location	-0.030	-0.047	-0.041	-0.068
R-squared	0.789	0.793	0.796	0.794
ρ	0	0.128	0.170	0.214

Notes: *** 1% level of significance; ** 5% level of significance;
* 10% level of significance. Staten Island is used as the baseline.

of immigration, reported significance as in our previous model. Surprisingly, the number of rooms reported a negative relationship with household rent because many sub-boroughs in Manhattan and some in Brooklyn tend to rent less rooms at a much higher amount compared to the rest of NYC. This claim makes sense when we look at the location variables where the Brooklyn and Manhattan estimates are the lowest among all.

Table E.3 shows the regression output for the housing cost percentage model. We notice that all SAR methods had a very similar R-squared and significant variables. Householder sociodemographic variables such as householder immigration status and sex are highly significant; being a female or immigrant householder is associated with a higher housing cost percentage. In addition, the Bronx borough reported a higher housing cost burden.

Table 4.2: OLS and SAR model estimates for log median household rent model

Variable	**OLS**	**SAR: Queen**	**SAR: knn**	**SAR: max. dist**
Intercept	-0.534	-0.831	-0.921	0.466
Householder Sociodemographics				
Householder's Birthplace	0.104	0.107	0.107	0.078
Householder's Age	-0.009	-0.009*	-0.009*	-0.009*
Householder's Sex	0.599	0.620*	0.636*	0.595*
Log Household Gross Income	0.545***	0.542***	0.549***	0.540***
Housing Value and Features				
Walls Condition	1.100	1.125	1.050	0.886
Stairs Condition	-0.454	-0.479	-0.436	-0.437
Floors Condition	1.583	1.629*	1.602*	1.750*
Number of Rooms	-0.100***	-0.100***	-0.100***	-0.104***
Location Variables				
Brooklyn Location	0.155*	0.157**	0.159**	0.163**
Bronx Location	0.174*	0.179**	0.184**	0.163*
Manhattan Location	0.134	0.139*	0.136*	0.128*
Queens Location	0.170*	0.172**	0.173**	0.178**
R-squared	0.842	0.843	0.842	0.844
ρ	0	0.037	0.045	-0.121

Notes: *** 1% level of significance; ** 5% level of significance;
* 10% level of significance. Staten Island is used as the baseline.

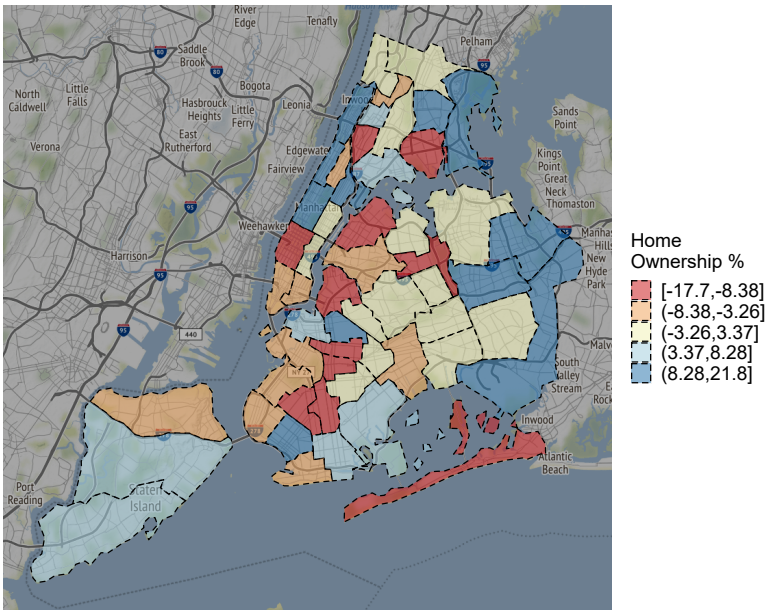

Fig. 4.5: NYC home ownership percentage SAR model: maximum distance residuals map for 2017. The data are broken into quintiles, with about 11 sub-boroughs in each of the quintiles

Table 4.3: OLS and SAR model estimates for housing cost percentage

Variable	**OLS**	**SAR: Queen**	**SAR: knn**	**SAR: max. dist**
Intercept	-0.614*	-0.621**	-0.606**	-0.660**
Householder Sociodemographics				
Householder's Birthplace	0.111**	0.116***	0.113***	0.106***
Householder's Age	0.001	0.002	0.002	0.001
Householder's Sex	0.303***	0.313***	0.306***	0.306***
Housing Value and Features				
Walls Condition	0.287	0.297	0.284	0.261
Stairs Condition	0.014	0.012	0.014	0.026
Floors Condition	0.405	0.417	0.408	0.440
Number of Rooms	-0.015	-0.015	-0.014	-0.015*
Location Variables				
Brooklyn Location	0.032	0.034	0.033	0.031
Bronx Location	0.067**	0.071***	0.069***	0.062**
Manhattan Location	-0.010	-0.008	-0.009	-0.013
Queens Location	0.022	0.024	0.023	0.021
R-squared	0.673	0.674	0.673	0.675
ρ	0	-0.071	-0.034	0.115

Notes: *** 1% level of significance; ** 5% level of significance;
* 10% level of significance. Staten Island is used as the baseline.

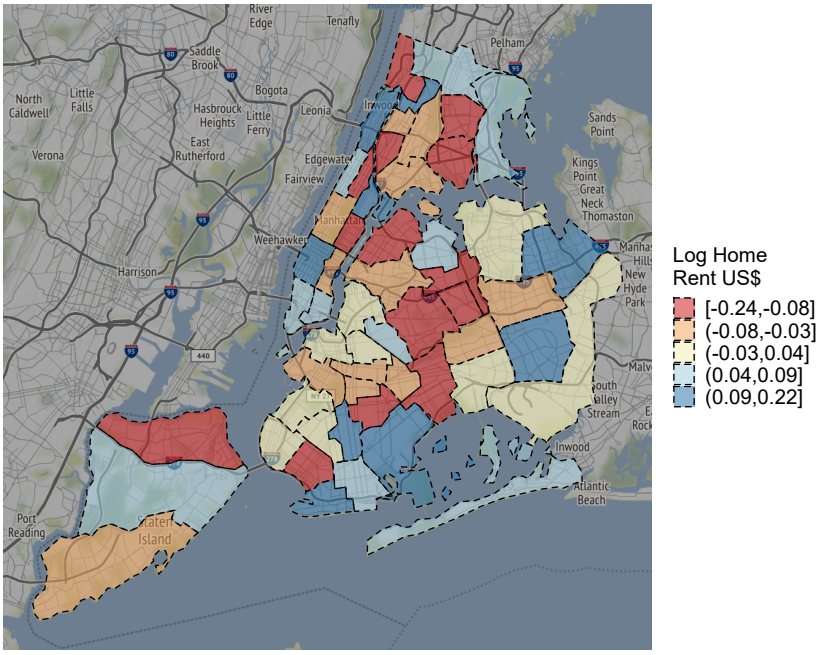

Fig. 4.6: NYC log median household rent SAR model: three nearest neighbors residuals map for 2017. The data are broken into quintiles, with about 11 sub-boroughs in each of the quintiles

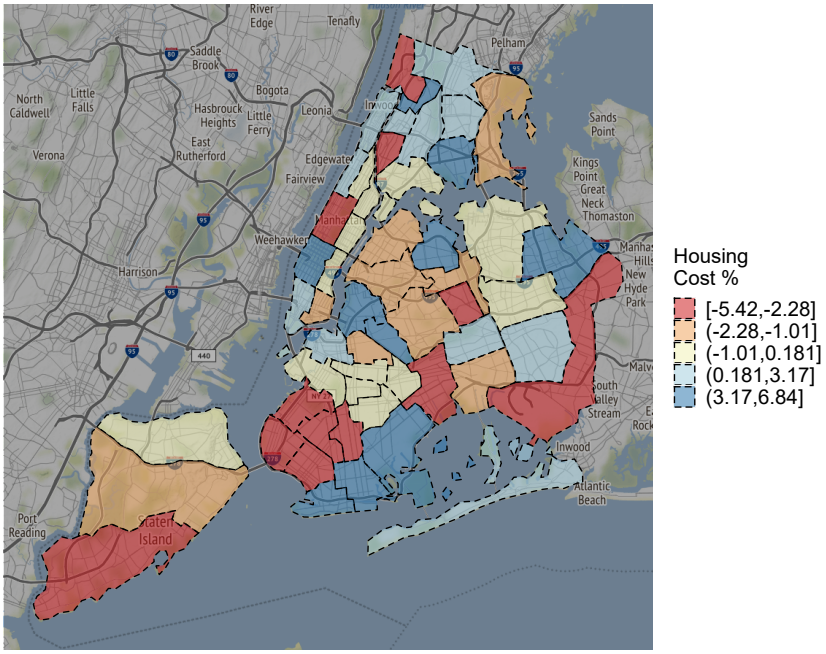

Fig. 4.7: NYC housing cost percentage SAR model: "Queen" residuals map for 2017. The data are broken into quintiles, with about 11 sub-boroughs in each of the quintiles

Figures 4.5, 4.6, and 4.7 show that the residuals don't seem to be spatially autocorrelated as we can't clearly identify any cluster. We confirmed this observation by running spatial autocorrelation tests where we got Moran's I statistics of 0.610, 0.222, and -0.053 for the home ownership percentage, log home renting, and housing cost percentage residuals respectively. We failed to reject the null hypothesis that indicates that the residuals do not show spatial autocorrelation in every model. In this case, the alternative is one-sided and indicates that the data is spatially clustered.

When looking at immigration, we can highlight a couple of findings based on these models. First, immigrant households tend to own a home at a smaller percentage than US citizens. Previous research indicates that the rate of advancement into homeownership of immigrants, relative to native borns, is usually slower because the assimilation process usually takes place in the long term (Myers and Lee, 2018; Sinning, 2010). In addition, other demographic and economic factors as well as length of residence tend to influence owner occupation (Nygaard, 2011).

Second, as discussed in Chapter 3.2, immigration status doesn't seem to be a relevant variable when explaining rent, since other variables such as income and location could be playing a more

explanatory role. Previous research suggests that immigration flows had a significant positive impact on housing rent in the long term (Latif, 2015); however, some studies that found a positive effect of immigration inflows on rent separated the effects, by variables such as education level, location, or crime rate, to deal with the inherent variability of the data (Mussa et al., 2017; Ottaviano and Peri, 2012).

Finally, our results suggest that immigrant households bear a higher housing cost burden compared to native borns. Previous research has found that human capital characteristics, stage in the life course, traditional assimilation indicators, and contextual variables are associated with immigrant housing cost burden, even though the proportions could differ among immigration groups (McConnell and Akresh, 2010). This burden exacerbates if we take into account the immigrant lawful status (Allen, 2020). Our analysis doesn't make a major differentiation among immigrants, since it also provides some insights in each borough.

A similar analysis was conducted using the spatial conditional and simultaneous autoregression model estimation. This method takes family and weights arguments for SAR model estimation via Maximum Likelihood, using dense matrix methods. The implementation is Generalized Least Squares (GLS) using a single spatial coefficient value, usually defined as λ. To estimate these models, we made use of the *spautolm* function, also, from the "sp" R package (Bivand et al., 2013). For further details, check Appendix E. Results were very similar, but resulted in a slightly lower r-squared value and significance, specially in the location variables. We leave it to the reader to try different variables or spatial estimation techniques (Bivand et al., 2021).

CHAPTER 5

Interactive Shiny R User Application

Because housing and sociodemographic variables could have a complex relationship, static visualizations are limited. Therefore, we developed a virtual interface through Shiny R (Chang et al., 2021), so any user can run a more customized data exploration. This Shiny app, inspired in the work of Parker (2021) and Medri et al. (2021), can be accessed at:

https://github.com/asa-stat-computing-and-graphics/COST-DataExpo-2019.

The app has currently four menu options that we will discuss in this chapter. Those features include the creation of a choropleth map (Section 5.1), a linked micromap (LM) plot (Section 5.2), a smoothed scatterplot (Section 5.3), and a spatial autocorrelation statistical test (Section 5.4). All these features of the app allow the user to make assessments for any of the ten years (ranging from 1991 to 2017) that are available in the data set. See Chapter 3 for more details about the choropleth maps, LM plots, and smoothed scatterplots methods, and Chapter 4 for more details about the spatial autocorrelation tests methods.

5.1 Choropleth Maps

The first menu option allows the user to create a customized overlayed choropleth map, like the ones shown in Figures 3.1, 3.5, and 3.9. Figures 5.1 and 5.2 show an example of this menu option. There are five selection features in which the user can customize the overlayed choropleth map. Those features include:

- A selection of eight variables such as home ownership percentage (percent of tenants that own a home in a sub-borough), household monthly rent (current US$), housing cost percentage, immigrant households percentage (percent of households with a householder whose birthplace is outside the U.S.), household annual income (current US$), householder sex (percent of female householders), and median householder age. See further details in Table 2.1.

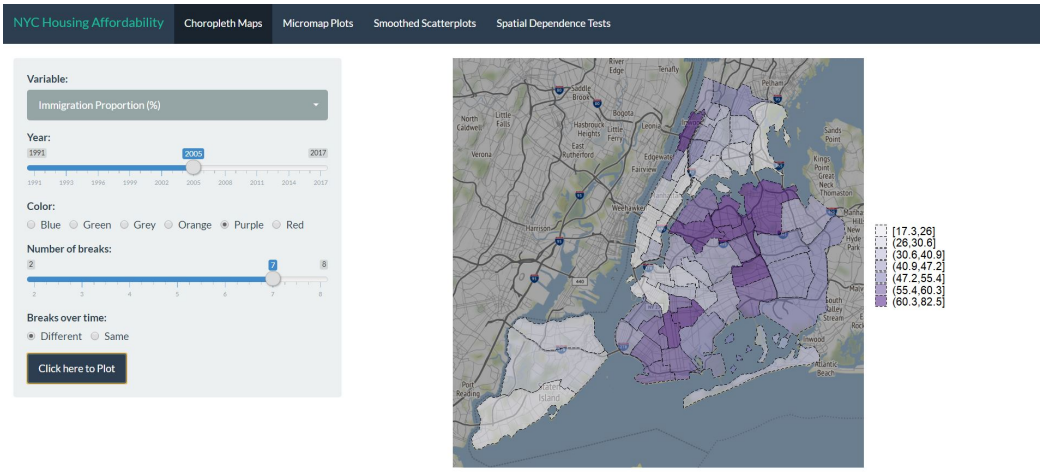

Fig. 5.1: Shiny R overlayed choropleth map interface. There are five selection features on the left side, and the map plot on the right. This is a choropleth map for percentage of immigrant households in the 2005 NYCHVS survey. It was shaded in different tones of purple and broken in septiles, with about seven or eight sub-boroughs in each septile. The breaks are calculated based on the information of that particular year. We can see that in the year 2005 the percentage of immigrant households is greatest in the eastern Queens sub-boroughs (above 60%), but lower in all Staten Island and most Manhattan sub-boroughs (below 30%)

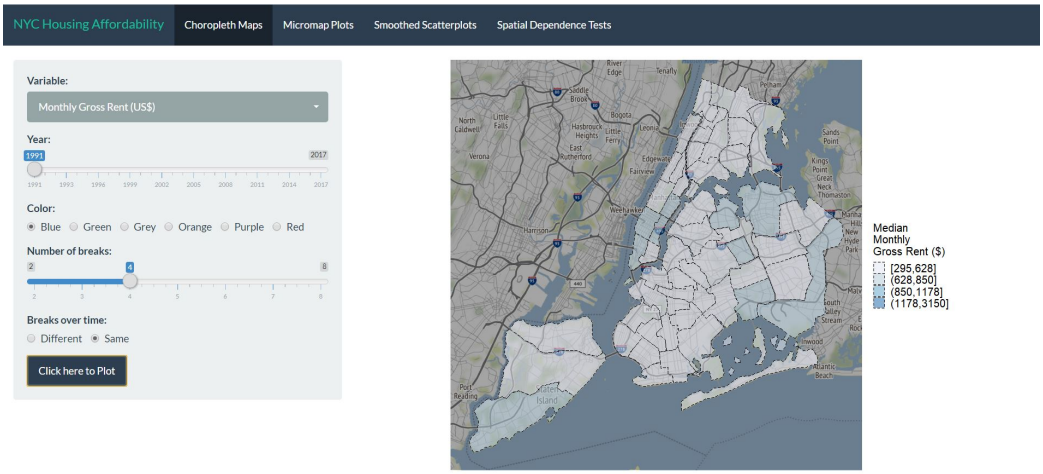

Fig. 5.2: Shiny R overlayed choropleth map interface. There are five selection features on the left side, and the map plot on the right. This is a choropleth map for monthly rent for the 1991 NYCHVS survey. It was shaded in different tones of blue and broken in quartiles, with about 13-14 sub-boroughs in each quartile. The breaks are calculated based on the information of all years. We can see that in the year 1991 the monthly rent was the lowest in the last 26 years, with sub-boroughs median values between $295 and $850

- Any of the last ten NYCHVS surveys that took place in the years 1991, 1993, 1996, 1999, 2002, 2005, 2008, 2011, 2014, and 2017.
- A selection of six colors schemes that includes blue, green, grey, orange, purple, and red. The selected color scheme will be used to shade the choropleth map. Colors schemes are based on the work of Brewer et al. (2003) and can be accessed at https://colorbrewer2.org/.
- A feature to calculate the amount of intervals or breaks to be considered in the choropleth map. Options include breaking the data by median (2), tercile (3), quantile (4), quintile (5), sextile (6), septile (7), or octile (8).
- An option to define if the breaks will be calculated based on the data of a particular year (different breaks over time) or on the data of all ten surveys (same breaks over time).

