Spring 5-3-2024

A Spatial Decision Support System for Rent Estimation of Retail Spaces in Manhattan Using Geographically Weighted Regression and Spatial Regression

Andie M. Migden Miller

A Spatial Decision Support System for Rent Estimation of Retail Spaces in Manhattan Using Geographically Weighted Regression and Spatial Regression

by

Andie Migden Miller

Submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Geoinformatics, Hunter College
The City University of New York

2024

May 3, 2024 — Dr. Jochen Albrecht

Date — Thesis Sponsor

May 3, 2024 — Dr. Shipeng Sun

Date — Second Reader

i. Abstract

This report outlines an automated, three-phase Spatial Decision Support System that creates models to estimate rent of retail spaces across Manhattan. First, enrich data with predictors using an ArcPy-based custom geoprocessing tool. Second, optimize spatially aware neighborhood-level models by combining GWR, spatial regression, and non-spatial regression in RStudio. Finally, visualize results in an Esri-based WebApp.

Keywords:

Geographically Weighted Regression, Retail Commercial Real Estate, Neighborhoods, SDSS

ii. Acknowledgements

All my thanks to my wife, Jess, for her unwavering support throughout my journey in completing this degree and thesis. I extend my thanks to Dr. Albrecht, Dr. Sun, and Dr. Ahearn for equipping me with the essential tools and knowledge to develop this SDSS. Additionally, a special thank you to my colleagues at work, William Parra, Jeremy Singh, and Kelsey Troyer, whose collaborative spirit and supportive presence have been invaluable in reinforcing my learning and contributing to a positive team environment.

Table of Contents

iii. List of Figures

iv. List of Tables

1. Introduction

Manhattan is one of the most complex retail markets due to its unique neighborhoods, the people that live and work there, as well as the density and variety of existing retail stores. In general, property prices are a product of many factors, of which location is arguably the most important (Cellmer, 2013). Similarly, rent prices of retail spaces are determined by their surroundings. This includes the rent of nearby similar spaces known as *comparables*, the mix of neighboring retailers known as *co-tenancy*, proximity to transit, demographics, among others. In Manhattan's urban market, however, the space that the retailer occupies must also be considered when determining the rent value. This includes the frontage, which is the width of the space along the street, ceiling heights, whether the space is built out for a food or dry user, and countless others. This echoes hedonic price theory, often used to value real estate, which is defined by Koschinsky et al. (2012) and Tomal and Helbich (2022) as valuing an object based on its utility-bearing characteristics related to structural, social and neighborhood, and location. Accurate rent price valuations are essential to make sound real estate decisions. In the private sector, when negotiating a deal with clients, having a statistically backed spatial analysis solution to justify a pricing strategy can be a key differentiator in favor of your business. On a broader city and state level, accurate assessments result in equitable property taxes and sound economic policy.

Neighborhoods play a key role in rent valuation, where each neighborhood exhibits its unique appeal. They are broadly categorized based on their dominant features: residential, retail, or office spaces, each with its own distinctions. Neighborhoods dominated by offices tend to attract businesses catering to commuters and the local daytime population, creating a vibrant atmosphere during work hours. On the other hand, residential areas are tailored to meet the needs of the local population, offering a retail mix more catered to those needs. Finally, retail hubs like SoHo and the Upper East Side's Fifth Avenue adapt to the atmosphere of their surroundings. The quaint cobblestone streets of SoHo contrast sharply with the luxury high-rises along Fifth Avenue, highlighting the diverse neighborhood personalities across Manhattan. This variation plays a significant role in shaping rent strategy throughout the city. Figure 1 illustrates the neighborhood boundaries according to the New York City Primary Land Use Tax Lot Output (PLUTO), with adjustments made to encompass retail areas such as Union Square and Little Italy.

