Computational Statistics manuscript No.
(will be inserted by the editor)

A statistical framework for analyzing housing quality

A case study of New York City

Damien Chambon · Jacob Gerszten

Received: date / Accepted: date

Abstract The physical condition of a person's home plays a significant role in determining the dweller's overall quality of life. This paper provides a statistical framework for measuring housing quality in an urban area through a standardized index. Using demographic, geographic, and economic factors from the New York City Housing and Vacancy Survey, this index is constructed using principal component analysis. Differences in housing quality based upon ownership status were tested and demonstrated that renters face more housing quality issues than owners. Several of the variables driving these differences were found to have varying effects on housing quality over time, in part due to the 2008 financial crisis. Using this novel statistical framework, housing quality indices can be constructed for other cities to investigate housing disparities and inform policies aimed at improving overall quality of life for urban residents.

Keywords Housing quality index · Ownership status · Principal component analysis · Linear regression · Mann-Whitney test

1 Introduction

Houses with cracked walls and broken windows are eyesores to some observers, but the impact on residents goes beyond outward cosmetics. Housing quality is an important determinant of health, and substandard housing conditions represent a major public health issue in the United States and other countries. Poor housing conditions are associated with a wide variety of mental and physical health issues for residents (Krieger and Higgins, 2002). For instance, it was shown that living in housing without functioning heat can cause respiratory health issues (Evans et al., 2000). In addition, housing serves a crucial purpose in producing social identity. Poor housing quality can therefore have adverse social outcomes for entire communities (Dunn and Hayes, 2000). Having a standardized method for measuring

Damien Chambon
E-mail: dlc8mt@virginia.edu

Jacob Gerszten
E-mail: jeg6bk@virginia.edu

housing quality in a large city such as New York is a beneficial tool when analyzing housing disparities across communities. This paper provides a statistical method for constructing a housing quality index using publicly available survey data that can be used to identify drivers of differences across urban communities.

New York City has faced many challenges related to high levels of poverty, crime, and homelessness throughout its history (Neckerman et al., 2016). To combat these issues, New York City has relied on a "broken windows" policy since the 1990s, where the city police inferred that relatively small issues, such as broken windows, provide an environment that enables more crime (Wilson and Kelling, 1982). This policy suggests that lower levels of housing quality are directly linked to higher levels of crime. While the results of the broken windows policing policies have been mixed (Corman and Mocan, 2005), the theory is still prominent, especially with the continued use of "stop-and-frisk" measures. In localities that employ this policing style, measuring and addressing housing quality directly relates to reducing crime. Using a standardized measure of housing quality, various structural, demographic, and geographic variables can be tested to determine the influential predictors of housing conditions.

This paper is organized as follows. First, a novel method is proposed for creating a housing quality index (HQI) which characterizes various features of an individual building's overall condition. Second, the differences between the HQI distributions for renters and owners are tested using a non-parametric statistical test. Finally, regression models are built to determine the most relevant factors impacting the HQI, testing to see if certain factors affect renters and owners in different ways and how their effects have evolved over time.

2 Data

The data used in this study come from the New York City Housing and Vacancy Survey (NYCHPD, 2017), which contains survey results of the New York City housing stock and population from 1991 to 2017, collected every three years. The ten surveys, sponsored by the New York City Department of Housing Preservation and Development, are representative of housing across all five boroughs of New York City. The dataset contains roughly 200 variables and 10,000 unique buildings per year. The data reflect New York City's size and diversity and encompass a wide range of responses designed to capture detailed micro-data for each dwelling. The exact number of variables and observations differ between years due to changes in survey questions and design, but are similar enough over the entire data collection period to be suitable for analysis.

The data were cleaned prior to analysis, which included selecting relevant variables that appeared in all surveys, re-coding values for some categorical variables, and removing observations with missing values. Several variables were also transformed into a more useful format, such as constructing *length of stay* from the year that the householder moved into the apartment. In the end, 34.57% of the observations were removed and 31 variables were retained for subsequent analysis.

3 Methodology

3.1 Constructing a Housing Quality Index

To create the Housing Quality Index (HQI), only variables related to housing quality were selected: *wall severity*, *window severity*, *stairway severity*, *floor severity*, *building condition*, *toilet breakdowns*, *kitchen functioning*, *mice and rats*, *cracks in walls*, *holes in floors*, *broken plaster*, and *water leakage*. These housing condition variables were one-hot-encoded so that one binary variable was created for each possible value. Principal component analysis (PCA) was then used to reduce the dimensionality of the aforementioned variables into one index. This method weighs the various housing features depending on how much new information they add: if they have low variance, they have a small impact on the final index. By outputting a single value, PCA allows housing quality to be compared between different buildings. In addition, it addresses differences in how often the issues occur, so less common conditions are weighted more heavily. In other words, for two households with the same overall number of issues, the one that faces less common issues would have a higher HQI.