5.2 Linked Micromap (LM) Plots

In this second menu option, the user can generate a LM plot to see the relationship of two or three variables at a time, like the ones shown in Figures 3.4, 3.8, and 3.12. Figure 5.3 shows an example of this menu option. There are four, potentially five, selection features in which the user can customize the LM plot. Those features include:

- Any of the last ten NYCHVS surveys that took place in the years 1991, 1993, 1996, 1999, 2002, 2005, 2008, 2011, 2014, and 2017.
- The number of panels to be included in the LM plot. Every panel will depict one variable. The user can either choose two (2) or three (3) panels to be displayed.
- A first variable that includes a selection of eight variables such as home ownership percentage (percent of tenants that own a home in a sub-borough), monthly rent (current US$), housing cost percentage, immigrant households percentage (percent of households with a householder whose birthplace is outside the U.S.), household annual income (current US$), householder sex (percent of female householders), and median householder age. See further details in Table 2.1. This variable will be displayed in the primary panel in the far left and the sub-boroughs will be ranked, in descending order, by this variable.

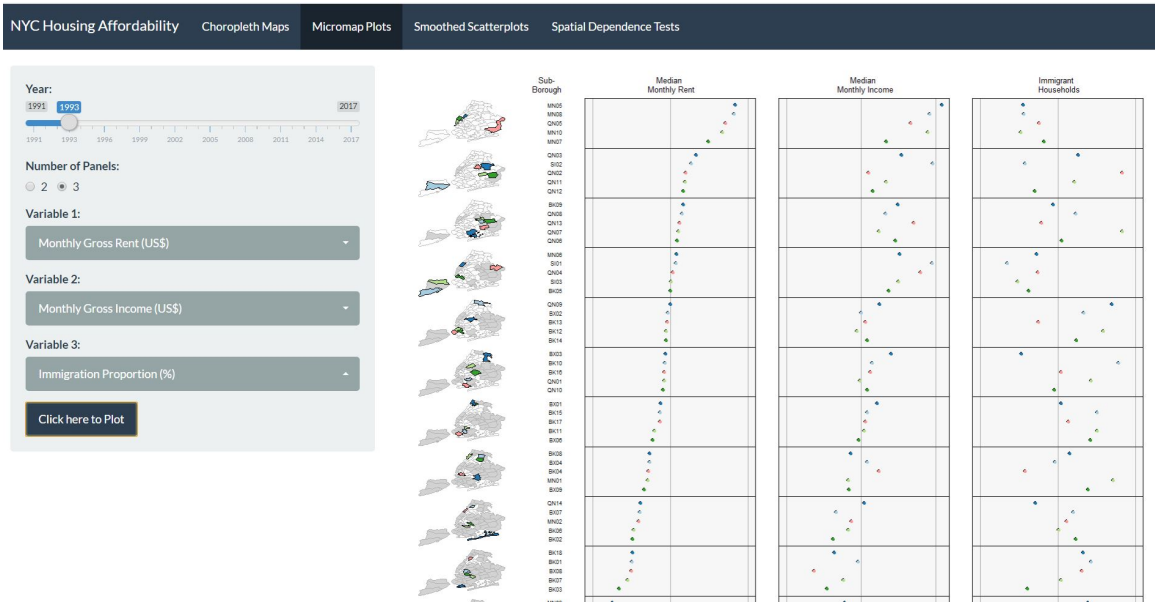

Fig. 5.3: Shiny R LM plot interface. There are four selection features, potentially five if the user selects three variables, on the left side, and the LM plot on the right. This is a LM plot with three panels for median monthly rent (primary panel in the far left), median monthly income (secondary panel in the middle), and immigrant households percentage (secondary panel in the far right) in the 1993 NYCHVS survey. We can see that four sub-boroughs in Manhattan and one in Queens have the highest median monthly rent. In addition, the median monthly rent and median monthly income variables are somewhat positively associated since they follow a similar pattern, which doesn't happen when we look at the relationship between rent and immigrant households

- A second variable that includes the same selection of eight variables as in the first variable. This variable will be displayed in a secondary panel next to the primary panel.
- A third variable that includes the same selection of eight variables as in the first variable. The option to select this variable will appear if the user selected the number of panels to be three in the second feature and will be displayed in a secondary panel in the far right.

5.3 Smoothed Scatterplots

In the third menu option, the user can plot a smoothed scatterplot for two variables, but also differentiate the information by up to two discrete variables, like in Figures 3.3, 3.7, and 3.11. Figure 5.4 shows an example of this menu option where we produced a scatterplot to show the relationship between household monthly income and rent, and differentiate each point by immigration status (shape) and borough (color). This menu option includes the following features.

- Any of the last ten NYCHVS surveys that took place in the years 1991, 1993, 1996, 1999, 2002, 2005, 2008, 2011, 2014, and 2017.

- A variable for the x-axis that includes a selection of eight variables such as householder age, household monthly rent (in current US$ and log_{10} current US$), house value (in current US$), mortgage interest rate percentage, household monthly income (in current US$ and log_{10} current US$), and housing cost percentage.

- A variable for the y-axis that include the same selection of variables as for the x-axis.

- A discrete variable represented by different colors that includes a selection of five variables such as householder sex, immigration status, home ownership status (owned or rented home), type of home ownership status (under mortgage or loan-free), and borough. User can also select "None" to omit this feature. This variable is also used to calculate a smoother.

- A discrete variable represented by different shapes that includes the same selection of variables as for the color variable. The user can also select "None" to omit this feature.

- An option to plot the entire data (Extended) or the information between the percentiles 5 and 95 (Middle 95%). Since the data have some extreme outliers, this option allow the user to customize the plot.

- An option to select the smooth method. Options available are "lm" (Linear Model), "glm" (Generalized Linear Model), "gam" (Generalized Additive Model), or "loess" (Locally Weighted Least Squares Regression). See Chapter 3 for further details.

Fig. 5.4: Shiny R smoothed scatterplot interface. There are seven features on the left side and the scatterplot on the right side. This is a smoothed scatterplot of Log household monthly rent and Log monthly income in the 2002 NYCHVS survey. The graph uses a color selection to depict the boroughs and a shape selection to depict the immigration status. The plot considers the middle 95% of the data and uses generalized linear methods (glm) as a smoother. The plot indicates the association between log household income and log household rent is positive and different among the main boroughs. While the Manhattan borough seems to report the highest amounts for both rent and income in the upper quantiles, the Bronx reports the lowest. However, in the lower quantiles, we can see the Manhattan borough also has the lowest values among all boroughs, which could be explained by their bimodal distribution discussed in Chapter 3.2

Fig. 5.5: Shiny R smoothed scatterplot interface. There are seven features on the left side and the scatterplot on the right side. This is a smoothed scatterplot of housing cost percentage and Log monthly income in the 1996 NYCHVS survey. The graph uses a color selection to depict the householder sex and a shape selection to depict the immigration status. The plot considers the middle 95% of the data and uses the loess method as a smoother. The plot indicates the association between housing cost percentage and log household income is, as expected, negative. We can see female householder reported a higher housing cost percentage and household income value than male householders in the data, but that difference is visually small. In addition, the smoother takes into account the non-linearity of the data to depict the relation

5.4 Spatial Autocorrelation Tests

In the fourth menu option, the user can run a hypothesis test using the Moran's I and Geary's C statistics to determine spatial autocorrelation, similar to the results shown in Figures 4.1, 4.2, 4.3, and 4.4. Figures 5.6 and 5.6 show an example of this menu option. There are five selection features, potentially six depending on the selected proximity method, in which the user can customize the spatial autocorrelation test. Those features include:

- A selection of eight variables such as home ownership percentage (percent of tenants that own a home in a sub-borough), monthly rent (current US$), housing cost percentage, immigrant households percentage (percent of households with a householder whose birthplace is outside the U.S.), household annual income (current US$), householder sex (percent of female householders), and median householder age. See further details in Table 2.1.

- Any of the last ten NYCHVS surveys that took place in the years 1991, 1993, 1996, 1999, 2002, 2005, 2008, 2011, 2014, and 2017.
- The statistic to be used i.e., either Moran's I or Geary's C.
- The alternative hypothesis to be used either greater, less, or two-sided. In this test, the null hypothesis represents H_0: "data does not show spatial autocorrelation". Possible alternative hypotheses are the one-sided alternatives H_a : "data is spatially clustered" and H_a : "data is spatially dispersed", and the two-sided alternative H_a : "data shows spatial autocorrelation". See Chapter 4.1.1 for further details.
- The proximity method to be used based on the weight matrix. Options are "Queen" contiguous neighbors, k-nearest neighbors, and maximum proximity distance.
- If k-nearest neighbor is selected, the user can input the number of neighbors as an integer between one (1) and 54, the number of sub-boroughs minus one.
- If maximum proximity distance is selected, user can type the maximum number of kms greater than 7.89km, the largest minimum distance between two sub-boroughs in NYC.

In addition, the results section in the right side include the following features.

- Number of observations. This number will always be 55 as it includes the number of NYC sub-boroughs.
- Sample average of the variable selected.
- Standard deviation of the variable selected.
- Null hypothesis. Data is not spatially autocorrelated.
- Alternative hypothesis. Data is spatially autocorrelated.
- Test Statistic. Numerical value of the Moran's I or Geary's C statistic.
- P-value of the hypothesis test.

- Conclusion, either we reject or fail to reject the null hypothesis.
- Brief interpretation of the obtained results.

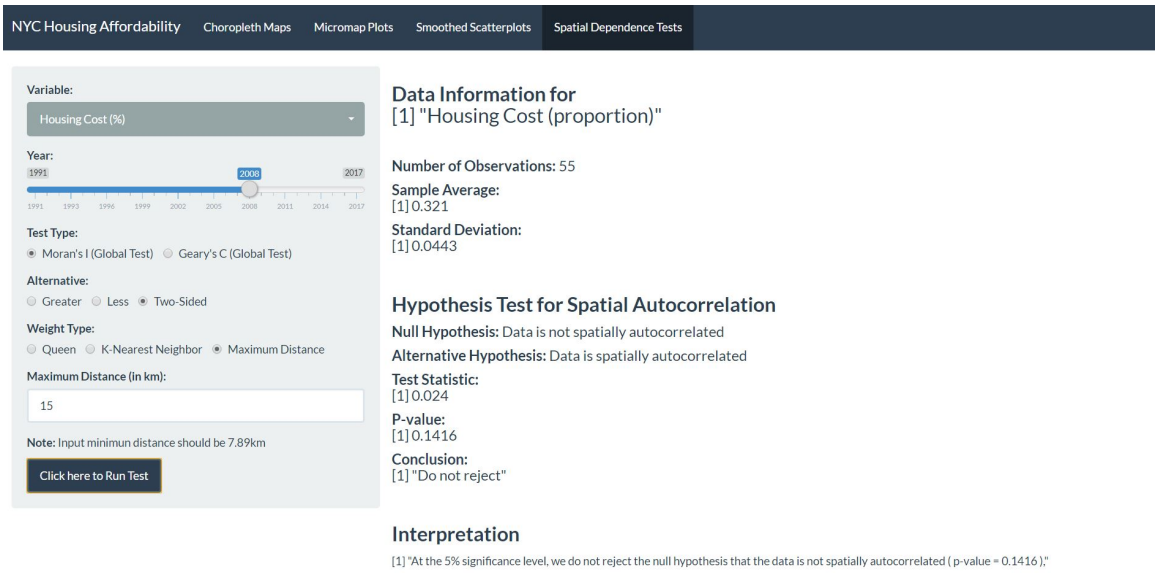

Fig. 5.6: Shiny R spatial autocorrelation test interface. There are five selection features, potentially six if the user selects the k-nearest neighbor or maximum distance weight type, on the left side, and the data summary and test results on the right side. This is a spatial autocorrelation test of the housing cost percentage in the 2008 NYCHVS survey that uses the Moran's I statistic. For the weight W matrix we selected the maximum distance proximity method, and specified 15km as the maximum distance. The results in the right indicate that we fail to reject the null hypothesis of spatial autocorrelation in the data

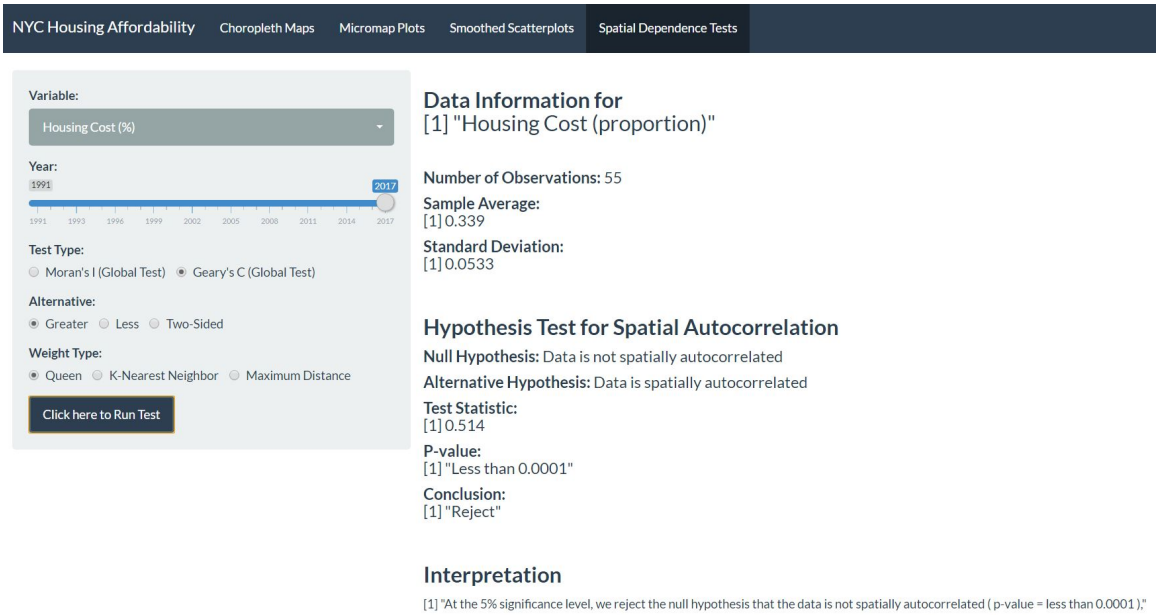

Fig. 5.7: Shiny R spatial autocorrelation test interface. There are five selection features, potentially six if the user selects the k-nearest neighbor or maximum distance weight type, on the left side, and the data summary and test results on the right side. This is a spatial autocorrelation test of the housing ownership percentage in the 2017 NYCHVS survey that uses the Geary's C statistic. For the weight W matrix we selected the the "Queen" proximity method. The results in the right indicate that we reject the null hypothesis of no spatial autocorrelation in the data, which matches the choropleth map shown in Figure 3.1

CHAPTER 6

Conclusions and Further Research

This thesis provided exploratory analyses of housing variables in New York City. We analyzed the five different boroughs and the 55 sub-boroughs, and considered the residents' immigration status. We selected home ownership percentage, home renting, and housing cost percentage as the housing variables of interest. All these housing and immigration variables have reported spatial autocorrelation throughout the years, which were addressed in our data analysis.

In the case of home ownership percentage, we didn't see any clear relationship between home ownership percentage or households with mortgage among NYC sub-boroughs in Section 3.1. In addition, mortgage rates were independent from immigration status. In some boroughs, such as Manhattan and the Bronx, being an immigrant household has a negative association with owning a home, and even immigrant households have less access to a mortgage. When running the SAR models, we also noticed that sociodemographic and housing variables, including immigration, where relevant to explain home ownership percentage among NYC boroughs.

New York City, as a city where renting a home is more popular than owning one, shows a moderate association between household rent and income in all NYC boroughs. In contrast, the association of immigration status, represented by the percentage of immigrants, with household rent or income is weak in most boroughs. The only scenario where immigration shows some stronger association with household rent is in the high-income sub-boroughs in Manhattan and Brooklyn, which represent the upper quantiles of median household rent. When running the SAR models, we noticed immigration and housing variables weren't significant to explain home renting as they were when looking at home ownership percentage.

Housing cost percentage also provided some general insights. Immigrant households reported a moderate correlation with housing cost percentage at an aggregated level, specially in the Bronx borough. Unfortunately, most boroughs reported a high variability which came with a low correlation with rent, income, and immigration status when looking at the raw data. Although we got some

results in the SAR models where we notice that being an immigrant may explain having a higher housing cost percentage, the variability in the data seems to suggest we may need to run a more specific analysis, for example implementing models by specific area or income group, to get more conclusive results.

Finally, the spatial autocorrelation in housing and immigration variables is positively strong and significant at the 5% significance level for most years in all four variables (home ownership percentage, median household rent, housing cost percentage, and immigrant household percentage). This matches with what can be seen in the choropleth maps in Figures 3.1, 3.5, and 3.9 that show some strong clustering.

Suggestions for further research include modeling these housing variables based on specific income and immigration groups at a sub-borough level as this may address the variability in the data. Developing more sophisticated spatial autoregressive methods may help identify patterns within certain groups. In this thesis, we used SAR models, but there are other methods, such as spatial autoregressive geographically weighted regression (GWR-SAR) developed by Geniaux and Martinetti (2018), to be explored. Regarding the Shiny R app, including a spatial autoregressive models feature as well as more variables could help answering more questions regarding housing and related variables. Some examples of these questions, listed in no particular order, are:

- Does the housing data report spatial autocorrelation in 2005?
- Does the relationship between household rent and income differ among immigration groups in 1991?
- Has the Bronx borough reported the highest housing cost percentage in the last 26 years?