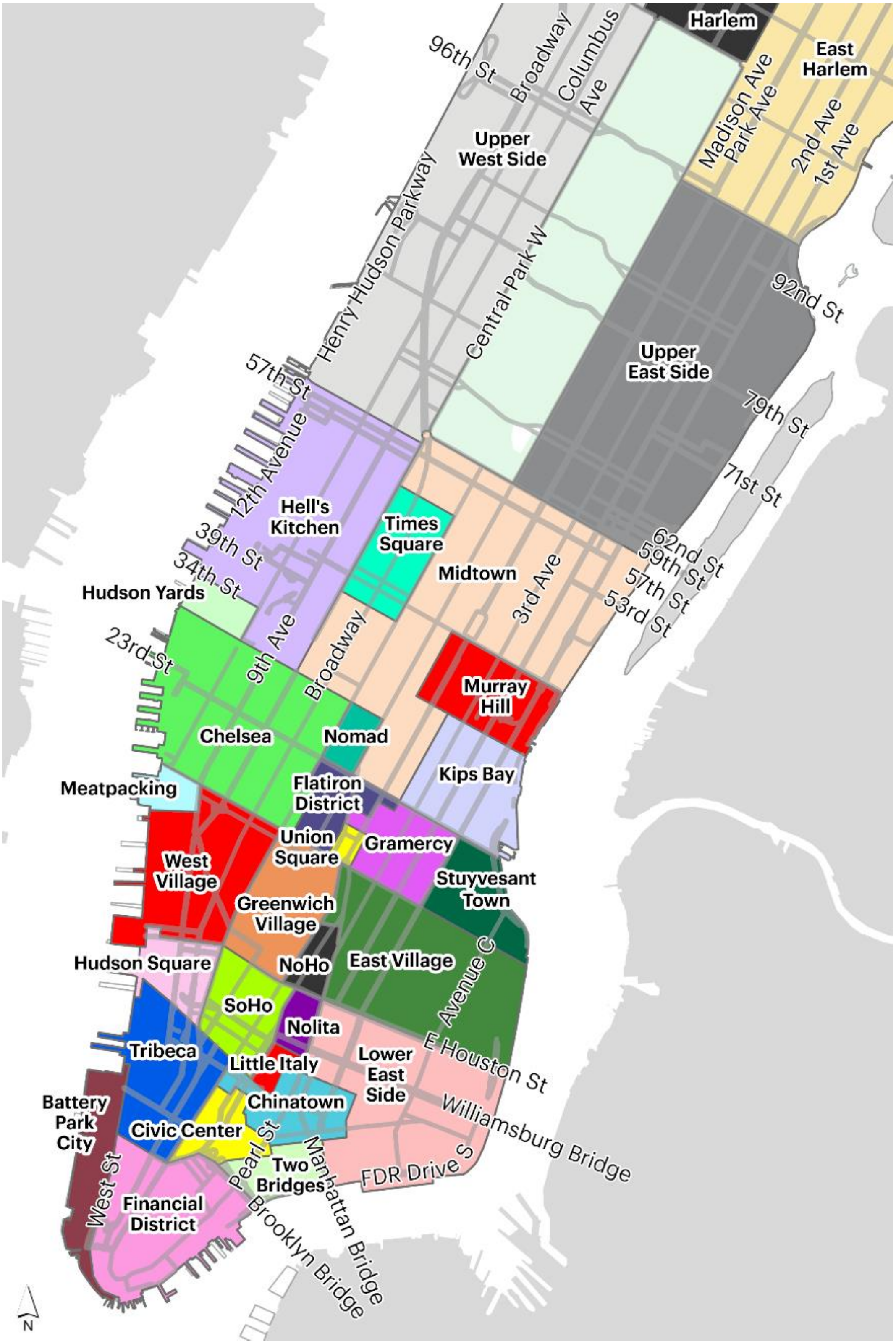


Figure 1 Manhattan neighborhoods, adapted from PLUTO.

Diverse Manhattan neighborhoods necessitate flexible rent valuation methods. A single rent prediction model is as limiting as summarizing Manhattan with one number. Consider Manhattan's 2023 median household income and average assessed land value at $89,885 and $1.3 million, respectively (New York City Department of City Planning, 2023). This is helpful for a countywide snapshot, yet insufficient for understanding neighborhood trends. Rent models therefore need to incorporate both similarities and differences across areas, spatial effects known as spatial autocorrelation and heterogeneity. Spatial autocorrelation refers to the tendency of areas close together to exhibit similar values, while spatial heterogeneity acknowledges the variation in values due to differing neighborhood characteristics. Therefore, rent valuation in Manhattan demands a neighborhood-specific set of factors, requiring the exploration of numerous variables to identify those essential for accurate rent estimation.

This paper presents a streamlined Spatial Decision Support System (SDSS) to optimize neighborhood-level rent estimation models in three broader phases. First, a custom geoprocessing tool configured with Esri’s ArcPy enriches data with predictors. Second, an RStudio script optimizes spatially aware neighborhood-level hedonic models using a combination of Geographically Weighted Regression (GWR), spatial regression, and ordinary least squares (OLS) regression. Finally, the results are visualized in an Esri-based web application where the user can drop a point and estimate rent. The paper is structured by reviewing previous studies, outlining the conceptual model, describing the methodology and data sources, analyzing results, discussing limitations, and finally, summarizing conclusions, discussion, and further improvements to the study.

2. Literature Review

Spatial econometric models have enhanced traditional hedonic modeling over the last four decades. This literature review highlights the reason to incorporate spatial effects into rent estimation, examples of ways to do this, and the importance of the neighborhood for rent decisions. The review includes examples of GWR, Multiscale Geographically Weighted Regression (MGWR), and other spatial regression methods. The majority of the below are studies within the residential sector rather than retail focused.