While PCA is an efficient means of creating a new index from a large number of existing variables, the index is only meant to be used for comparing households and cannot be directly interpreted. It is difficult to determine whether a high HQI value for a given household results from having many small issues or from fewer but more heavily weighted issues. Initially, the HQI values ranged from -0.403 to 2.935. The index was then scaled from 0 to 10, with a value of 0 representing no issues and a value of 10 corresponding to a household which reports every possible issue.

3.2 Comparing housing quality between owners and renters

To determine whether renters face additional housing issues, the differences between owner and renter HQI distributions are tested using a Mann-Whitney Test (Mann and Whitney, 1947). Because the test is non-parametric, it is not impacted by the logarithmic distributions of the HQIs. The null hypothesis of the Mann-Whitney test states that a randomly selected value from one population is equally likely to be greater as it is to be less than a randomly selected value from a second population. In this case, our null hypothesis states that the HQI levels of renters are similar to those of owners. Our alternative hypothesis posits that the HQI levels of owners are lower than those of renters. In this test, the test statistic is equal to the number of pairs x_i, y_j such that $x_i < y_j$ and where x_i belongs to the set of the renters HQIs and y_j belongs to the set of owners HQIs. For this test, our null hypothesis is the following:

$$H_0 : F_{renters} = F_{owners} \tag{1}$$

where $F_{renters}$ is the distribution of the HQI for renters and F_{owners} is that of the owners. Our alternative hypothesis was the following:

$$H_a : F_{renters} \leq F_{owners} \; for \; all \; x \tag{2}$$

3.3 Determining relevant factors impacting housing quality

In addition to ownership status, other factors can explain variation in housing quality across dwellings. Determining additional significant predictors of housing quality is essential for creating a framework to assess housing quality in a particular city. A linear regression model was used to test whether or not other factors were significant predictors of HQI, the response variable. The linear regression model is as follows:

$$\begin{aligned} HQI = \beta_0 + \beta_1 HouseholderSex + \beta_2 HouseholderAge + \\ \beta_3 HouseholderHispanicOrigin \\ + \beta_4 DurationOfStay + \beta_5 NumberOfUnits + \\ \beta_6 OwnerInBuilding + \beta_7 NumberOfStories + \beta_8 NumberOfRooms + \\ \beta_9 PlumbingFacilities + \beta_{10} KitchenFacilities + \beta_{11} LengthOfLease + \\ \beta_{12} ResidentRating + \beta_{13} HouseholderIncome \\ + \beta_{14} Year + \beta_{15} Borough + \beta_{16} Status \end{aligned} \tag{3}$$

3.4 Testing the impact of ownership status on housing quality

To test the significance of the impact of ownership status on different factors, a linear regression model was built for the entire dataset with *ownership status* as an interaction term and the HQI as the response variable. This method identifies which variables had a significantly different impact on the HQI between the two groups. In addition, two linear regression models with the HQI as the response variable were created for each status. The differences between the coefficients of the two regression models were computed for the variables which were previously identified to have different impacts on owners and renters. The percentage difference in coefficients for each significantly changing variable was computed so that the differences could be compared on the same scale.

3.5 Analyzing change in factor importance over time

A similar process to the one for testing the impact of ownership status on housing quality was employed to track how the relative importance of each predictor evolves over time. First, a linear regression model with the year as an interaction term and the HQI as the response variable was built incorporating both owners and renters. This model identifies which variables had a significantly different impact on the HQI over time. A linear regression model with the HQI as the dependent variable was then created for each year in the dataset so that the varying coefficient values for each significant variable could be analyzed throughout the years of the survey.

4 Results and Discussion

4.1 Creating the Housing Quality Index

The HQI generated meaningful insights into the distributions of housing quality for both renters and owners. Summary statistics for the index are provided in Table 1, which provides several important facts about the new index. First, a large number of both renters and owners face no condition issues, while some individuals face much higher levels, including some renters who face every possible housing issue. The overall mean of 1.207 shows that most people have few issues, on average. The quartiles show that 25% of the surveyed population have HQIs between 1.7 and 10.0 suggesting a heavy right tail. Most residents have few issues but there are some residents who experience many housing problems. These trends are visible in Figure 1, histograms of the HQIs for both categories, which shows that both distributions are heavily right skewed and appear to be logarithmic.