APPENDICES

APPENDIX A

Four-year Comparison Figures

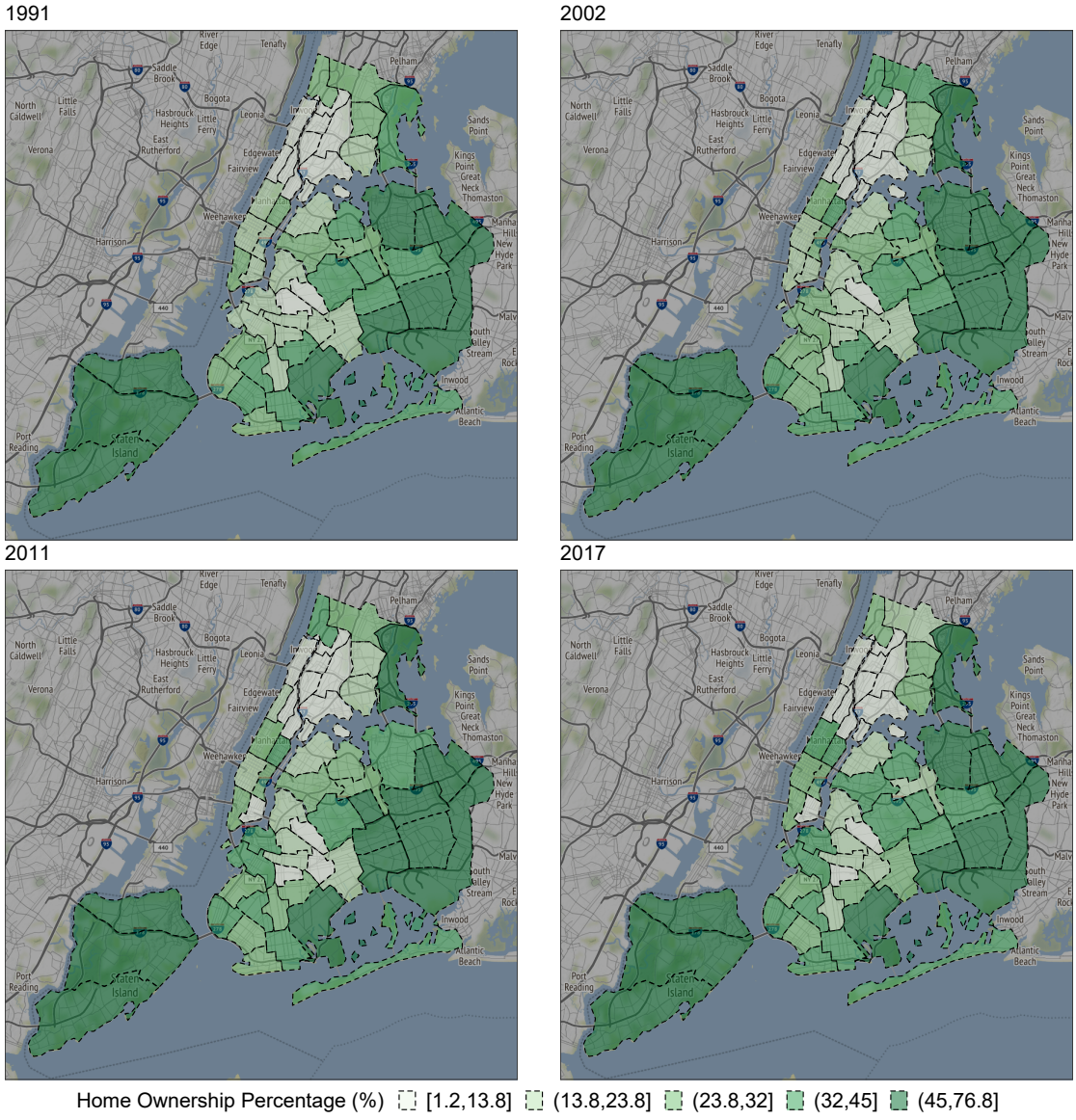

Fig. A.1: NYC home ownership percentage choropleth map for the years 1991, 2002, 2011, and 2017. The data are broken into quintiles, and there is a common scale for all four maps. About 44 sub-boroughs across the four years fall into each of the five quintiles. The geographic distribution of sub-boroughs with relatively low and relatively high home ownership percentages did not change much over the past 26 years

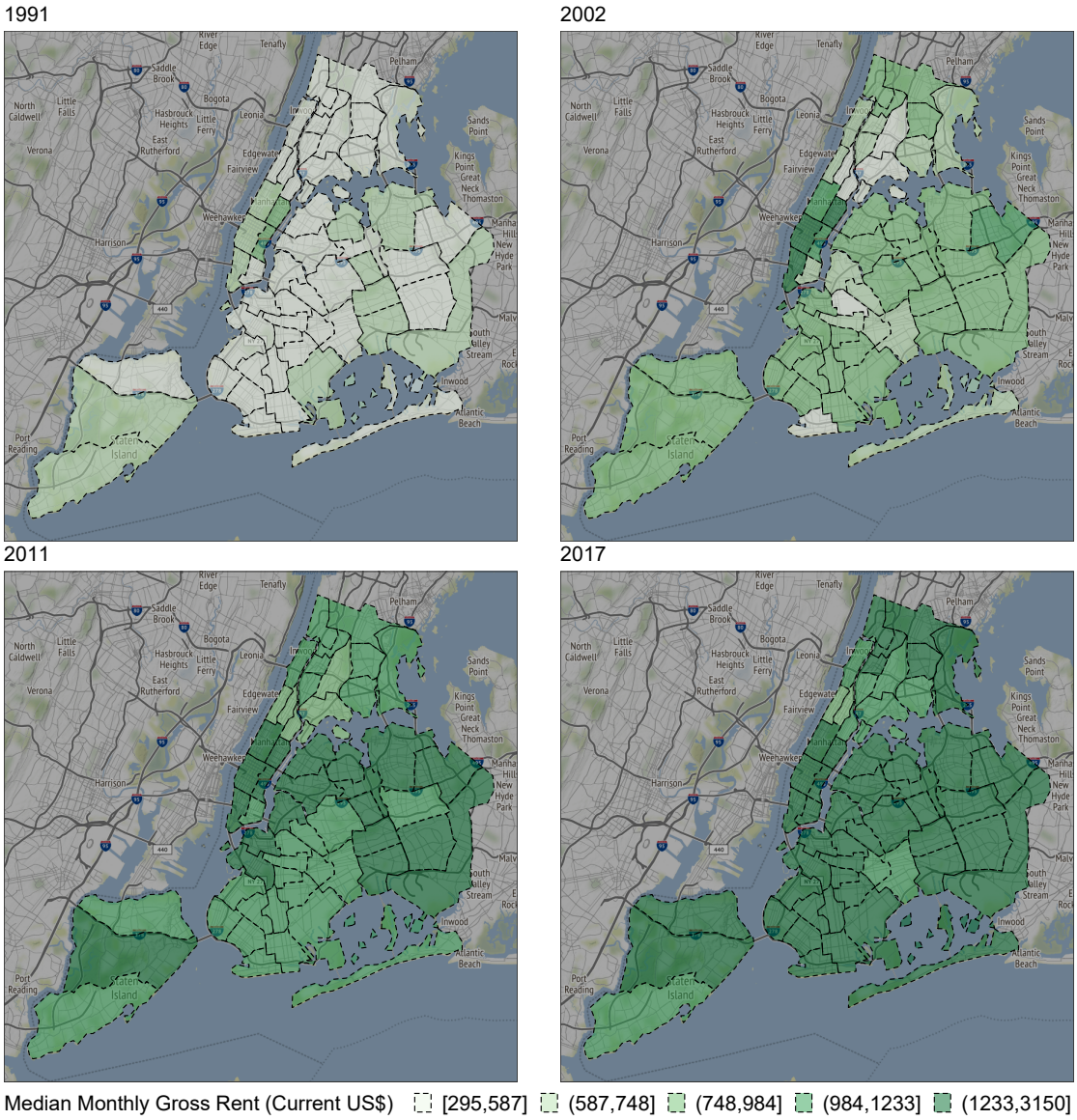

Fig. A.2: NYC household median monthly rent choropleth map for the years 1991, 2002, 2011, and 2017. The data are broken into quintiles, and there is a common scale for all four maps. About 44 sub-boroughs across the four years fall into each of the five quintiles. This raw US$ amounts have not been adjusted by inflation

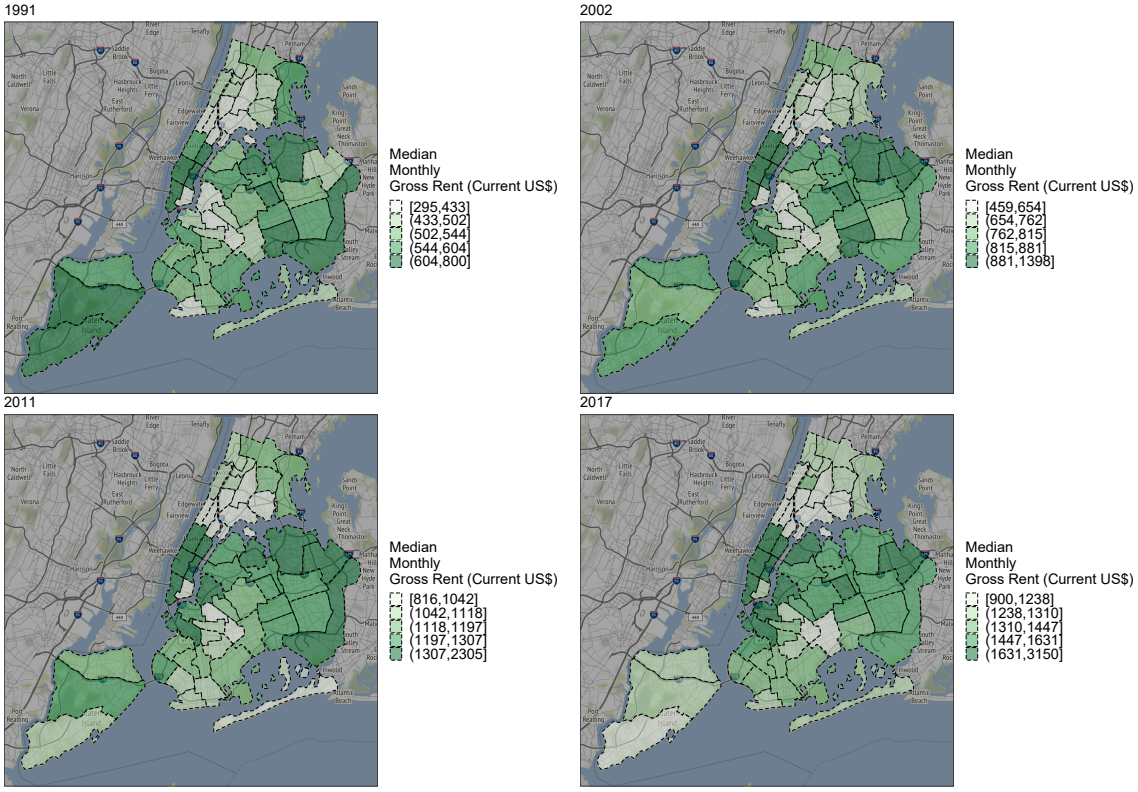

Fig. A.3: NYC household median monthly rent choropleth map for the years 1991, 2002, 2011, and 2017. The data are broken into quintiles and the scale is different in each map, with about 11 sub-boroughs in each of the five quintiles.We notice Staten Island has a high rent in 1991, but a low rent in 2017 because the rent increase has been lower compared to all other boroughs

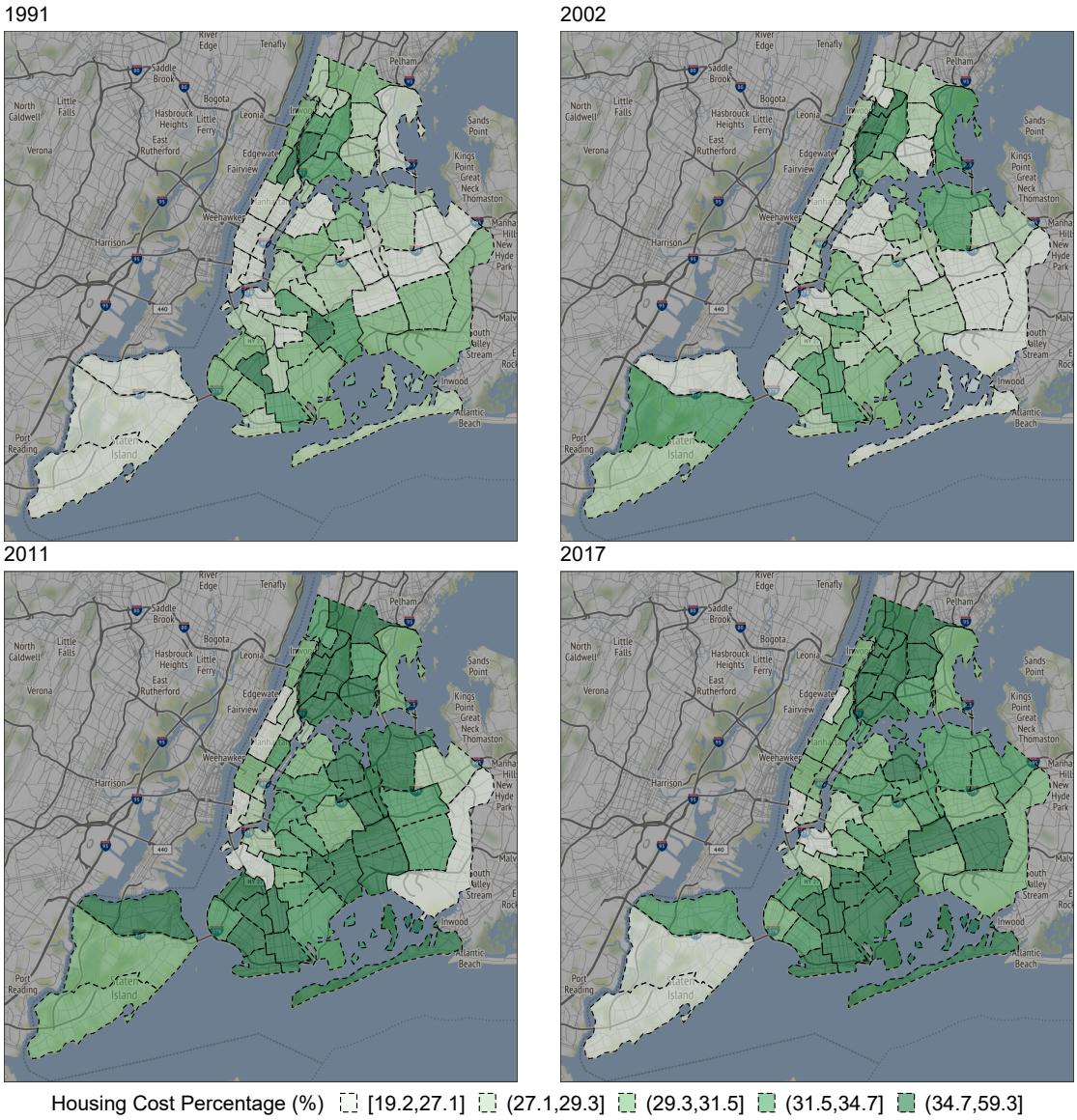

Fig. A.4: NYC housing cost percentage percentage choropleth map for the years 1991, 2002, 2011, and 2017. The data are broken into quintiles, and there is a common scale for all four maps. About 44 sub-boroughs across the four years fall into each of the five quintiles. Most NYC sub-boroughs show a noticeable increase in housing cost percentage

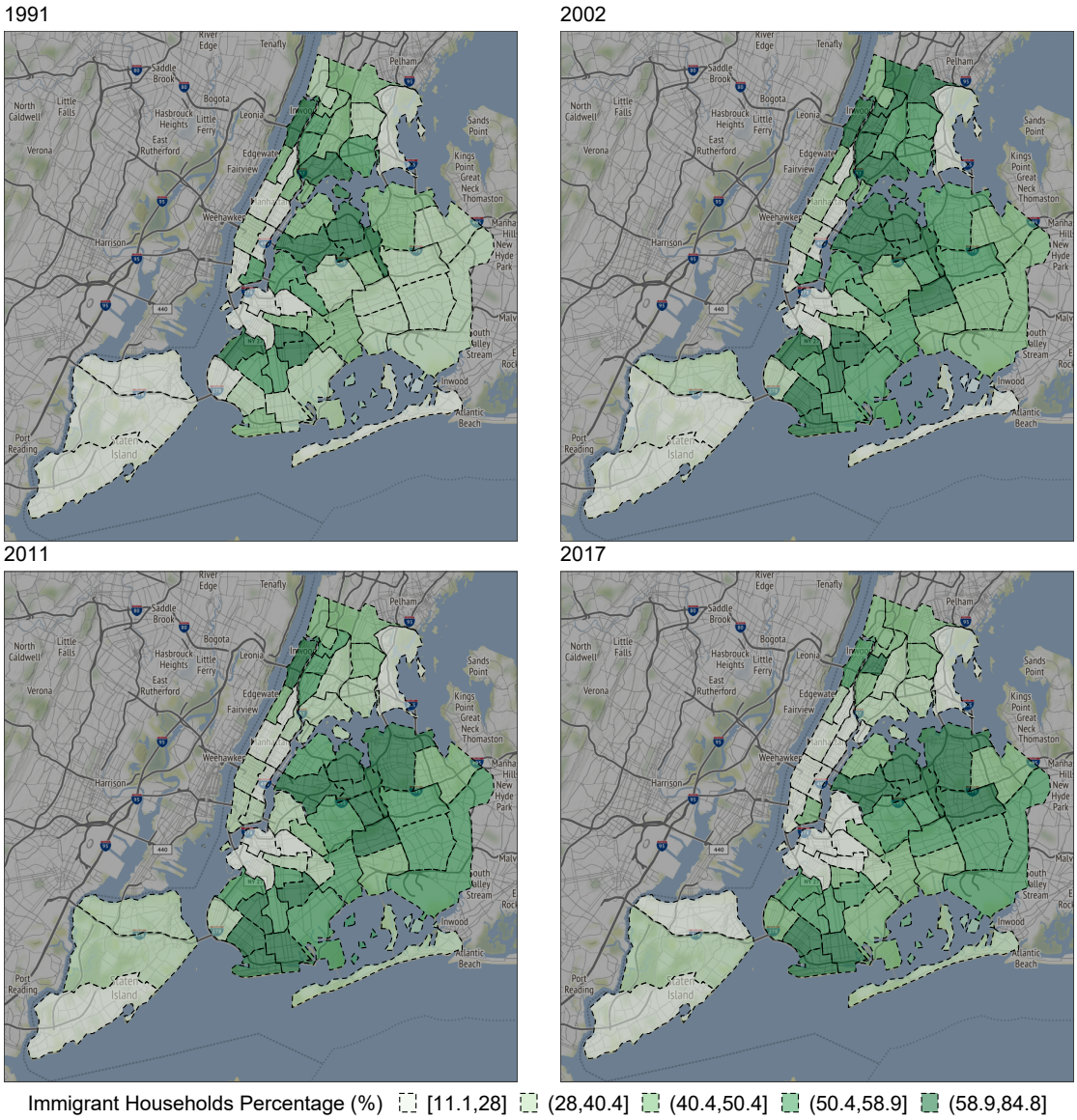

Fig. A.5: NYC immigrant households percentage choropleth map for the years 1991, 2002, 2011, and 2017. The data are broken into quintiles, and there is a common scale for all four maps. About 44 sub-boroughs across the four years fall into each of the five quintiles. The percent of immigrant households has increased throughout the years in NYC, specially in the Queens borough