2.1 Why should spatial effects be incorporated into rent value?

"Real estate data and spatial statistics naturally complement each other" (Pace et al., 1998, p. 10). The basic hedonic model assumes that property value is a function of structure, social and neighborhood characteristics, and location characteristics; all of which exhibit spatial effects. Researchers have argued that the OLS-based hedonic pricing model is sufficient because spatial effects can be effectively modeled through location-based predictor values such as proximity to main highways or subways, distance to city centers, shopping centers, or employment centers (Valente, 2005; Koschinsky et al., 2012). Others have made the case that there is a more direct need to handle them (Anselin, 1998; Fotheringham et al., 2002; Pace et al., 2009; Furtado, 2011).

The incorporation of spatial effects into property valuation addresses a significant limitation of traditional hedonic price modelling: that it omits spatial effects (Anselin, 1998, Fotheringham et al., 2002, Tomal & Helbich, 2022). Spatial data violates the assumption of classic non-spatial statistics, that observations are random and independent (Pace et al. 2009). This assumption is problematic because predicting attributes such as building quality, accessibility, shared local amenities (Ismail 2006), and space characteristics depend on location.

For spaces sharing similar characteristics, rents tend to be similar, with higher rents in some areas and lower rents in others. Using traditional regression methods on data that has spatial autocorrelation leads to over- or under-estimation in the relationship between predictors and rent prices across the study area (Fotheringham et al., 2002). Plotting the residuals of a non-spatial model on a map reveals their spatial autocorrelation, indicating that non-spatial regression leads to biased and inaccurate models of the average relationship between a particular attribute and a property value.

2.2 How can spatial effects be incorporated into rent value?

2.2.1. Spatial Lag and Spatial Error Models

Spatial lag and spatial error modeling incorporate spatial autocorrelation by adding a spatial weights matrix to a regression model (Anselin 1998, Getis 2009). Spatial error autocorrelation exists when error terms are correlated across space and spatial lag autocorrelation exists where the explanatory variable terms are correlated across space. There is no clear way to determine the proper weight matrix, however, Ismail (2006) echoes Gillen et. al, 2001 to incorporate the concept of accessibility. Cellmer (2022) tests the effectiveness of his spatial weight matrices using k = 5 nearest neighbors, among other additional distance-based weight matrices, whereas Fotheringham et al. (2002) uses k = 10. The below models show how spatial effects are accounted for using the basic spatial lag model (top) and error model (bottom),

$$y = \beta X + \lambda W\gamma + \varepsilon$$

$$y = \beta X + \lambda We + u$$

where β represents the coefficients for the intercept and predictor variables, X is the predictor variables, $W\gamma$ is the spatial lag of the dependent variables with W as the spatial weights matrix and γ as the value of the dependent variable for the neighbors, and λ as the coefficient for the spatial lag term. Similarly, We is the spatially autocorrelated error term and λ is the error components influenced by neighboring points. Any remaining non-spatial error is captured by u. The second model below incorporates a condensed example of Fotheringham's (2002) spatial lag for London housing prices.

$$P_i = \mathrm{K} + \mathrm{p}P_i{}^* + \alpha_1 FLRAREA_i + \alpha_2 BLDGPOSTW_i + \alpha_3 TYPDETCH_i + \mathcal{E}_i$$

where P_i represents the price a house sold for at location i. K and p parameters are estimated, as well as P_i* as the average of the prices of the 10 nearest houses. The parameter p is a measure of spatial autocorrelation for housing prices to describe the average degree of spatial autocorrelation across the area. FLRAREA is the floor area of the property in square meters, BLDGPOSTW is a binary variable where 1 is if the property was built between 1940 and 1959 and a 0 otherwise, and TYPDETCH is another binary variable where 1 is if the property is detached, 0 otherwise. This model allows us to understand the extent that an attribute is spatially autocorrelated given the distribution of other attributes. However, the spatial lag and spatial error models are still only global ways to incorporate spatial effects into hedonic price modelling.

2.2.2. Geographically Weighted Regression

Geographically weighted regression incorporates local spatial relationships into what would otherwise be one global regression framework. It allows geographically varying relationships to be modelled through spatially varying parameter estimates rather than globally through the error term (spatial error) or the predictor variables (spatial lag). GWR determines an optimized distance bandwidth, i.e., number of neighboring points, to use in the regression on each point (Fotheringham et al., 2002). The distance bandwidth inversely weighs each explanatory value of the neighbor based on distance. When calibrating multiple GWR models, the lowest Akaike Information Criterion (AIC) score is the better fit (Fotheringham et al., 2002).