Table 1 House Quality Index statistics. Values were rounded up to 3 decimals.

Statistic	Total	Owners	Renters
Mean	1.207	0.56	1.541
Std. Dev.	1.639	1.005	1.796
Min	0	0	0
Q1	0	0	0
Median	0.28	0	1.08
Q3	1.7	1.08	2.677
Max	10.0	9.286	10.0

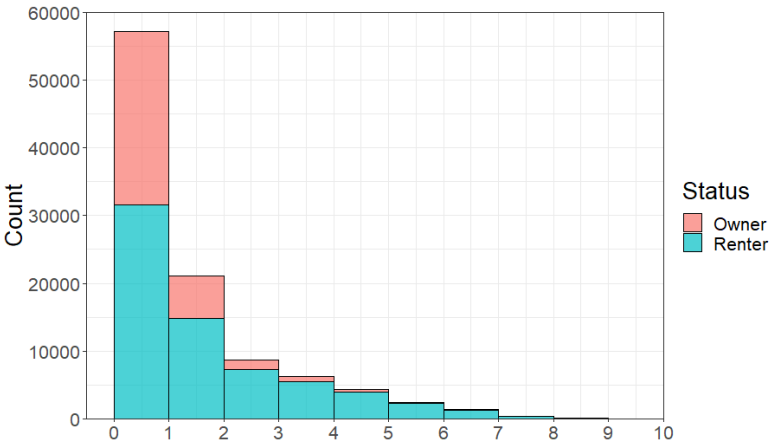

Fig. 1 HQI counts by ownership status

4.2 Comparing housing quality between owners and renters

One of the biggest choices an individual who resides in a large urban area makes is whether to own or rent a home. There are almost twice as many renters as there are owners in this data set (34648 owners vs 67135 renters), showing that New York City is a city where most people rent rather than own their home. As presented in Table 1, the mean value for renters is almost a full unit higher than mean value for owners, which suggests that renters face more housing issues. The median of zero for owners indicates that over half of them did not report a single housing issue. The median value of the index for renters is close to 1, which is relatively low on the scale despite the long right tail. The distributions of both groups are presented in Figure 2.

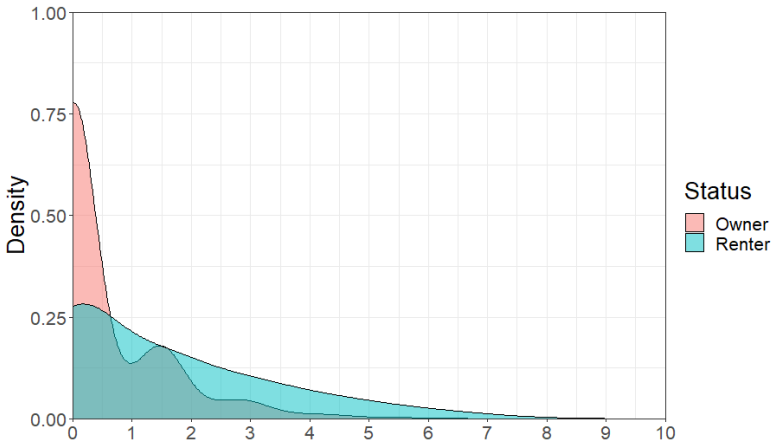

Fig. 2 HQI density plot by ownership status

The distributions for both groups have a similar shape. However, a major difference is that the distribution for renters is more right-skewed, which suggests that renters tend to experience more housing quality issues. That is confirmed with the Mann-Whitney test since the test statistic is equal to 780330000 and the p-value is very close to 0. Therefore, the HQI distribution for renters is greater than that for owners, meaning renters do tend to have more issues with their dwellings.

4.3 Determining relevant factors impacting housing quality

To identify other important factors that may impact housing quality, a linear regression model with HQI as the response was constructed.