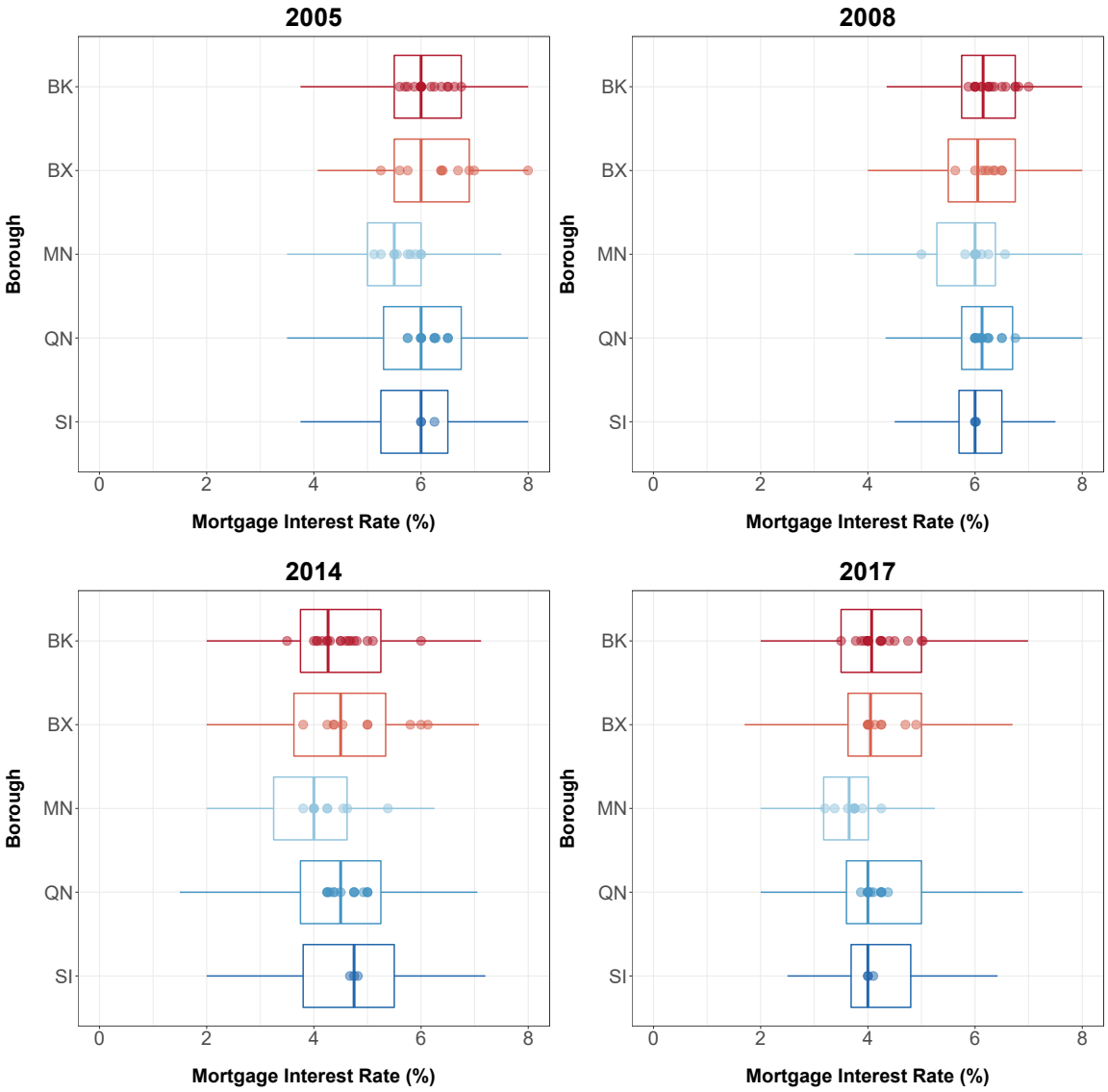

Fig. A.6: NYC mortgage interest rates by borough for the years 2005, 2008, 2014, and 2017. The dots represent the aggregated data of the 55 sub-boroughs, and the boxplot is constructed based on the raw survey data. This variable started to be included in the 2005 survey. We can see an overall reduction in mortgage interest rates in all boroughs

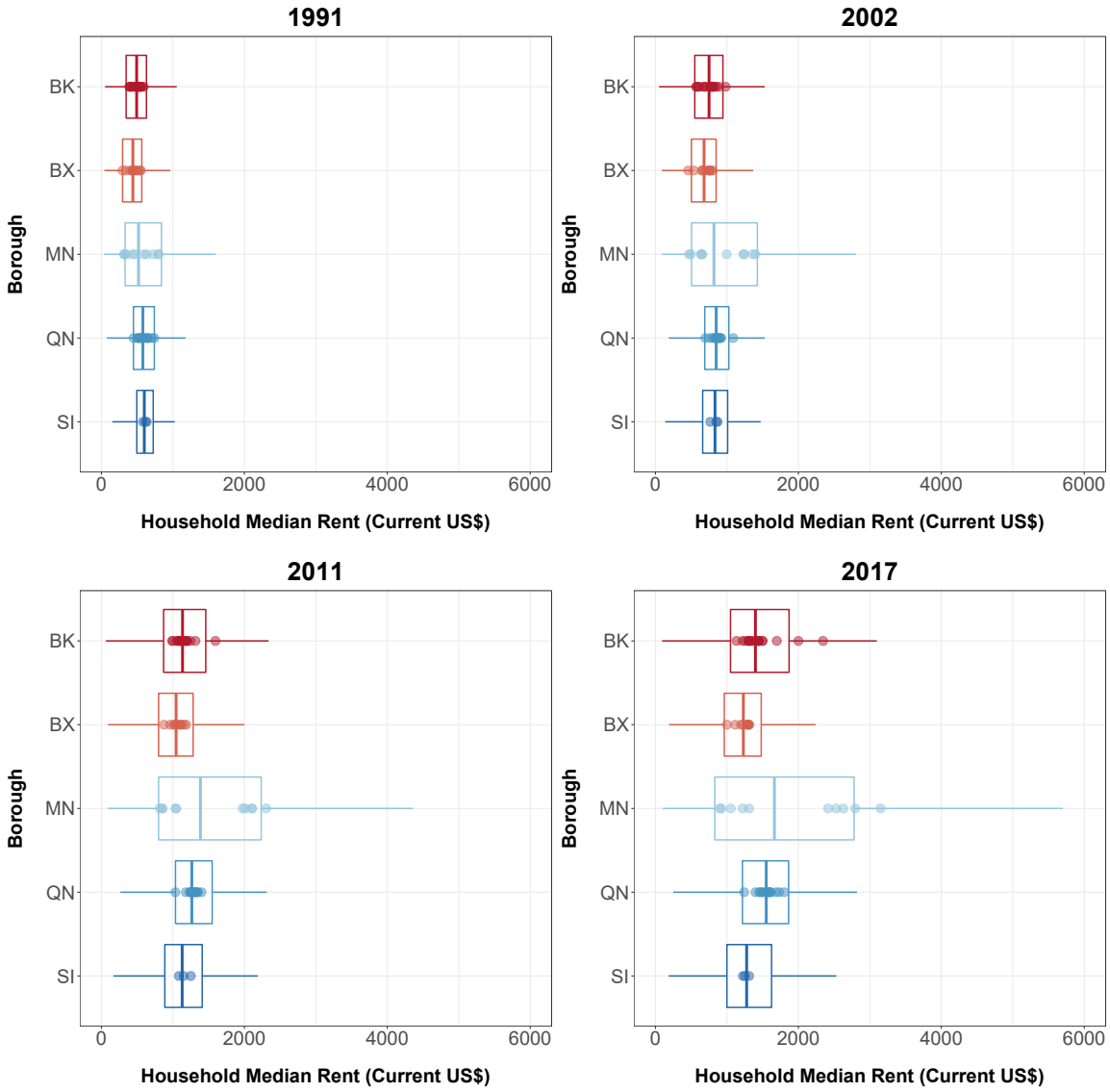

Fig. A.7: NYC monthly rent (current US$) by borough for the years 1991, 2002, 2011, and 2017. The dots represent the aggregated data of the 55 sub-boroughs, and the boxplot is constructed based on the raw survey data. The increase in the rent and its spread include the increase in the Consumer Price Index, where Manhattan shows the biggest increase

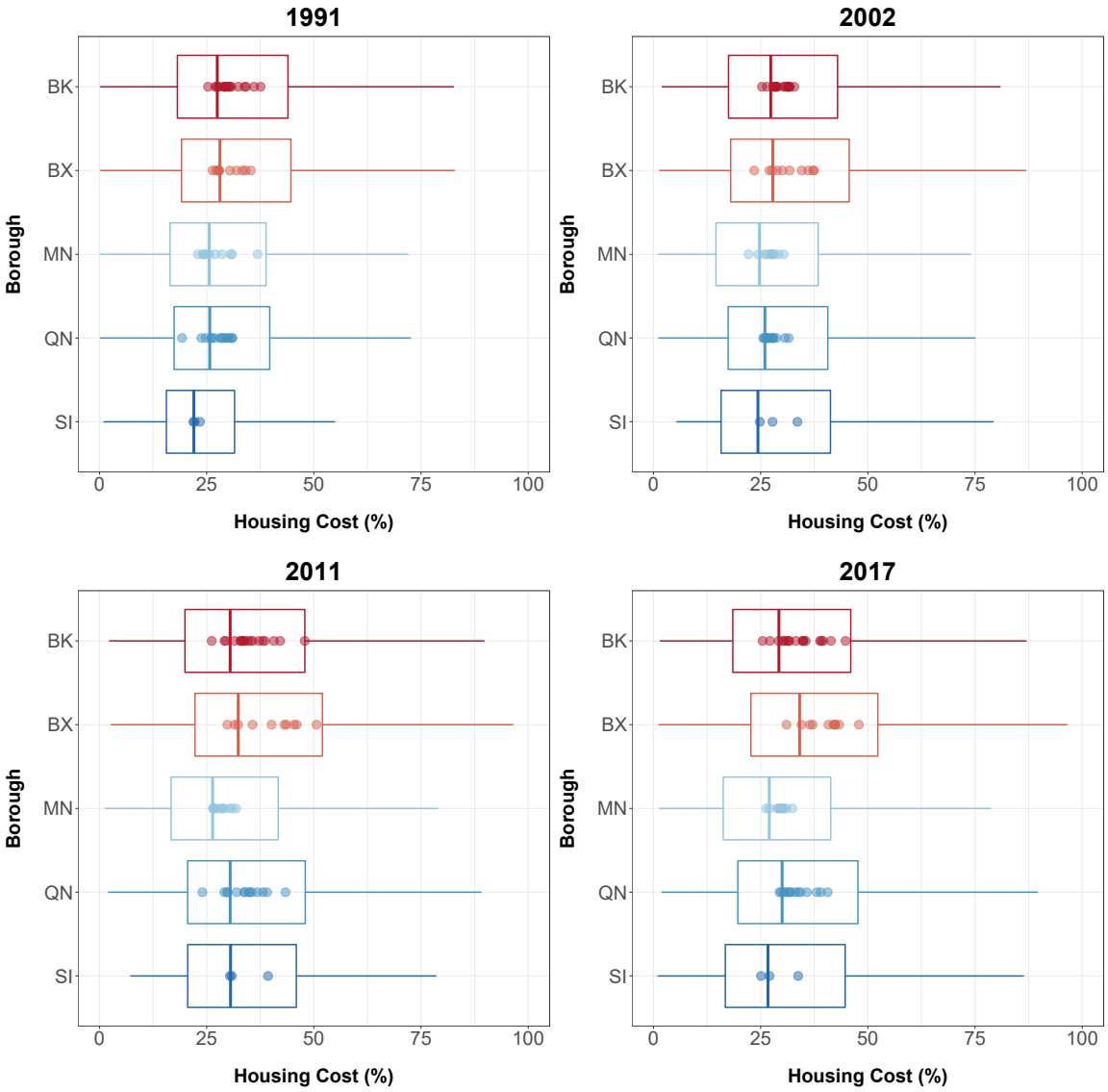

Fig. A.8: NYC housing cost percentage by borough for the years 1991, 2002, 2011, and 2017. The dots represent the aggregated data of the 55 sub-boroughs, and the boxplot is constructed based on the raw survey data. We can see an overall increase in the housing cost percentage in all boroughs

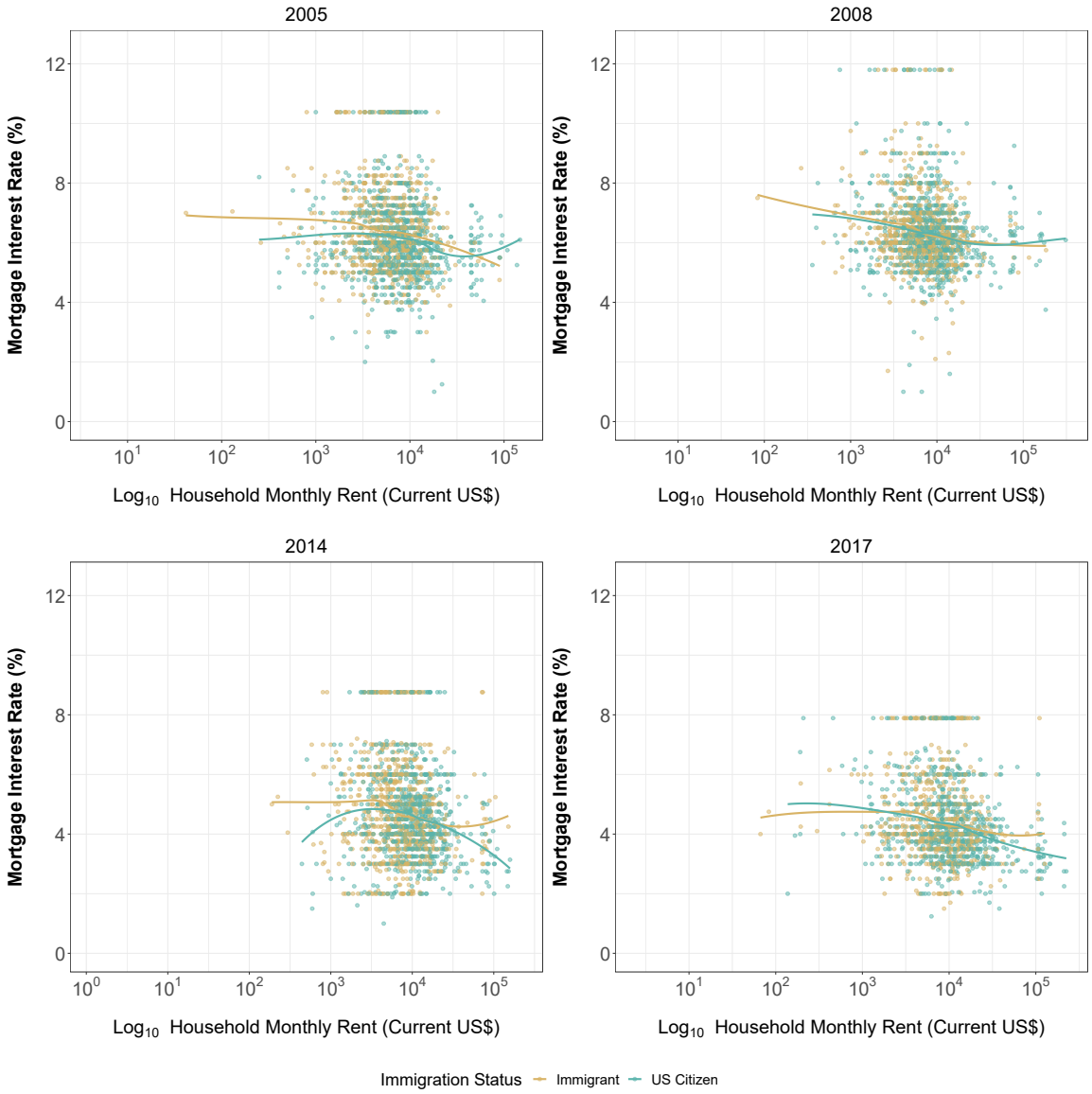

Fig. A.9: Monthly income and mortgage interest rates scatterplot for the years 2005, 2008, 2014, and 2017. This variable started to be included in the 2005 survey. We can notice difference in upper and lower income quantiles, specially in 2014 and 2017; high income immigrants report a slightly higher mortgage rate than US citizens