In his study, GWR highlights spatial autocorrelation within London houses sold in 1990. Echoing the spatial lag model above, Fotheringham produces a surface of local estimates of spatial autocorrelation in housing prices as a function of property-related attributes using the below model. This is a condensed version of the original model.

$$P_i = К(u_i, v_i) + р(u_i, v_i)P_i^* + \alpha_1(u_i, v_i)FLRAREA_i + \alpha_2(u_i, v_i)BLDGPOSTW_i + \alpha_3(u_i, v_i)TYPDETCH_i + Ɛ_i$$

Plotting the variables' coefficients, residuals, and local R-squared values on a map shows how they vary across space and how nearby areas form clusters of similarities. Multiscale Geographically Weighted Regression (MGWR) takes this idea a step further where the bandwidth adjusts for each predictor variable with the argument each influences the dependent variable at a different scale (Oshan, 2019). Oshan also scales predictors so that coefficients can be looked at with the same scale.

2.2.3. Examples of Spatial Model Applications to Real Estate

Numerous examples demonstrate the application of GWR, spatial regression, and other advanced methods as techniques to mitigate the spatial effects in modeling the value of residential and commercial real estate. This paper contributes further by exploring the nuances of rent values for retail real estate, an area that has been far less studied. Basu and Thibodeu (1998) compare OLS and spatial models to estimate housing prices in eight different submarkets in Texas, Zhang, et al. (2014) examine an improved spatial error model for commercial properties in Shenzhen, China; and Valente (2005) explores a spatial error model across different markets in United States. Kiely and Bastian (2019) compare several machine learning algorithms using spatial lag, semi-spatial, and non-spatial models to predict gentrification. In all studies, the results of using the spatial model are at least in part better than the non-spatial version.

2.3 The Importance of Neighborhoods in Real Estate

Neighborhoods influence both the real estate landscape and the social fabric of communities. Neighborhoods are social constructs shaped by the combination of architectural styles, demographic profiles, public services, environmental considerations, accessibility, and even symbolic meanings (Golab 1982). The characteristics and amenities of a neighborhood significantly affect rent prices and marketability. Prospective tenants consider the mix of surrounding co-tenants, demographics, proximity to services such as hospitals and schools, as well as the overall aesthetic of an area as key factors in their deal making decisions. Landlords consider these factors when trying to determine the best use type to maximize occupancy and returns. This complex interaction between real estate dynamics and social cohesion highlights the essential role neighborhoods have in deal making.

3. Methodology

"…geographers have long…debated whether, and in what circumstances, we might expect geographic processes to remain stable, thereby allowing the study of those processes to be replicable. A schism arose…with adherents of a "place-based", largely humanistic, idiographic geography on one side and those who believe in regularities across space can be reliably identified and measured through generally quantitative, nomothetic approaches on the other… The latter…can be identified from quantitative modeling and spatial analytics." (Fotheringham, Li, 2023, p. 1)

This study involves a streamlined Spatial Decision Support System (SDSS) to optimize neighborhood-level rent estimation models in three broader phases. The first step uses an ArcPy-based geoprocessing tool to enrich the point layer of retail spaces with predictor variables. The second step uses RStudio to calibrate the optimal GWR model and the optimal neighborhood-level models. GWR determines which predictor variables contribute to the value of a space and to what extent. Additionally, it uses local R-squared values to identify neighborhoods where the relationships between these predictor variables and spaces are similar. OLS, spatial lag, and spatial error models are then tested on each of these neighborhood subsets. The RStudio code sifts through a user-defined number of combinations accounting for multicollinearity, variable significance, outliers, and heteroskedasticity to provide a best-fit model. The third step of the SDSS visualizes the results in an interactive web application using Esri's Experience Builder and Survey 123. Figure 2 below presents a high-level conceptual model of the SDSS.