In Figure 3, the distributions of *householder age* and *number of rooms* appear normal. Variables *household value* and *monthly rent* contained placeholder values, denoted as *9999999*, that needed to be removed. In fact, renters show a household value denoted as *9999999* while the monthly rent is given. That is the opposite for owners: the household value is given but the monthly rent contain the value

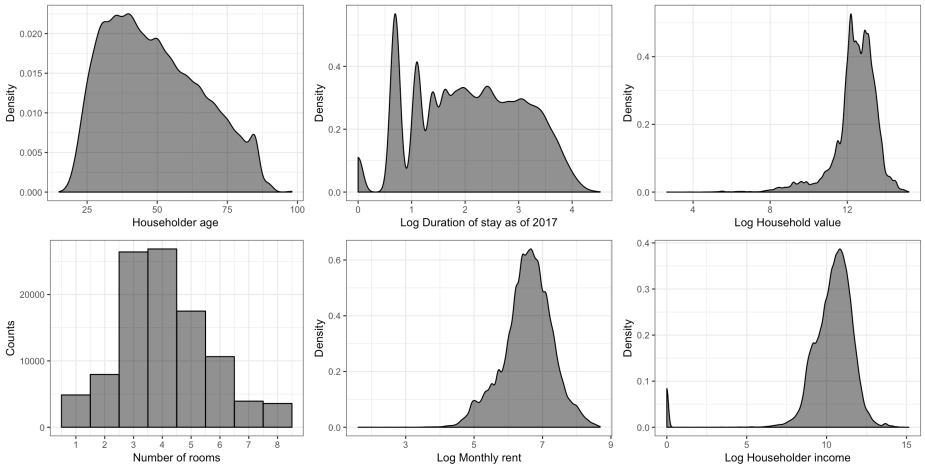

Fig. 3 Distribution plots of the numerical variables included in the model

9999999. After removing the placeholder value, both variables were logged. Finally, the variables *duration of stay as of 2017* and *householder income* first appeared to be close to a logarithmic distribution. A log transformation was also applied on them. Their resulting distribution can be seen in Figure 3 as well.

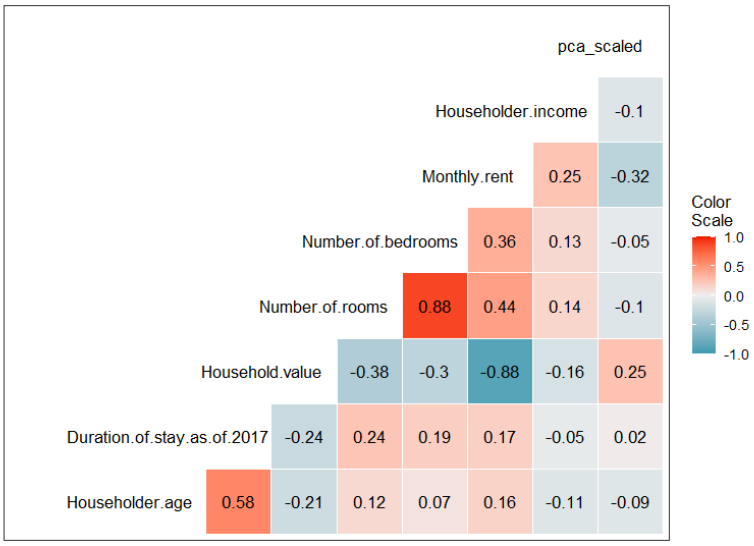

Fig. 4 Correlation matrix of the numerical variables included in the model

To check for multicollinearity, a correlation matrix of the predictors was created, as seen in Figure 4. Two pairs of variables exhibit strong evidence of multicollinearity: *household value* with *monthly rent*, and *number of rooms* with *number of bedrooms*. For the first pair, this issue is quite easy to deal with. As previously

stated, no observation has two actual values for those variables at the same time since one of the two variables has a placeholder variable based on the status of the resident, e.g. *999999* for *household value* for renters and *999999* for *monthly rent* for renters. Therefore, in the analysis, the two variables are never included at the same time. When the entire dataset is studied, those variables are not taken into account. When the renters are studied, *monthly rent* is included in the analysis but not *household value*, and vice versa for owners. Thus, placeholder values will not hinder analysis for the whole sample or either subgroup.

Multicollinearity makes sense for the second pair, *number of rooms* and *number of bedrooms*, since having additional bedrooms increases the number of total rooms. As the correlation is high (around 0.9), one of the two variables had to be removed. Given that *number of bedrooms* was less normally distributed than *number of rooms*, *number of bedrooms* was removed. Finally, *householder age* and *duration of stay* are positively correlated, which is logical since older householders have had more time to stay in the same dwelling. Both variables were kept for analysis as they did not impact the performance of the models.

A linear regression was fitted with all observations and all remaining predictors against the HQI. As mentioned before, *monthly rent* and *household value* were removed. Furthermore, variable *length of lease* was removed so that it does not conflict with the variable *status*. Indeed, the value *owner-occupied* for the length of the lease necessarily implies a value of *owner* for the ownership status. As the variable *length of lease* is not as important in the analysis, we decided to keep *status* instead.