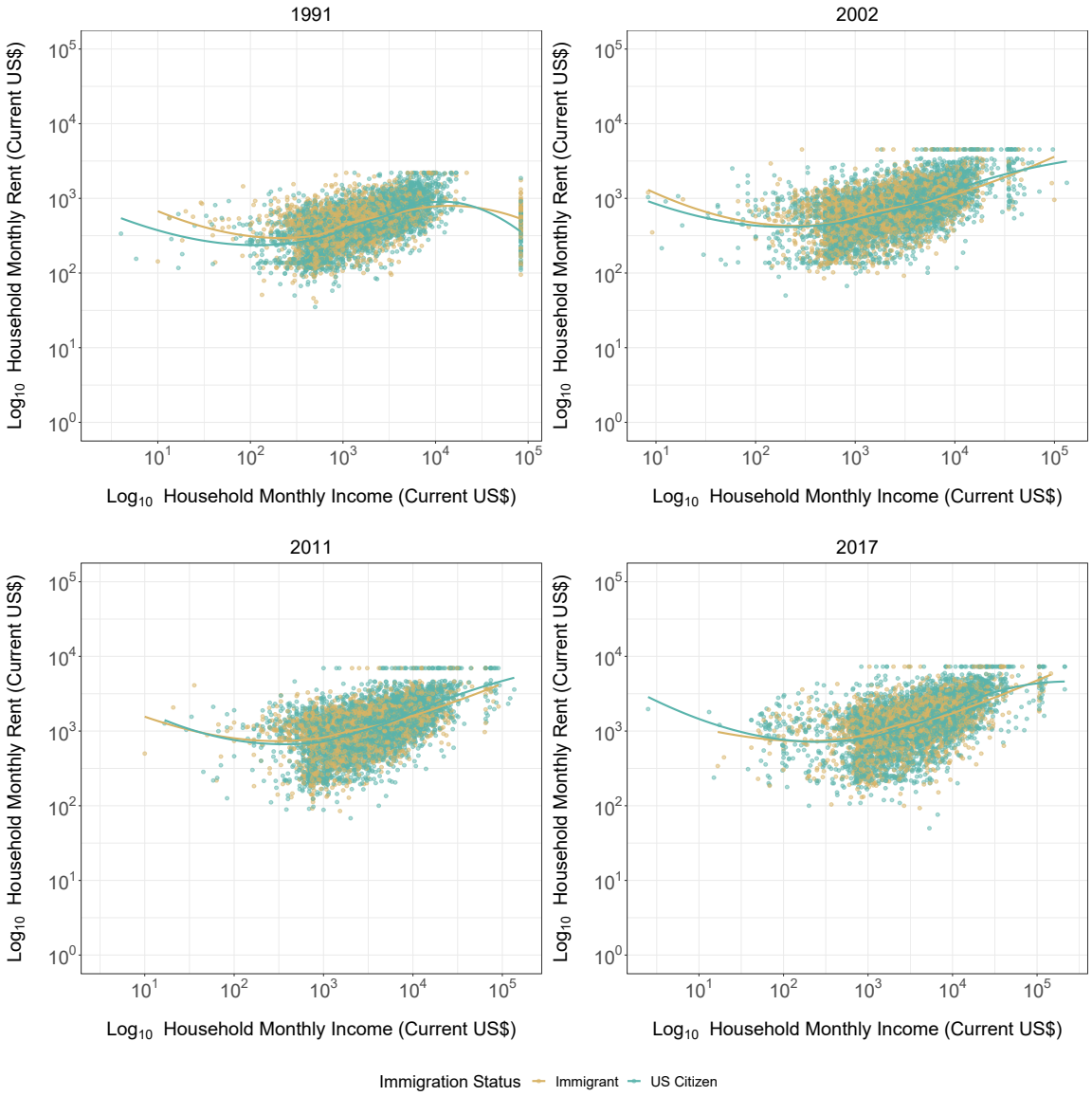

Fig. A.10: NYC household monthly income and rent scatterplot for the years 1991, 2002, 2011, and 2017. Part of the increase in the rent and its spread includes the increase in the Consumer Price Index. Graphs don't show any noticeable difference among immigration groups

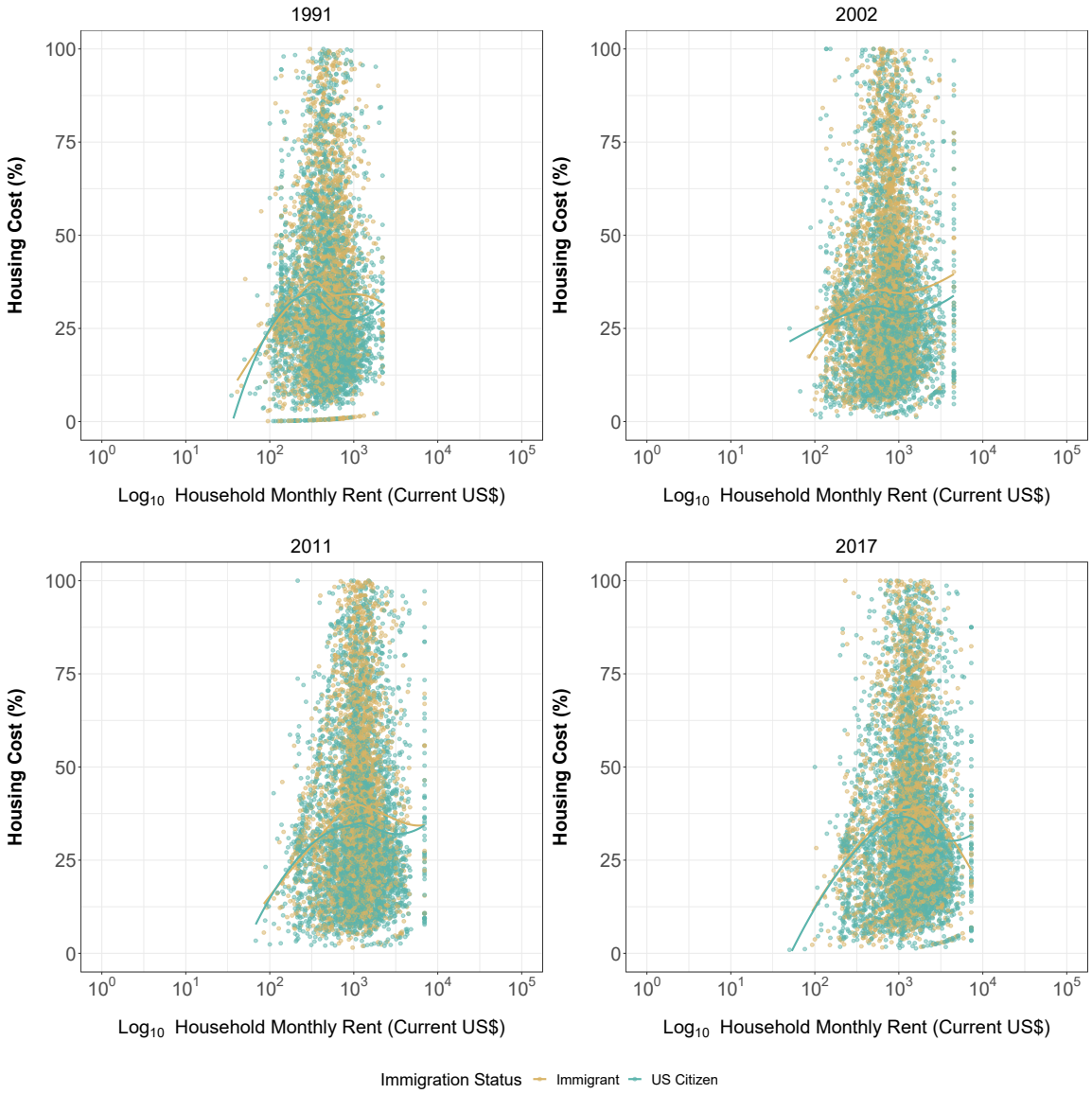

Fig. A.11: NYC household housing cost percentage and monthly rent scatterplot for the years 1991, 2002, 2011, and 2017. Between 8 and 13% of the data represent cases where the housing cost percentage is greater than 100%. These cases are not shown in the scatterplot. The NYCHVS survey had a maximum listed amount for monthly rent, and it assigned a certain value to households that listed a monthly rent above that listed maximum depending on the year. Even though there is high variability in the data, middle rent immigrants have higher housing cost percentage in every plot

APPENDIX B

Neighborhood Matrices

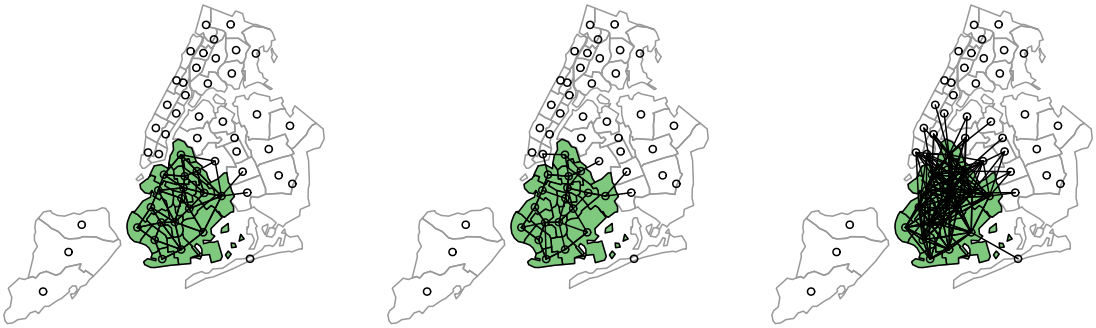

Fig. B.1: Graphical representation of Brooklyn neighborhood matrices. The left plot refers to the “Queen” method, the middle plot refers to the three nearest neighbor method, and the right plot refers to the ”Maximum Distance” method with 8 km. White dots represent the centroids of each sub-borough

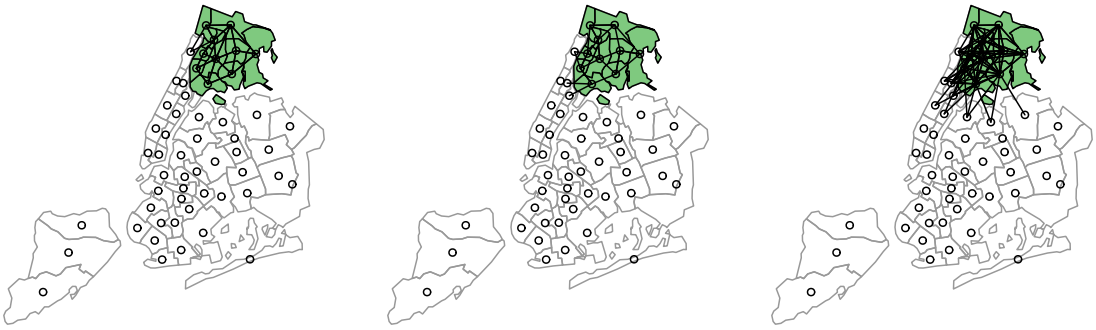

Fig. B.2: Graphical representation of the Bronx neighborhood matrices. The left plot refers to the "Queen" method, the middle plot refers to the three nearest neighbor method, and the right plot refers to the "Maximum Distance" method with 8 km. White dots represent the centroids of each sub-borough

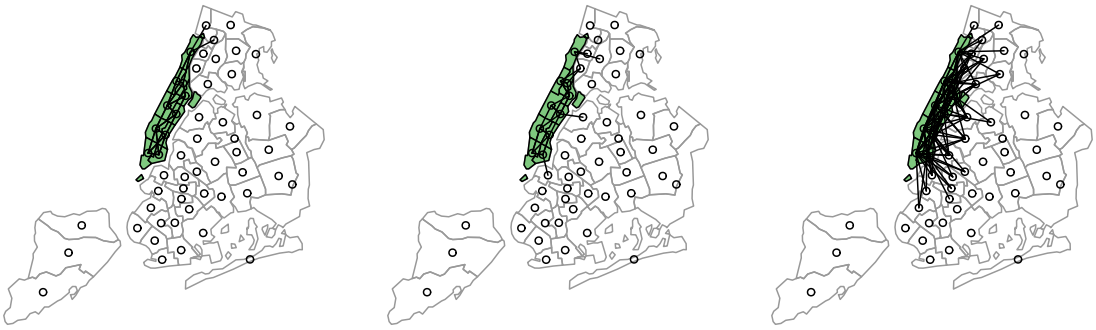

Fig. B.3: Graphical representation of Manhattan neighborhood matrices. The left plot refers to the "Queen" method, the middle plot refers to the three nearest neighbor method, and the right plot refers to the "Maximum Distance" method with 8 km. White dots represent the centroids of each sub-borough

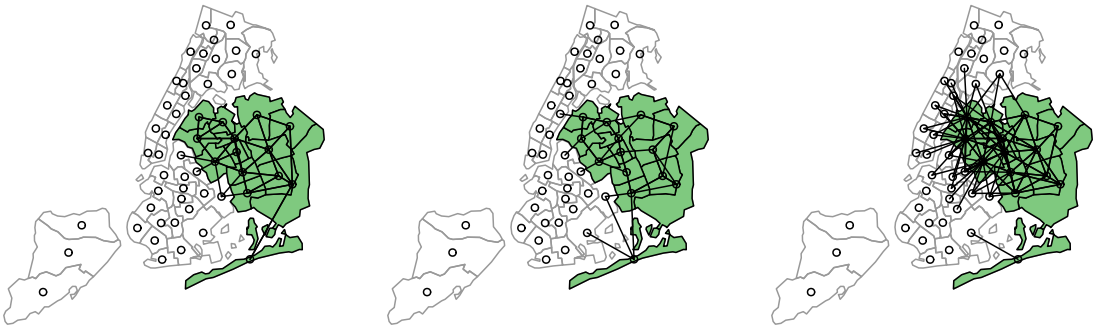

Fig. B.4: Graphical representation of Queens neighborhood matrices. The left plot refers to the "Queen" method, the middle plot refers to the three nearest neighbor method, and the right plot refers to the "Maximum Distance" method with 8 km. White dots represent the centroids of each sub-borough

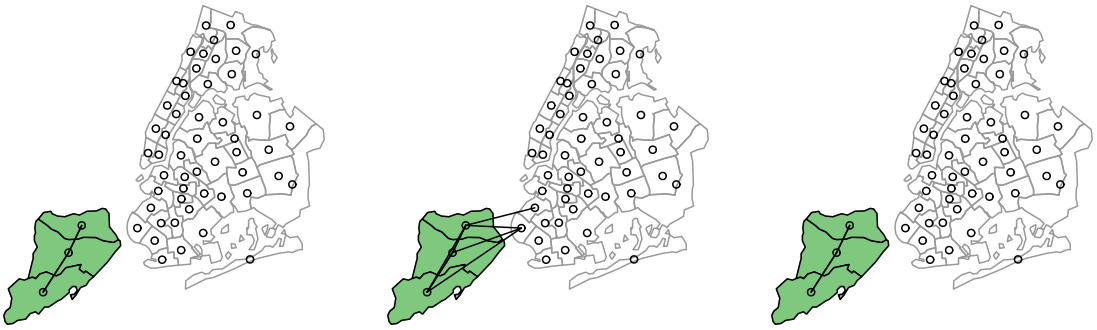

Fig. B.5: Graphical representation of Staten Island neighborhood matrices. The left plot refers to the "Queen" method, the middle plot refers to the three nearest neighbor method, and the right plot refers to the "Maximum Distance" method with 8 km. White dots represent the centroids of each sub-borough

APPENDIX C

Supplementary Spatial Autocorrelation Results

In the following figures, the top left graph shows the values of Moran's I statistic. The top right graph shows the values of Geary's C statistic. The bottom graph shows the p-values obtained from the hypothesis tests for spatial randomness using both statistics.