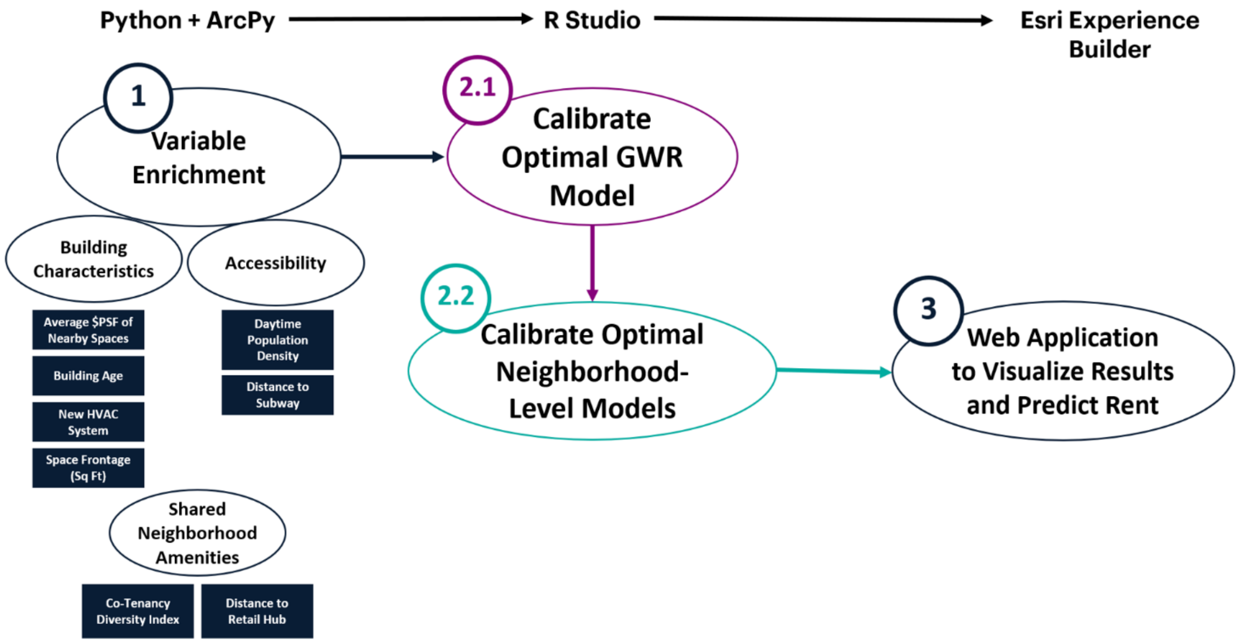


Figure 2 High-level conceptual model of the entire process to estimate rent prices.

3.1 Unique Factors that Contribute to Rent of Retail Spaces

Rent determination involves a combination of predictors involving both space and location. In-space characteristics include factors such as use type (food or dry user), ceiling heights, building age, upgraded HVAC systems, total square feet, and frontage. Location and neighborhood factors include the rent of nearby comparable spaces, transit proximity, and surrounding co-tenancy. This conceptual model is pictured in Figure 3.

This study introduces a novel co-tenancy diversity index score to quantify the impact of surrounding retailers, with its calculation method illustrated in Figure 4. First, retailers are organized into broader umbrella categories. For each space, the number of nearby retailers within each category is counted and then divided by the maximum count observed for that group near any other space. These ratios are weighted and summed to calculate the final co-tenancy score for each space. High foot traffic-driving retailers Trader Joe's, Whole Foods, and Target are labeled under their own *special* group to ensure they contribute more substantially to the co-tenancy diversity index score than other retailers contribute. A more detailed discussion on co-tenancy is in section 3.1.1. It is important to note, however, that the retail real estate market is inherently imperfect where there are immeasurable extrinsic social factors of rent negotiation. This includes both high level economic and market trends, as well as nuanced elements such as relationships among stakeholders, landlord willingness to take on the costs of tenant improvements, and general preferences of a landlord or tenant.

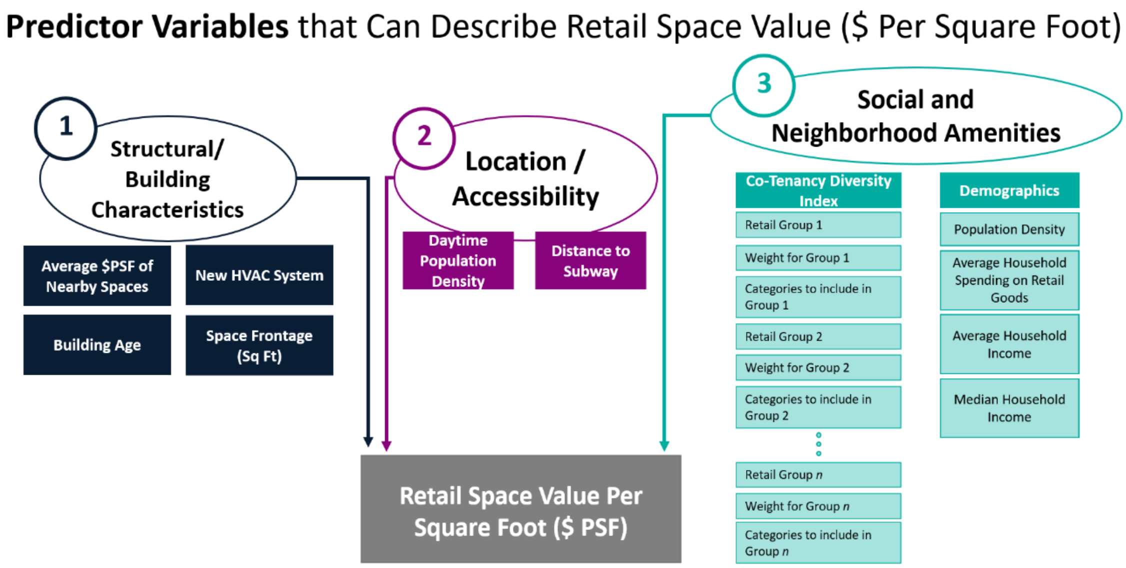


Figure 3 Conceptual model that rent prices are composed of structural characteristics, location, and social and neighborhood amenities.