Table 2 Subset of the fitted linear regression model. The baseline for the number of stories is *Number of stories: 1-2*. The baseline for the borough is *Borough: Bronx.*. The R-squared value is 0.2649. Values were rounded up to 3 decimals.

Predictor	Estimate	p-value
Householder age	-0.008	<2e-16
Duration of stay as of 2017	0.173	<2e-16
Owner in building	-0.349	<2e-16
Number of stories: 3-5	0.118	2.28e-12
Number of stories: 6-10	-0.086	2.52e-4
Number of stories: 11-20	-0.482	<2e-16
Number of stories: 21+	-0.639	<2e-16
Householder income	-0.001	0.590
Borough: Brooklyn	-0.114	4.67e-15
Borough: Manhattan	0.017	0.262
Borough: Queens	-0.397	<2e-16
Borough: Staten Island	-0.343	<2e-16
Status: Renter	0.210	<2e-16

Most of the predictors are considered significant, such as *householder age*, *number of stories in the building*, *length of the lease*, and *borough*. Table 2 shows a subset of the coefficients of the regression model. Only variables that are significant at the 0.001 threshold are considered to have an impact on the HQI. With this threshold taken into consideration, *householder income* surprisingly does not play a role in determining HQI. The coefficient of status being equal to *renter* is significant and has a positive value. This indicates that for dwellings with the

exact same characteristics, the average HQI will be 0.210 higher for renters than owners. That value supports the previous finding that renters have higher housing condition index levels, meaning that they have more issues.

Specifically looking at the number of stories in a building, some insights can be extracted. Compared to building with 1 or 2 floors, buildings that have between 3 and 5 stories have worse housing conditions (average of 0.118 of the increase in HQI), holding other factors constant. However, as the number of stories keeps increasing, housing conditions improve on average, holding other factors constant. Indeed, the coefficient is significantly more and more negative for an increasing number of stories. That phenomenon could be explained by the fact that taller buildings tend to contain households with higher incomes, therefore providing higher rents to the building managers, who can afford to fix potential housing issues. Another reason could be that such buildings are managed by larger companies which can better monitor a building's condition as well as invest proper housing maintenance.

4.4 Testing the impact of ownership status on housing quality

After measuring the impact of variables on the HQI, it is important to see whether the impact of certain variables on the HQI changes depending on ownership status. To do so, a linear regression was run for the entire dataset with the HQI as the response variable, this time adding ownership status as an interaction term. With this model, an ANOVA test was performed to determine whether adding the interaction term made the predictor significant or not. Table 3 contains a subset of the results.

Table 3 Subset of the ANOVA of the linear regression model. Values were rounded up to 3 decimals.

Predictor with *Status* as an interaction term	p-value
Householder age	1.308e-04
Householder sex	0.104
Duration of stay as of 2017	<2e-16
Number of units	<2.2e-16
Number of stories	1.269e-13
Householder income	0.371
Borough	<2.2e-16

An important finding is that *ownership status* does not change the impact of a household's income on the HQI. Indeed, the p-value is quite large (more than 0.37). That implies that an owner and a renter who have the same income will have, on average, the same intensity of housing issues. Conversely, an owner and a renter who live in the same borough of the city will have, on average, a very different number of issues given the very low p-value for that predictor.

To quantify the discrepancies between the impact of the predictors, the differences between the coefficients of the regression for the renters and the regression for the owners were computed. The percentage that the difference represents com-

pared to the value of the coefficient for the renters was calculated so that the differences could be compared. The results are displayed in Table 4.

Table 4 Subset of the differences in coefficients of different predictors. Percentages of differences are computed compared to renters. Values were rounded up to 3 decimals.

Predictor	Coeff. renters	Coeff. owners	Percentage difference (in %)
Householder age	-0.010	-0.003	-64.387
Duration of stay as of 2017	0.224	0.04	-82.207
Nb. of units: 2 units	0.078	-0.040	-151.498
Nb. of units: 3-5 units	0.244	-0.089	-136.653
Nb. of units: 6-9 units	0.577	0.187	-67.541
Nb. of units: 10-19 units	0.642	0.229	-64.254
Nb. of units: 20-49 units	0.718	0.297	-58.603
Nb. of units: 50-99 units	0.624	0.284	-54.497
Nb. of units: 100+ units	0.596	0.243	-59.263
Nb. of stories: 3-5 stories	0.176	0.103	-41.731
Nb. of stories: 6-10 stories	-0.019	-0.059	197.837
Nb. of stories: 11-20 stories	-0.499	-0.266	-46.724
Nb. of stories: 21+ stories	-0.718	-0.327	-54.449
Number of rooms	0.103	0.021	-79.593
Borough: Brooklyn	-0.128	-0.016	-87.670
Borough: Manhattan	-0.002	0.127	-5,812.815
Borough: Queens	-0.492	-0.174	-64.766
Borough: Staten Island	-0.371	-0.210	-43.506

First of all, all significant percentage differences, except one, are negative. That means that housing conditions of renters are more strongly impacted, positively or negatively, by any factor compared to owners'. Those differences in absolute value are large, meaning that the difference in impact is quite strong.