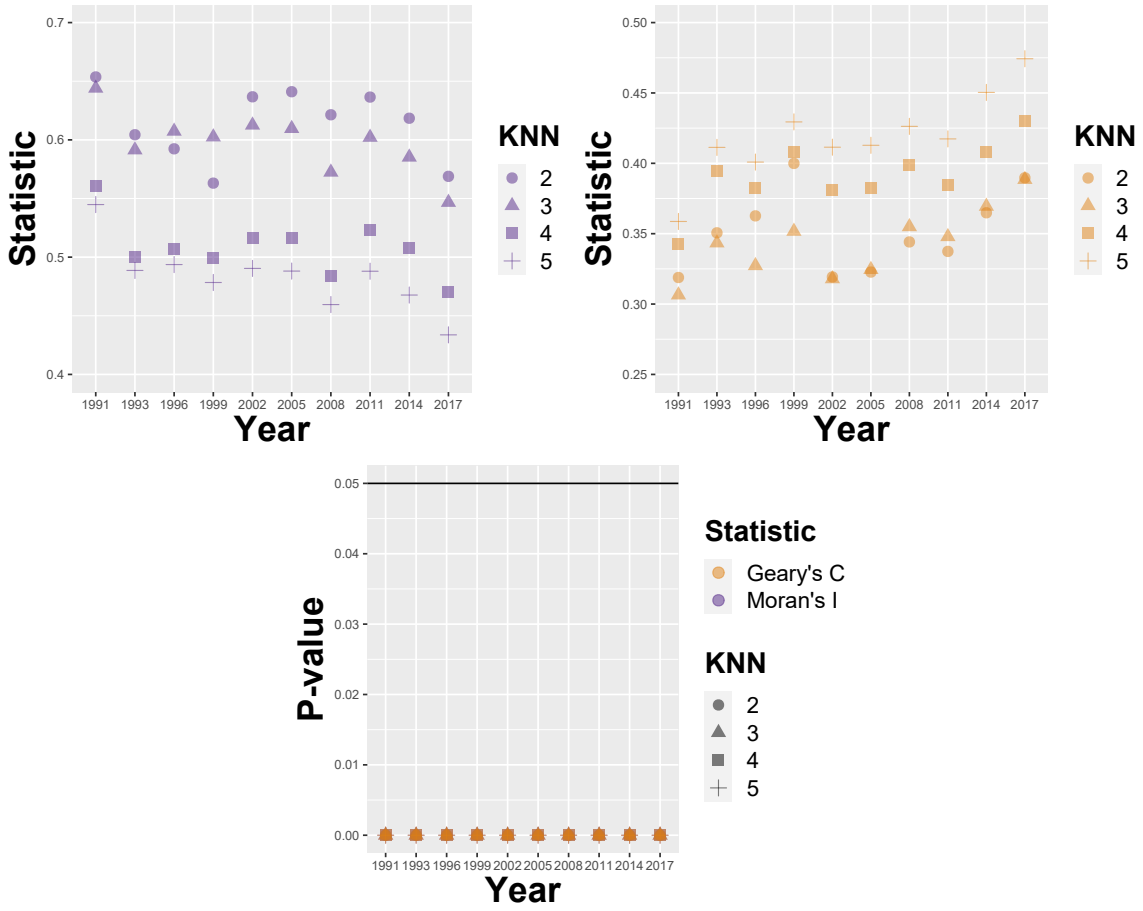

Fig. C.1: NYC home ownership percentage spatial autocorrelation tests 1991 - 2017 for the k-nearest neighbor method

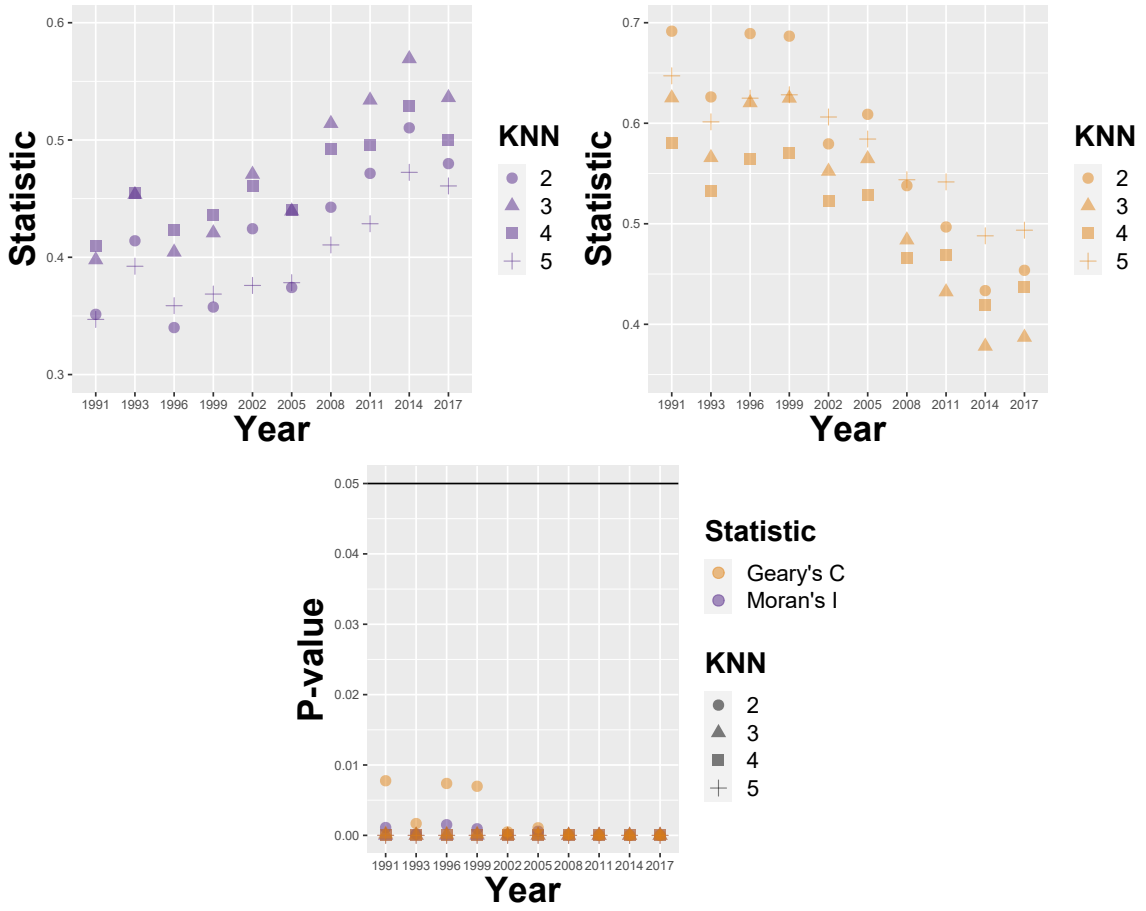

Fig. C.2: NYC median household rent spatial autocorrelation tests 1991 - 2017 for the k-nearest neighbor method

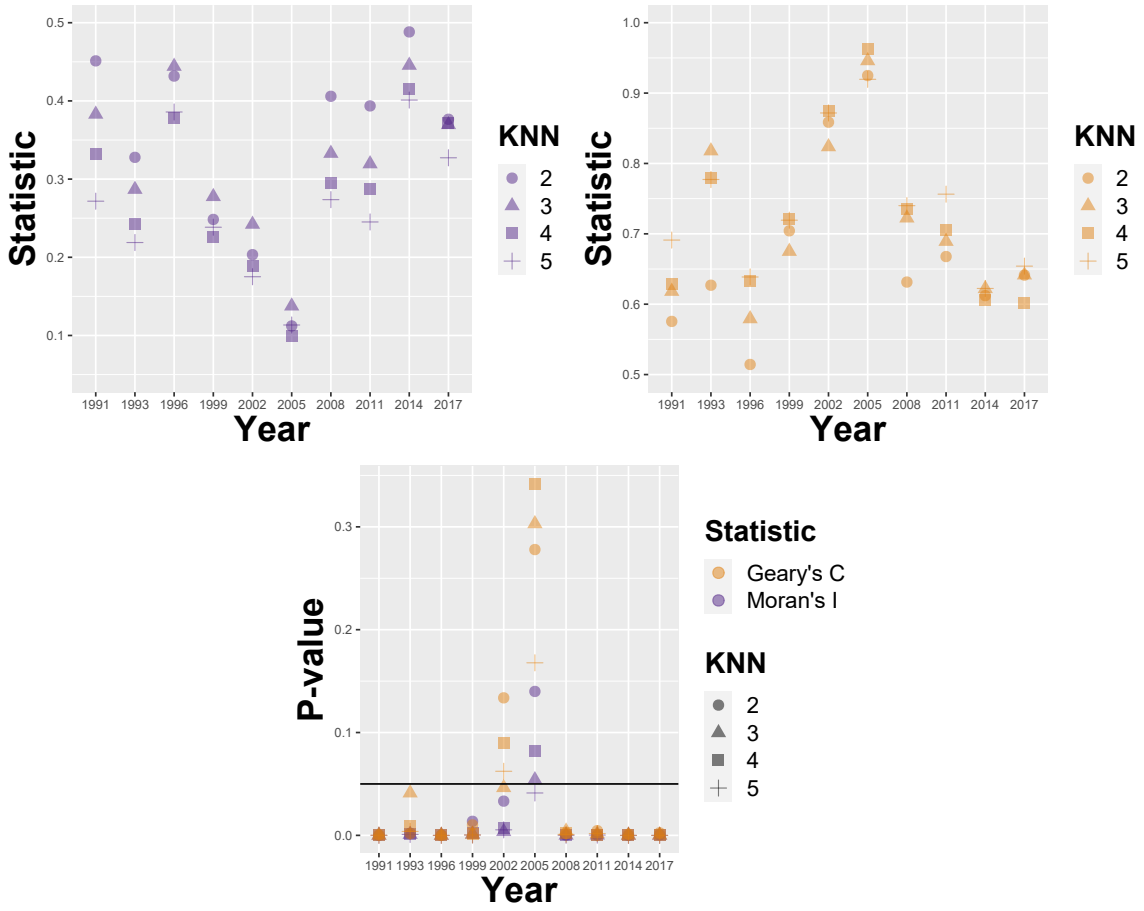

Fig. C.3: NYC housing cost percentage spatial autocorrelation tests 1991 - 2017 for the k-nearest neighbor method

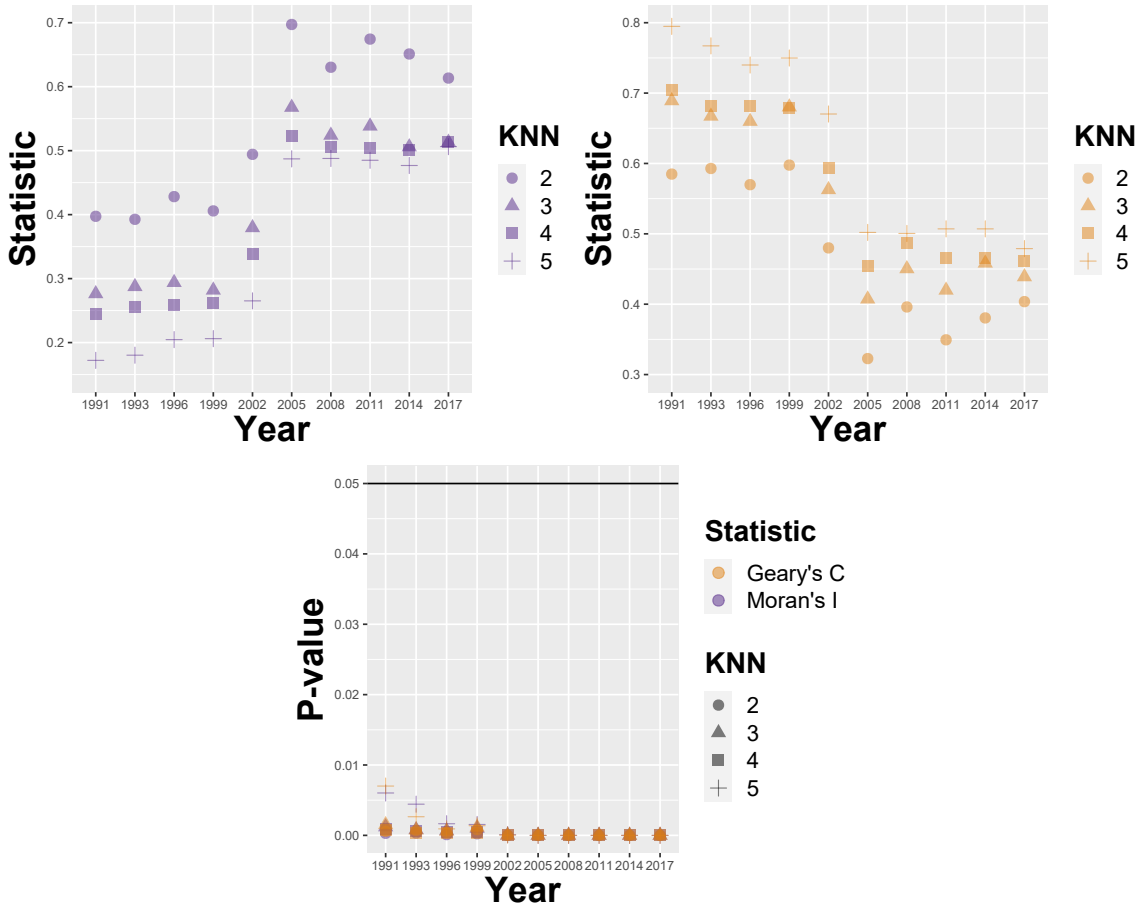

Fig. C.4: NYC immigrant household percentage spatial autocorrelation tests 1991 - 2017 for the k-nearest neighbor method

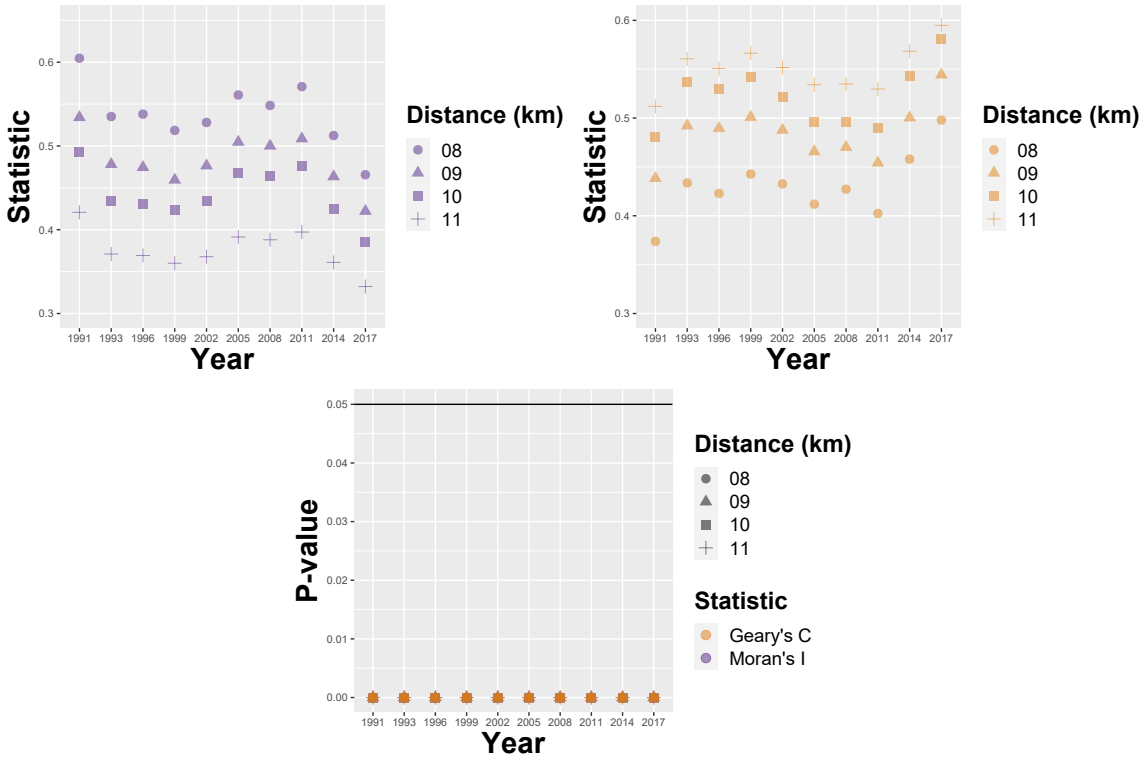

Fig. C.5: NYC home ownership percentage spatial autocorrelation tests 1991 - 2017 for the maximum distance proximity method

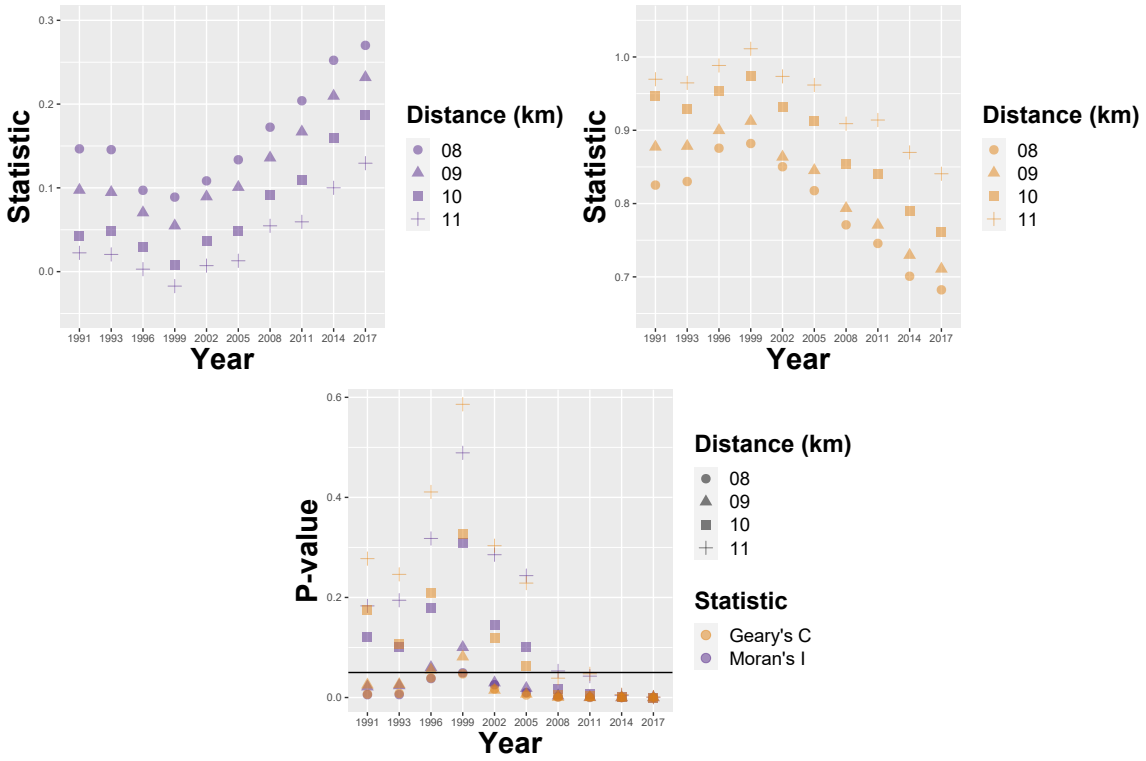

Fig. C.6: NYC median household rent spatial autocorrelation tests 1991 - 2017 for the maximum distance proximity method