$$index\ score = \sum \frac{No.\,Retailers\ within\ 5\ Mins}{Denominator} x\ Weight * 100$$

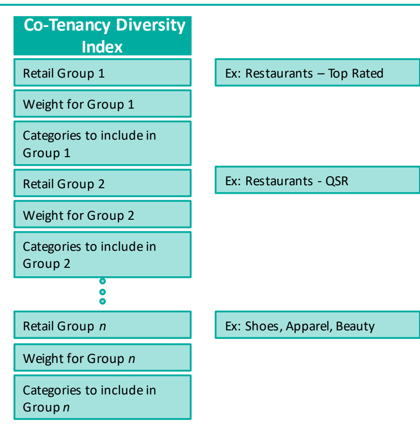


Figure 4 The co-tenancy diversity index score compares how each observation performs relative to each other.

3.1.1. The Co-Tenancy Diversity Index Score

The creation of a co-tenancy diversity index score is, by nature, subjective. It should categorize and assign weights to retail groups based on their significance, aiming to reflect the health and diversity of a retail corridor. The higher the score, the stronger the retail corridor. This is the first way of viewing co-tenancy. The second perspective of co-tenancy focuses on the mix of retailers relevant for the neighborhood's intended purpose. As mentioned in the introductory section, neighborhoods serve different purposes based on their dominant theme: retail, office, or residential, with aspects of each often overlapping. The retail theme can be further broken down to reflect regional shopping and trendy shopping areas, resulting in a total of four themes. Regional shopping areas are big-box focused (i.e., Times Square) whereas trendy shopping centers have both local and name-brand tenants (i.e., SoHo). In this second perspective, all retail groups are considered equally important (and therefore weighted equally) because a neighborhood has a healthy co-tenancy as long as serves its intended population. This approach emphasizes the overall functionality of a neighborhood. Therefore, the higher the score in co-tenancy perspective two, the more blended a neighborhood, taking characteristics from all four themes. This flexibility acknowledges the dynamic nature of a neighborhood. In both co-tenancy perspectives, each retailer use type can only be assigned to one broader category so that it is not double counted. Table 1 illustrates the first approach to co-tenancy, while Table 2 presents the alternative perspective, with each dominant theme labeled accordingly.

A survey was created for this study to categorize use types into favorable co-tenancy categories. All six survey responses from real estate professionals reflect these two ways of thinking about co-tenancy, with the survey template in Appendix A. The survey responses and demographic variable selection echo both abductive reasoning and theory-free reasoning (Miller

& Goodchild, 2014). The analyst needs to understand which parameters are likely important for rent value as well as which mix of nearby retailers has a bigger influence on rent.

Table 1 Co-tenancy perspective one considering the overall health of any given retail corridor.

Group 1 - 40%	Group 2 - 30%	Group 3 - 15%	Group 4 - 10%	Group 5 - 5%
grocery – 0	apparel - 1	apparel - 0	art - 1	florist - 0
grocery – 1	drugstore - 1	beauty_and_hair - 0	book_store - 0	other_retailer - 0
special – 1	home_goods_store - 1	beauty_and_hair - 1	convenience_store - 0	other_retailer - 1
	hospital - 1	book_store - 1	department_store - 0	bicycle_store - 0
	office - 1	convenience_store - 1	electronics_store - 0	
	restaurant - 0	department_store - 1	experiential - 0	
	restaurant - 1	drugstore - 0	furniture_store - 0	
	Restaurant - bakery - 0	electronics_store - 1	jewelry_store - 0	
	Restaurant - bakery - 1	experiential - 1	jewelry_store - 1	
	Restaurant - casual - 0	finance - 1	liquor_store - 0	
	Restaurant - casual - 1	furniture_store - 1	liquor_store - 1	
	Restaurant - coffee - 0	gym - 0	music - 0	
	Restaurant - coffee - 1	gym - 1	music - 1	
	Restaurant - fast casual - 1	hardware_store - 0	pet_store - 0	
	Restaurant - qsr - 1	hardware_store - 1	Restaurant - dessert - 1	
	Restaurant - top rated - 0	home_goods_store - 0	Restaurant - fine - 1	
	Restaurant - top rated - 1	hotels - 1	Restaurant - juice - 1	
	schools - 1	medical - 0	shoe_store - 0	
	shopping_mall - 1	medical - 1		
		pet_store - 1		
		Restaurant - bar - 1		
		Restaurant - qsr - 0		
		shoe_store - 1		
		veterinary_care - 0		
		veterinary_care - 1		

Table 2 Co-tenancy perspective two considering the purpose of the neighborhood.