The largest percentage difference in absolute value corresponds to *borough* being equal to "Manhattan". It has a value of -5,812%, meaning that living in Manhattan is 5,812% more beneficial for renters than owners on average. In other words, for a renter living in Manhattan, the average increase of the HQI of their dwelling will be 5,812% smaller than that of owners. We posit that for an owner, living in Manhattan could end up being detrimental for housing quality because life is quite expensive in that area, and they would not spend as much money on their own dwelling as if they were living in a cheaper area such as Staten Island, where the difference in impact is smaller. It is interesting to note that it appears to be more beneficial to be a renter in all areas of New York City, though it contradicts our previous finding that being an owner was related to a lower housing condition index. A possible explanation could be that living in the Bronx (the baseline) may be much better for owners compared to renters, impacting the total average HQI for owners.

The second largest percentage, 197.837%, corresponds to the number of stories being between 6 and 10. As the percentage is positive, it means that the impact of living in such a building on the HQI is greater for owners compared to renters. That is the only factor in this case. As the coefficient is negative (-0.019 for renters and -0.059 for owners), that shows that owners living in buildings with 6-10 stories face fewer housing issues. However, the coefficients for that variable are very small for both ownership status, suggesting a minimal impact on the HQI.

4.5 Analyzing change in factor importance over time

A final advantage of the novel framework provided in this study is also the ability to understand how the coefficients of the predictors of the previous linear regression evolved over the years. Similar to what was used for ownership status, a linear regression model was created with the HQI as the response variable and the year as an interaction term. An ANOVA test was performed on the interaction terms to determine whether the specific year had a significant impact. In other words, it was used to show which coefficients had significantly different values over the years. The results of the ANOVA are in Table 5.

Table 5 Subset of the ANOVA of the linear regression model. Values were rounded up to 3 decimals.

Predictor with *Year* as an interaction term	p-value
Householder age	<2.2e-16
Duration of stay as of 2017	4.185e-6
Number of units	<2.2e-16
Number of stories	1.297e-9
Resident rating of the neighborhood	5.195e-15
Householder income	0.068
Borough	<2.2e-16

Some insights can be extracted from Table 5. First, it appears that the *number of units* and *number of stories* have had varying impacts over the years as well as *householder age* and the borough where a particular dwelling is located. Another interesting fact is that *householder income* has a similar impact on HQI over time, though that impact is limited as demonstrated in the previous sections.

After identifying the significant changes in coefficients, linear regression models containing those coefficients were fitted with the HQI as the response variable for each year. By doing so, the coefficients for each predictor across the years could be obtained. The graphs showing the evolution of coefficients of a selection of significant coefficients regarding New York City housing are featured next.

In Figure 5, the changes in coefficients for each category of *Number of Units* all have similar changes over the years: they tend to increase and decrease at the same time. The overall trend is decreasing, meaning that they are representative of a lower HQI, compared to the baseline which is "1 unit". In other words, living in a single-unit dwelling is progressively less related to a lower HQI, meaning that it is an increasing source of dwelling problems. There is also a noticeable increase in coefficient in 2008 for the *number of units* being between 6 and 9. We can interpret this value by saying that the buildings comprising between 6 and 9 units were more impacted by the 2008 financial crisis compared to single-unit dwellings.

Figure 6 demonstrates how the *number of stories* impact the HQI in different ways over time. Buildings with *number of stories* between 3 and 5 have a roughly constant impact on the HQI over the years. In contrast, buildings with more than 21 stories encounter more issues over time as the coefficient has been steadily increasing since 1994. It is worth noting that its increase appears to slow down in recent years. Over time, the conditions in buildings with more than 21 stories have been deteriorating.