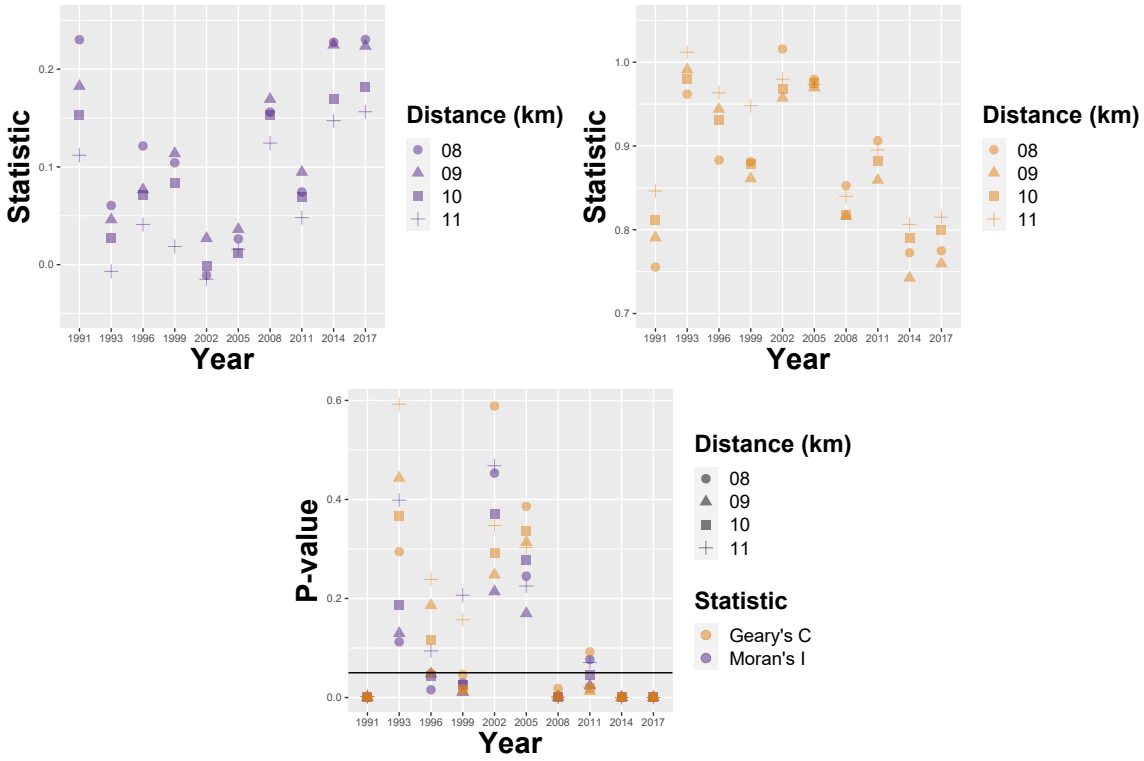

Fig. C.7: NYC housing cost percentage spatial autocorrelation tests 1991 - 2017 for the maximum distance proximity method

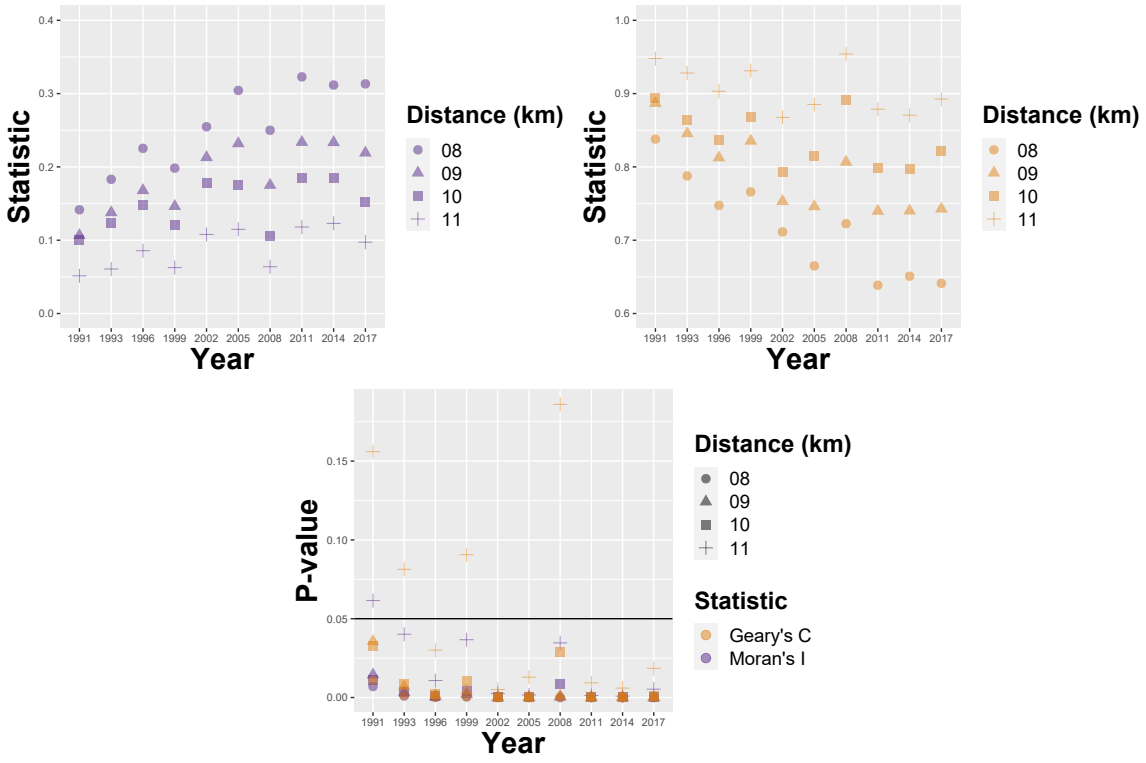

Fig. C.8: NYC immigrant household percentage spatial autocorrelation tests 1991 - 2017 for the maximum distance proximity method

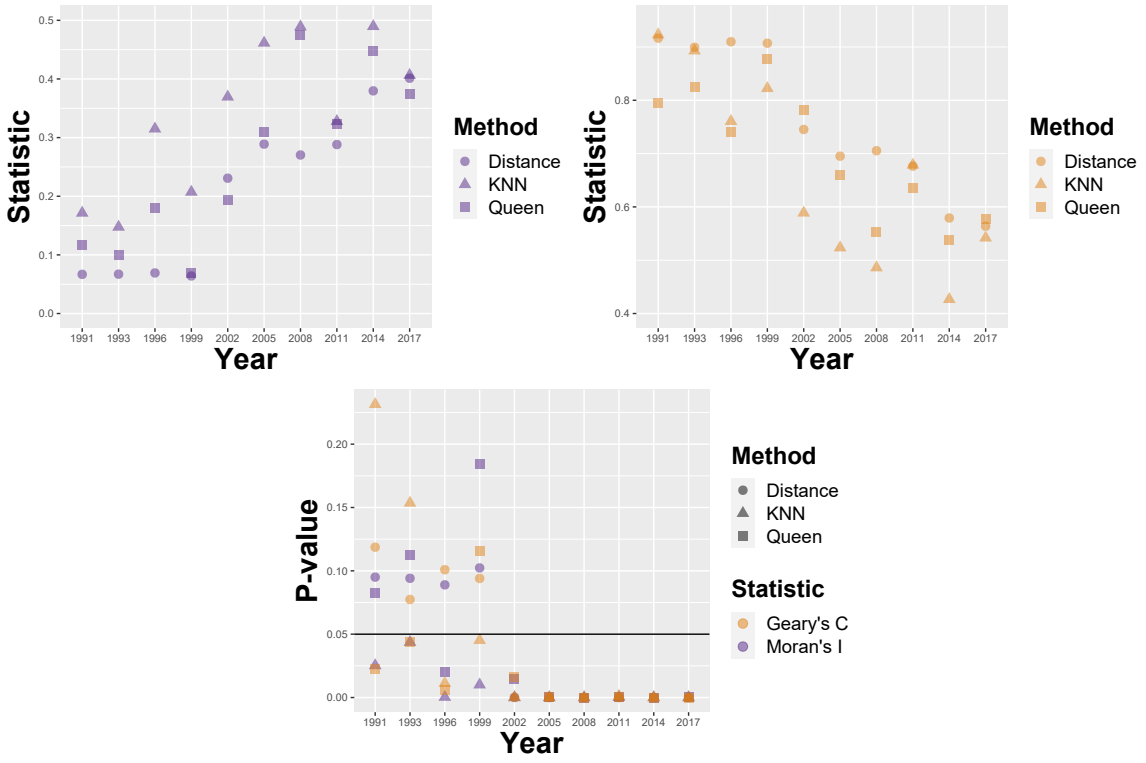

Fig. C.9: NYC median housing value spatial autocorrelation tests 1991 - 2017

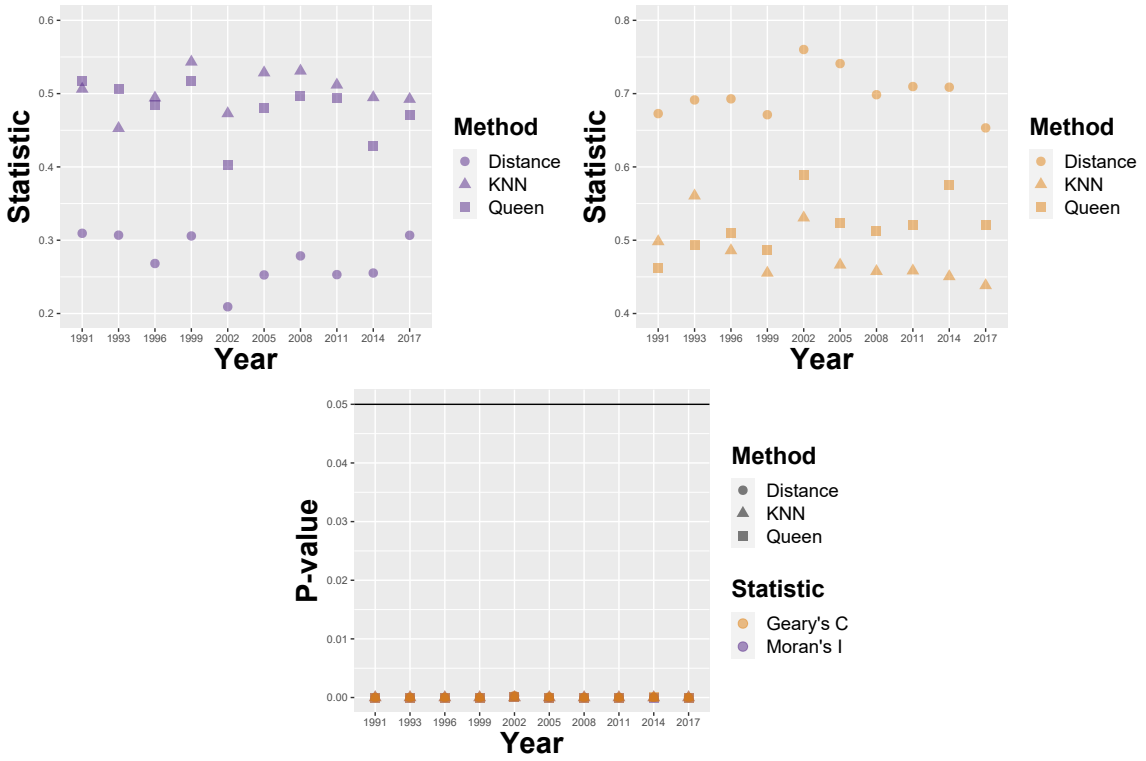

Fig. C.10: NYC median household income spatial autocorrelation tests 1991 - 2017

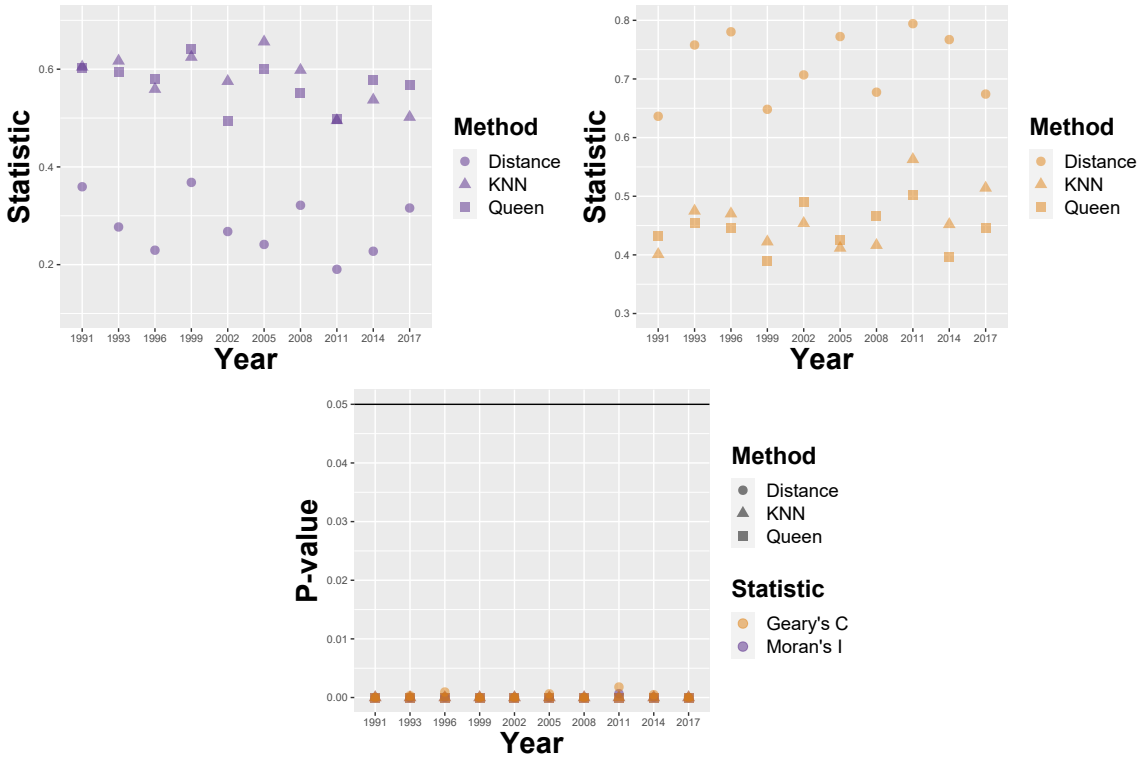

Fig. C.11: NYC female householder percentage spatial autocorrelation tests 1991 - 2017

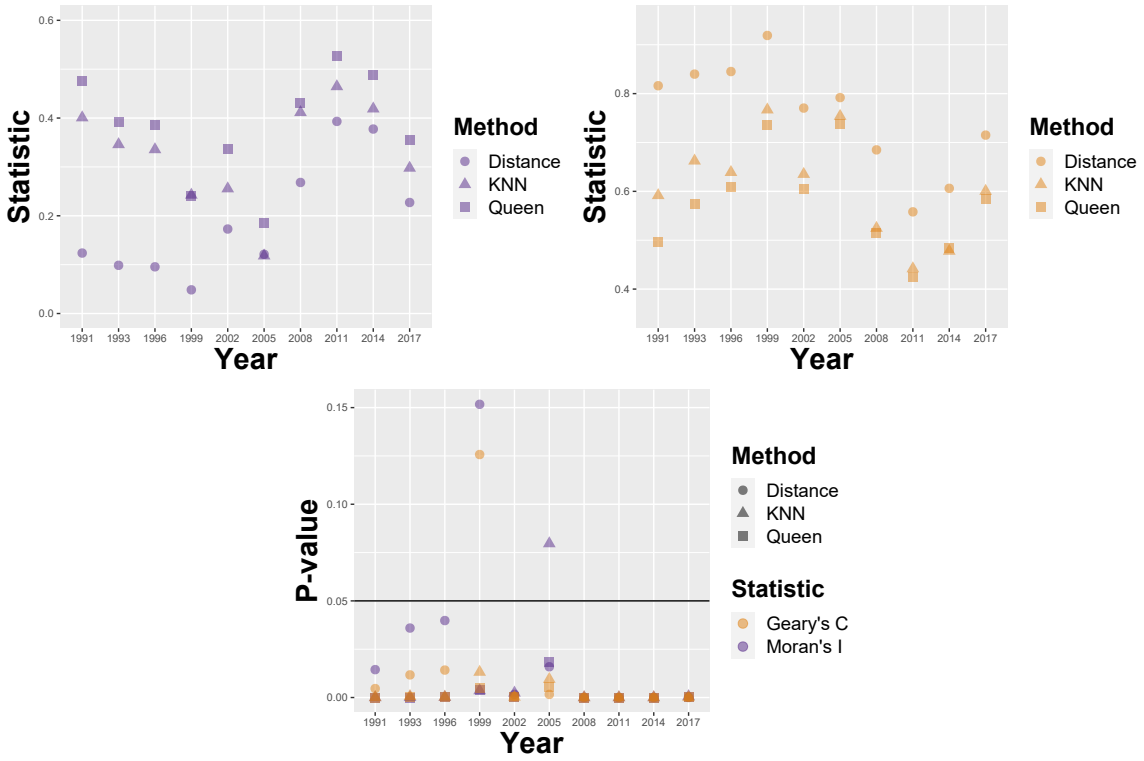

Fig. C.12: NYC median householder age spatial autocorrelation tests 1991 - 2017

APPENDIX D

Variance Inflation Index (VIF) of Selected Variables

Table D.1: Variance Inflation Factor (VIF) of the home ownership percentage, log home renting, and housing cost percentage explanatory variables

Variable	**Home Ownership %**	**Log Home Renting $**	**Housing Cost %**
Householder sociodemographics			
Householder's Birthplace	2.433	2.241	2.241
Householder's Age	2.671	2.566	1.765
Householder's Sex	1.574	3.226	3.037
Log Household Income	5.099	3.002	–
Housing value and features			
Log Housing Value	5.099	–	–
Mortgage Status	2.129	–	–
Mortgage interest rate	2.036	–	–
Walls Condition	2.036	1.957	1.957
Stairs Condition	2.007	1.511	1.457
Floors Condition	–	2.574	2.548
Number of rooms	3.081	1.404	1.392
Location variables			
Brooklyn Location	7.305	7.153	7.118
Bronx Location	6.072	5.392	5.383
Manhattan Location	7.716	4.555	4.111
Queens Location	6.477	6.015	5.943

Staten Island is used as the baseline.