Regional Shopping	Trendy Shopping	Residential	Office
art - 1	apparel - 0	beauty_and_hair - 0	drugstore - 1
book_store - 1	apparel - 1	bicycle_store - 0	finance - 1
department_store - 1	beauty_and_hair - 1	convenience_store - 0	gym - 1
electronics_store - 1	hotels - 1	convenience_store - 1	office - 1
experiential - 1	jewelry_store - 0	drugstore - 0	Restaurant - fine - 1
grocery - 1	jewelry_store - 1	florist - 0	Restaurant - qsr - 0
home_goods_store - 1	Restaurant - bakery - 1	furniture_store - 1	Restaurant - qsr - 1
music - 1	Restaurant - coffee - 0	grocery - 0	Restaurant - top rated - 1
pet_store - 1	Restaurant - dessert - 1	gym - 0	
Restaurant - bar - 1	Restaurant - juice - 1	hardware_store - 0	
Restaurant - fast casual - 1	Restaurant - top rated - 0	hardware_store - 1	
shopping_mall - 1	shoe_store - 0	liquor_store - 0	
special - 1	shoe_store - 1	liquor_store - 1	
		medical - 0	
		medical - 1	
		pet_store - 0	
		pet_store - 1	
		restaurant - 0	
		restaurant - 1	
		Restaurant - bakery - 0	
		Restaurant - casual - 0	
		Restaurant - casual - 1	
		Restaurant - coffee - 1	
		schools - 1	
		veterinary_care - 0	
		veterinary_care - 1	

3.2 GWR and Neighborhood-Level Model Optimization

Step two of the SDSS automates a two-tiered approach for rent estimation. It allows the analyst to explore and understand the interaction of parameters, identify spatial dependencies, and visualize patterns that would be otherwise overlooked. First, it generates an optimal GWR model and uses the resulting local R-squared values to create data-derived neighborhoods. This provides a way to quantifiably define neighborhood boundaries, allowing them to evolve dynamically as the underlying data shifts over time. This contrasts with existing fixed and rigid neighborhood boundaries as defined by PLUTO. Second, mirroring the code logic used in step one, it optimizes an OLS, spatial error, and spatial lag model on the subset of points within each newly defined neighborhood. The final neighborhood-level model is what can be used to predict rent prices.

3.2.1. Create Neighborhoods with GWR-Derived Local R-squared Values

The methods outlined in this paper create neighborhood boundaries based on quantitative data rather than using existing neighborhood boundaries such as well-known retail corridors Times Square or SoHo. GWR explores the spatial variation that exists throughout Manhattan and identifies areas of coherence across different predictor variables. With traditional OLS all the features in the study area are used to calibrate a model. With GWR each feature gets its own separate equation. That equation is calibrated using only the data from your neighboring features, defined by an optimal bandwidth. Spaces with similar local R-squared values clustered together are considered their own GWR-defined neighborhood. This reflects Fotheringham and Li’s 2023 quote at the top of section 3 where the R-squared values identify regularities across space, effectively removing or reducing spatial autocorrelation from within the sub-region. GWR’s local R-squared values also provide a summary statistic to delineate spatial regimes. However, these R-squared values are not being used in the statistical sense and do not suggest uniformity across all similar statistical moments. For example, an R-squared value of 0.6 in Chelsea and in Times Square does not mean it is the same spatial regime. The combinations of the other variables; while having the same explained variance and leading to the same outcome, are each contributing in a different way in Chelsea and in Times Square. Therefore, these two neighborhoods are separate. In contrast, similar local R-squared values in neighboring areas would be merged into one spatial regime.

3.2.2. Apply OLS, Spatial Lag, and Spatial Error Models on Data-Derived Neighborhoods

Identifying coherence amongst spaces significantly mitigates or eliminates spatial autocorrelation. This allows for the calibration of global, non-spatial hedonic models tailored to each neighborhood. Should any statistically significant spatial autocorrelation remain, as indicated by a Moran's I test on the residuals, it can be effectively managed with either a spatial lag or spatial error model. This approach ensures that each neighborhood benefits from an optimized model, featuring a tailored combinations of predictors and coefficients.