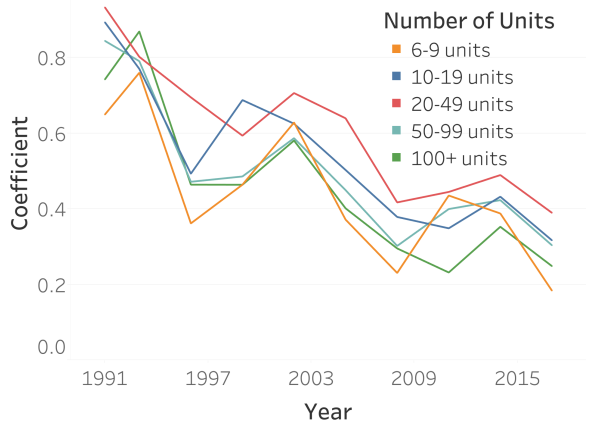

Fig. 5 Impact of *Number of Units* on the HQI over time

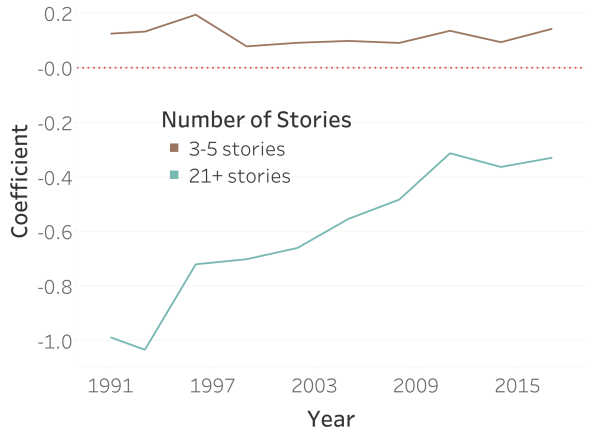

Fig. 6 Impact of *Number of Stories* on the HQI over time

Figure 7 displays that for previous years, living in Manhattan meant more issues for a dwelling compared to the Bronx. However, this trend became less pronounced in recent years. The value of the coefficient was around 0.3 and then steadily decreased. There is a sharp decrease in 2008 for that coefficient, the same year as the economic crisis that affected the world economy. Since the coefficient got smaller quite fast and eventually went negative, that shows that people living in Manhattan had a lower HQI than those in the Bronx, meaning that they were less affected by that economic crisis than the people living in the Bronx. Indeed, the populations in the Bronx are poorer and the economic crisis accentuated those financial issues. From 2008 onward, the coefficient stayed very low and even negative, which shows that the Bronx has never fully recovered from the 2008 crisis: from that point on, it is worse to live in the Bronx than in Manhattan in terms of housing conditions.

Finally, the changes in the coefficients for *age* over the years are represented in Figure 8. After successive increases and decreases, the coefficient has been in-

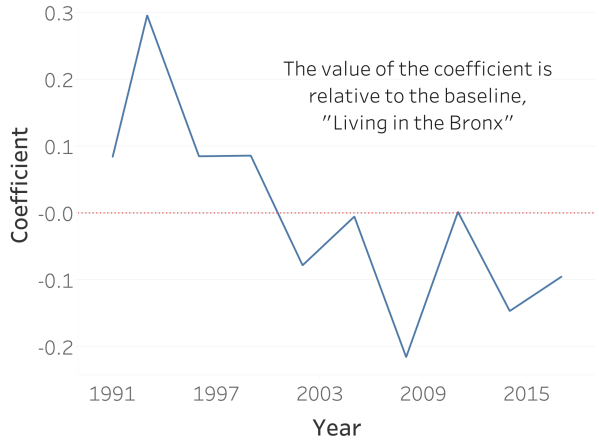

Fig. 7 Impact of *Living in Manhattan* on the HQI over time

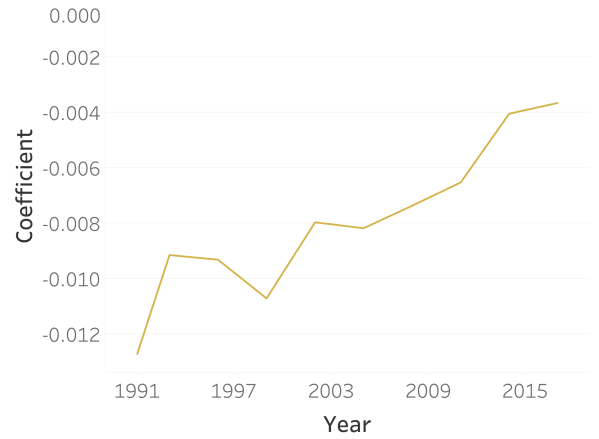

Fig. 8 Impact of *Age of Householder* on the HQI over time

creasing since 2005 and is getting closer to 0. This shows that the HQI is less and less negative as the householder age increases. In other words, after years of enjoying better housing conditions, older people are progressively less likely to experience less issues with their dwellings than young people. As the residents age, so does their dwelling, which could make the buildings more prone to physical issues. Therefore, the condition of old buildings in New York City should be examined by the City in order to help older residents live in better conditions. It is worth noting that the coefficient is relatively low, meaning that the impact of *age* is small, but it still is significant.