Variables marked with "–" were not considered in the model because they had a high VIF or they were hard to explain in the model

APPENDIX E

Spatial Conditional and Simultaneous Autoregression (SAR) Model Estimation Results

Table E.1: OLS and SAR model estimates for home ownership percentage model

Variable	**OLS**	**SAR: Queen**	**SAR: knn**	**SAR: max. dist**
Intercept	-4.232***	-4.229***	-4.233***	-4.592***
Householder Sociodemographics				
Householder's Birthplace	-0.352**	-0.359***	-0.352***	-0.322***
Householder's Age	0.029***	0.030***	0.029***	0.030***
Householder's Sex	-0.080	-0.097	-0.081	-0.064
Log Household Gross Income	0.139*	0.141**	0.139**	0.144**
Housing Value and Features				
Log Housing Gross Value	0.102	0.101*	0.102*	0.121**
Mortgage Status	-0.093	-0.085	-0.093	-0.126
Mortgage Interest Rate	-0.236	-0.285	-0.237	-0.142
Walls Condition	0.830*	0.833**	0.830**	0.973**
Stairs Condition	-0.007	-0.011	-0.007	-0.087
Number of Rooms	-0.027	-0.027	-0.027	-0.040
Location Variables				
Brooklyn Location	-0.190**	-0.191***	-0.190***	-0.200***
Bronx Location	-0.156*	-0.156**	-0.156**	-0.198**
Manhattan Location	-0.455***	-0.455***	-0.455***	-0.504***
Queens Location	-0.030	-0.029	-0.029	-0.100
R-squared	0.789	0.789	0.789	0.794
λ	0	-0.080	-0.002	0.544

Notes: *** 1% level of significance; ** 5% level of significance;

* 10% level of significance. Staten Island is used as the baseline.

Table E.2: OLS and SAR model estimates for log median household rent model

Variable	**OLS**	**SAR: Queen**	**SAR: knn**	**SAR: max. dist**
Intercept	-0.534	-0.608	-0.540	-0.567
Householder Sociodemographics				
Householder's Birthplace	0.104	0.106	0.104	0.105
Householder's Age	-0.009	-0.008*	-0.009*	-0.009*
Householder's Sex	0.599	0.534	0.329*	0.611*
Log Household Gross Income	0.545***	0.556***	0.545***	0.546***
Housing Value and Features				
Walls Condition	1.100	1.420	1.021	1.086
Stairs Condition	-0.454	-0.496	-0.437	-0.432
Floors Condition	1.583	1.238	1.651*	1.593*
Number of Rooms	-0.100***	-0.098***	-0.101***	-0.099***
Location Variables				
Brooklyn Location	0.155*	0.148*	0.154*	0.154*
Bronx Location	0.174*	0.169**	0.174**	0.173*
Manhattan Location	0.134	0.120	0.135*	0.134*
Queens Location	0.170*	0.155**	0.173**	0.169**
R-squared	0.842	0.842	0.842	0.842
λ	0	-0.128	0.046	0.045

Notes: *** 1% level of significance; ** 5% level of significance;

* 10% level of significance. Staten Island is used as the baseline.

Table E.3: OLS and SAR model estimates for housing cost percentage

Variable	**OLS**	**SAR: Queen**	**SAR: knn**	**SAR: max. dist**
Intercept	-0.614*	-0.651**	-0.598**	-0.641 **
Householder Sociodemographics				
Householder's Birthplace	0.111**	0.130***	0.110***	0.102***
Householder's Age	0.001	0.001	0.002	0.001
Householder's Sex	0.303***	0.321***	0.306***	0.300***
Housing Value and Features				
Walls Condition	0.287	0.551*	0.246	0.206
Stairs Condition	0.014	-0.028	0.015	0.021
Floors Condition	0.405	0.226	0.427	0.518*
Number of Rooms	-0.015	-0.017*	-0.015*	-0.015*
Location Variables				
Brooklyn Location	0.032	0.023*	0.032	0.039*
Bronx Location	0.067**	0.063***	0.066***	0.070**
Manhattan Location	-0.010	-0.014	-0.010	-0.008*
Queens Location	0.022	0.014	0.023	0.024
R-squared	0.673	0.685	0.673	0.679
λ	0	-0.363	0.055	0.291

Notes: *** 1% level of significance; ** 5% level of significance;

* 10% level of significance. Staten Island is used as the baseline.

REFERENCES

Allen, R. (2020). The Relationship Between Legal Status and Housing Cost Burden for Immigrants in the United States. *Housing Policy Debate*, 0(0):1–23. DOI: 10.1080/10511482.2020.1848898.

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27(2):93–115.

Anselin, L. (2013). *Spatial Econometrics: Methods and Models*. Volume 4. Springer Science & Business Media. Dordrecht, Netherlands.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. Longman Scientific & Technical, Harlow, U.K.

Bivand, R., Keitt, T., and Rowlingson, B. (2019). *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. R package version 1.4-3 (https://CRAN.R-project.org/package=rgdal).

Bivand, R., Millo, G., and Piras, G. (2021). A Review of Software for Spatial Econometrics in R. *Mathematics*, 9(11):1276.

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied Spatial Data Analysis with R, Second Edition*. Springer, New York, NY. (https://asdar-book.org/).

Bivand, R. S. and Rundel, C. (2014). *rgeos: Interface to Geometry Engine — Open Source (GEOS)*. R package version 0.3–4 (http://CRAN.R-project.org/package=rgeos).

Bogdon, A. S. and Can, A. (1997). Indicators of Local Housing Affordability: Comparative and Spatial Approaches. *Real Estate Economics*, 25(1):43–80.

Brewer, C. A., Hatchard, G. W., and Harrower, M. (2003). ColorBrewer in Print: A Catalog of Color Schemes for Maps. *Cartography and Geographic Information Science*, 30(1):5–32.

Campbell, J. Y. and Cocco, J. F. (2015). A Model of Mortgage Default. *The Journal of Finance*, 70(4):1495–1554.

Carr, D. B. and Pierson, S. M. (1996). Emphasizing Statistical Summaries and Showing Spatial Context with Micromaps. *Statistical Computing and Statistical Graphics Newsletter*, 7(3):16–23.

Chamberlain, S. and Teucher, A. (2019). *geojsonio: Convert Data from and to 'GeoJSON' or 'TopoJSON'*. R package version 0.7.0 (https://CRAN.R-project.org/package=geojsonio).

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2021). *Shiny: Web Application Framework for R*. (https://CRAN.R-project.org/package=shiny).

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). Local Regression Models. In Chambers, J. M. and Hastie, T. J., editors, *Statistical Models in S*, pages 309 – 373. Routledge, New York.

Dacquisto, D. J. and Rodda, D. T. (2006). *Housing Impact Analysis*. US Department of Housing and Urban Development, Office of Policy Development. Available at https://www.huduser.gov/Publications/pdf/hsgimpact.pdf.

De Jong, P., Sprenger, C., and Van Veen, F. (1984). On Extreme Values of Moran's I and Geary's C. *Geographical Analysis*, 16(1):17–24.

Dell'Olio, F. (2004). Immigration and Immigrant Policy in Italy and the UK: Is Housing Policy a Barrier to a Common Approach towards Immigration in the EU? *Journal of ethnic and migration studies*, 30(1):107–128.

DeSilva, S. and Elmelech, Y. (2012). Housing Inequality in the United States: Explaining the White-Minority Disparities in Homeownership. *Housing Studies*, 27(1):1–26.

Doling, J. and Ronald, R. (2010). Home Ownership and Asset-Based Welfare. *Journal of Housing and the Built Environment*, 25(2):165–173.

Elmelech, Y. (2004). Housing Inequality in New York City: Racial and Ethnic Disparities in Homeownership and Shelter-Cost Burden. *Housing, Theory and Society*, 21(4):163–175.

Geary, R. C. (1954). The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, 5(3):115–146.

Gebreab, S. Y., Gillies, R. R., Munger, R. G., and Symanzik, J. (2008). Visualization and Interpretation of Birth Defects Data Using Linked Micromap Plots. *Birth Defects Research (Part A): Clinical and Molecular Teratology*, 82(2):110–119.

Geniaux, G. and Martinetti, D. (2018). A New Method for Dealing Simultaneously with Spatial Autocorrelation and Spatial Heterogeneity in Regression Models. *Regional Science and Urban Economics*, 72:74–85.

George, U. and Chaze, F. (2009). Social Capital and Employment: South Asian Women's Experiences. *Affilia*, 24(4):394–405.

Getis, A. (1991). Spatial Interaction and Spatial Autocorrelation: A Cross-Product Approach. *Environment and Planning A*, 23(9):1269–1277.

Getis, A. (2010). Spatial Autocorrelation. In Fischer, M. and Getis, A., editors, *Handbook of Applied Spatial Analysis*, pages 255–278. Springer, Berlin, Heidelberg.

Getis, A. and Aldstadt, J. (2004). Constructing the Spatial Weights Matrix Using a Local Statistic. *Geographical Analysis*, 36(2):90–104.

Griffith, D. A. (1988). Estimating Spatial Autoregressive Model Parameters with Commercial Statistical Packages. *Geographical Analysis*, 20(2):176–186.

Hayes, A. F. and Cai, L. (2007). Using Heteroskedasticity-Consistent Standard Error Estimators in OLS Regression: An Introduction and Software Implementation. *Behavior Research Methods*, 39(4):709–722.

Herbert, C. E., Haurin, D. R., Rosenthal, S. S., and Duda, M. (2005). Homeownership Gaps Among Low-Income and Minority Borrowers and Neighborhoods. *Washington, DC: US Department*

of Housing and Urban Development. (http://www.huduser.org/publications/HOMEOWN/HGapsAmongLInMBnN.html).

Hubert, L. J. and Golledge, R. G. (1981). A Heuristic Method for the Comparison of Related Structures. *Journal of Mathematical Psychology*, 23(3):214–226.

Hubert, L. J., Golledge, R. G., and Costanzo, C. M. (1981). Generalized Procedures for Evaluating Spatial Autocorrelation. *Geographical Analysis*, 13(3):224–233.

Latif, E. (2015). Immigration and Housing Rents in Canada: A Panel Data Analysis. *Economic Issues*, 20(1):91–108.

LeSage, J. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Chapman and Hall/CRC. New York, NY.

Leung, Y., Mei, C.-L., and Zhang, W.-X. (2000). Testing for Spatial Autocorrelation Among the Residuals of the Geographically weighted regression. *Environment and Planning A*, 32(5):871–890.

Liu, R., Li, T., and Greene, R. (2020). Migration and Inequality in Rental Housing: Affordability Stress in the Chinese Cities. *Applied Geography*, 115(102138).

Mansfield, E. R. and Helms, B. P. (1982). Detecting Multicollinearity. *The American Statistician*, 36(3a):158–160.

McConnell, E. D. and Akresh, I. R. (2010). Housing Cost Burden and New Lawful Immigrants in the United States. *Population Research and Policy Review*, 29(2):143–171.

Medri, J., Probst, B., and Symanzik, J. (2019). Housing Affordability and Immigration: An Exploratory Analysis in New York City. In *JSM Proceedings, Statistical Computing Section*. Alexandria, VA: American Statistical Association. 2549-2564.

Medri, J., Probst, B., and Symanzik, J. (2021). "Housing Affordability and Immigration: An Exploratory Analysis in New York City". *Computational Statistics*, Under Review.

Moos, M. and Skaburskis, A. (2010). The Globalization of Urban Housing Markets: Immigration and Changing Housing Demand in Vancouver. *Urban Geography*, 31(6):724–749.

Moran, P. A. (1948). The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):243–251.

Mulder, C. H. and Smits, J. (1999). First-Time Home-Ownership of Couples: the Effect of Inter-Generational Transmission. *European Sociological Review*, 15(3):323–337.

Mussa, A., Nwaogu, U. G., and Pozo, S. (2017). Immigration and Housing: A Spatial Econometric Analysis. *Journal of Housing Economics*, 35:13–25.

Myers, D. and Lee, S. W. (2018). Immigrant Trajectories into Homeownership: A Temporal Analysis of Residential Assimuation 1. In Suarez-Orozco, M. M., Suarez-Orozco, C., and Qin-Hillard, D., editors, *Interdisciplinary Perspectives on the New Immigration*, pages 307–339. Routledge, New York.

Neter, J., Wasserman, W., and Kutner, M. H. (1983). *Applied Linear Regression Models.* Richard D. Irwin Inc, Homewood, Illinois.

Nuesch-Olver, D. (2002). Thank You for Asking: The Experiences of Latina Immigrant Professional Women. *Social Work and Christianity*, 29(1):31–53.

Nygaard, C. (2011). International Migration, Housing Demand and Access to Homeownership in the UK. *Urban Studies*, 48(11):2211–2229.

O'Dell, W., Smith, M. T., and White, D. (2004). Weaknesses in Current Measures of Housing Needs. *Housing and Society*, 31(1):29–40.

Ottaviano, G. I. and Peri, G. (2012). The Effects of Immigration on US Wages and Rents: A General Equilibrium Approach. In Nijkamp, P., Poot, J., and Sahin, M., editors, *Migration Impact Assessment*. Edward Elgar Publishing, Cheltenham, UK.

Owusu, T. Y. (1998). To Buy or Not to Buy: Determinants of Home Ownership Among Ghanaian Immigrants in Toronto. *The Canadian Geographer/Le Géographe Canadien*, 42(1):40–52.

Parker, E. (2021). *COVID-19 Tracker*. (https://shiny.rstudio.com/gallery/covid19-tracker.html).

Payton, Q. C. and Olsen, A. R. (2015). *micromap: Linked Micromap Plots*. R package version 1.9.1 (http://CRAN.R-project.org/package=micromap).

Pettit, C., Tice, A., and Randolph, B. (2017). Using an Online Spatial Analytics Workbench for Understanding Housing Affordability in Sydney. In Thakuriah, P., Tilahun, N., and Zellner, M., editors, *Seeing Cities Through Big Data*, pages 233–255. Springer, Cham.

Piantadosi, S., Byar, D. P., and Green, S. B. (1988). The Ecological Fallacy. *American Journal of Epidemiology*, 127(5):893–904.

Probst, B. D. (2020). *'LMshapemaker': Utilizing the 'Rmapshaper'R Package to Modify Shapefiles for Use in Linked Micromap Plots*. Master's thesis. *All Graduate Theses and Dissertations.* 7751. (https://doi.org/10.26076/j9yj-mm66).

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. (https://www.R-project.org/).

Schwartz, M. and Wilson, E. (2008). Who Can Afford to Live in a Home?: A Look at Data from the 2006 American Community Survey. *US Census Bureau*, pages 1–13.

Shier, M. L., Graham, J. R., Fukuda, E., and Turner, A. (2016). Predictors of Living in Precarious Housing Among Immigrants Accessing Housing Support Services. *Journal of International Migration and Integration*, 17(1):173–192.

Sinning, M. (2010). Homeownership and Economic Performance of Immigrants in Germany. *Urban Studies*, 47(2):387–409.

Symanzik, J. and Carr, D. B. (2008). Interactive Linked Micromap Plots for the Display of Geographically Referenced Statistical Data. In Chen, C., Härdle, W., and Unwin, A., editors, *Handbook of Data Visualization*, pages 267–294 & 2 Color Plates. Springer, Berlin, Heidelberg.

Teucher, A. and Russell, K. (2018). *rmapshaper: Client for 'mapshaper' for 'Geospatial' Operations*, 2 edition. R package version 0.4.1 (https://CRAN.R-project.org/package=rmapshaper).

U.S. Census Bureau (2019a). 2013-2017 American Community Survey 5-Year Estimates. "Table DP04; generated using American FactFinder; (25 September 2019). (https://www.census.gov/quickfacts/newyorkcitynewyork)".

U.S. Census Bureau (2019b). 2013-2017 American Community Survey 5-Year Estimates. "Table S0501; generated using American FactFinder; (25 September 2019). (https://www.census.gov/quickfacts/newyorkcitynewyork)".

Whittle, P. (1954). On Stationary Processes in the Plane. *Biometrika*, 41(3):434–449.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis, Second Edition*. Springer, New York, NY. (https://ggplot2-book.org).

Wickham, H., François, R., Henry, L., and Müller, K. (2019). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.3 (https://CRAN.R-project.org/package=dplyr).

Wright, S. P. (1992). Adjusted P-Values for Simultaneous Inference. *Biometrics*, 48(4):1005–1013.

Zou, Y. (2014). Analysis of Spatial Autocorrelation in Higher-Priced Mortgages: Evidence from Philadelphia and Chicago. *Cities*, 40 Part A:1–10.