5 Conclusion

To mitigate the impact of substandard housing conditions on public health, a standardized method to measure and assess housing quality can identify dispari-

ties across populations. Working with an example of a major city like New York shows the power of the statistical framework provided in this study. After building a housing quality index that encompasses a variety of housing features, studying differences based on ownership status shows that owners see significantly fewer housing quality issues than renters. To determine other factors driving such difference, linear regression models were fitted and featured additional significant variables, such as borough (i.e. section of the city) and number of stories. It was determined that the householder income cannot explain varying housing conditions. However, the model shown above comprised of economic, demographic, and physical variables could only explain about 26.5% of the variance in the HQI, implying that there may be other omitted factors not included in the analysis.

This statistical framework for analyzing housing conditions was used to investigate how demographic, economic, and geographic factors affected housing quality differently for renters versus homeowners. The model showed that while the typical New York renter faces more housing issues than owners, the opposite was true in the borough of Manhattan. Within the same borough, factors such as the height of the building are associated with different effects, with owners facing much less issues than renters in buildings with 6 to 10 stories. Additional research might utilize this variation to understand which types of buildings promote positive health for occupants, potentially influencing future housing development.

Finally, the statistical framework presented in this study provides a method to quantify the changes in impact of different factors on housing quality over time. Those different results lead to valuable insights for cities. For example, living in Manhattan suddenly became associated with a lower HQI compared to living in the Bronx after the 2008 crisis. Similarly, living in buildings with more than 21 stories has increasingly become a source of numerous housing issues. Observing and predicting trends in the factors that affect housing quality can help inform policies to curb the detrimental consequences of poor housing conditions.

6 Data availability

The datasets analysed during the current study are available in the Dataverse repository: https://doi.org/10.18130/V3/PMWP2C. These datasets were derived from the following public domain resources:

https://www.census.gov/data/datasets/2017/demo/nychvs/microdata.html.

Acknowledgements We acknowledge the help of the UVA Statistics Department, specifically Tianxi Li and Jordan Rodu, for their help and guidance. We want to thank the American Statistical Association for organizing the Data Challenge Expo. Finally, we would like to thank the city of New York City for providing the data used for this project.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Corman H, Mocan N (2005) Carrots, sticks, and broken windows. The Journal of Law and Economics 48(1):235–266, DOI 10.1086/425594, URL https://doi.org/10.1086/425594, https://doi.org/10.1086/425594

Dunn JR, Hayes MV (2000) Social inequality, population health, and housing: a study of two vancouver neighborhoods. Social Science Medicine 51(4):563 – 587, DOI https://doi.org/10.1016/S0277-9536(99)00496-7, URL http://www.sciencedirect.com/science/article/pii/S0277953699004967

Evans J, Hyndman S, Stewart-Brown S, Smith D, Petersen S (2000) An epidemiological study of the relative importance of damp housing in relation to adult health. Journal of Epidemiology & Community Health 54(9):677–686, DOI 10.1136/jech.54.9.677, URL https://jech.bmj.com/content/54/9/677, https://jech.bmj.com/content/54/9/677.full.pdf

Krieger J, Higgins DL (2002) Housing and health: Time again for public health action. American Journal of Public Health 92(5):758–768, DOI 10.2105/AJPH.92.5.758, URL https://doi.org/10.2105/AJPH.92.5.758, pMID: 11988443, https://doi.org/10.2105/AJPH.92.5.758

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist 18(1):50–60, DOI 10.1214/aoms/1177730491, URL https://doi.org/10.1214/aoms/1177730491

Neckerman KM, Garfinkel I, Teitler JO, Waldfogel J, Wimer C (2016) Beyond income poverty: Measuring disadvantage in terms of material hardship and health. Academic Pediatrics 16(3, Supplement):S52 – S59, DOI https://doi.org/10.1016/j.acap.2016.01.015, URL http://www.sciencedirect.com/science/article/pii/S1876285916000310, child Poverty in the United States

NYCHPD (2017) New york city housing and vacancy survey. URL https://www.census.gov/programs-surveys/nychvs/data/datasets.html

Wilson JQ, Kelling GL (1982) Broken windows. Atlantic monthly 249(3):29–38