3.3 Visualize Results

The results of this paper can be viewed in an Esri-based interactive map. Neighborhood-level models are stored in a tessellated hexagon layer and predictors are stored at the block group level. The user can drop a pin to calculate a rent estimation based on the output from the SDSS. The interactive map equips stakeholders with a visual tool to facilitate informed, data-driven decisions in real estate transactions.

Data Sources

4.1 Study Area, Data Sources, and Data Preparation

This study uses proprietary company data on blended rent, in price per square foot, for 2,673 spaces on closed deals since January 1, 2010, in New York County only, and below 96th Street.

4.1.1. Retail Spaces with Rent Information

The original data set included over 7,000 data points of deals done since the year 2000. After removing duplicates and processing data, 3,212 data points remained with 2,673 used in the study, removing the one percent of high and low extreme rent values. The final dataset includes all rent deals 2010 and later. Original columns include those listed below.

```
['index', 'BLG #', 'STREET', 'CROSS STREETS', 'SUBMARKET', 'SIZE',
       'ANNUAL RENT', 'PSF RENT', 'LEASE TERM', 'TENANT', 'USE TYPE',
       'COMMENTS', 'DATE EXECUTED', 'DATE ENTERED', 'SOURCE',
       'YEAR'],
```

Processing the data includes adjusting rent prices to 2023 values using the Bureau of Labor Statistic’s Consumer Price Index (CPI) (Bureau of Labor Statistics, 2024) and obtaining additional square footage information. The CPI is a measure of the average change over time in prices paid by urban consumers for market goods and is similarly applied in property assessments for taxes. First, calculate the percent change in CPI between the current year and the year the rent is in. Next, multiply that percent change by the original rent. Lastly, add that value to the original rent. The result is the new price in 2023 values. The below is the basic percent change formula for CPI.

$$\% \, Change = \frac{(\text{Current Index} - \text{Base Index})}{\text{Base Index}}$$

After processing, the data includes the below columns.

```
['index', 'FULL ADDRESS', 'BLDG NUM', 'STREET', 'CROSS STREETS',
'SUBMARKET', 'SIZE', 'ANNUAL RENT', 'PSF RENT', 'LEASE TERM',
'TENANT','USE TYPE', 'COMMENTS', 'DATE EXECUTED', 'DATE ENTERED', 'YEAR',
'TOTAL SF', 'GROUND SF', 'Lower SF', 'PSF RENT ADJ']
```

Isolating columns for total square feet, ground square feet, and lower square enhances data integrity when comparing those values with *PSF RENT* and *ANNUAL RENT* columns. It is important to note that the accuracy of this data is contingent upon the quality of its input. Consequently, it is not possible to verify each rent price and related attributes with absolute certainty. Rent prices are clustered together, illustrated in Figure 5.

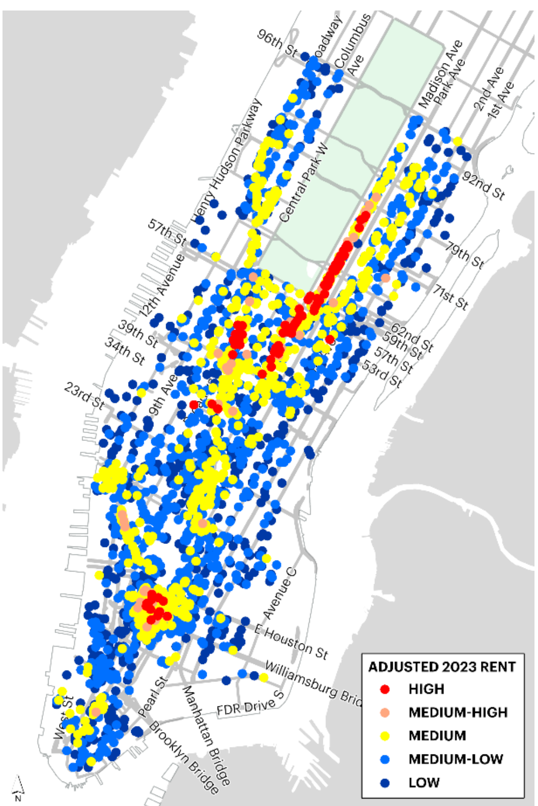


Figure 5 Low, medium, and high value retail spaces, plotted.

4.1.2. Retailer Points Data

Retailer point data was collected in January 2024 from Google's *Places Search* API, which gathers information on both local and nationally branded retailers, and from Chain XY, a privately-owned company that specializes in providing retailer point data for national brands. The *Places Search* API takes user-defined search points and pulls up to 60 of Google’s decidedly most relevant retailers per retail category, per point, per number of user-generated API keys. This study developed a *Google Points Cleanup* python script to clean up the data that was pulled from the *Places Search* API (Migden-Miller, 2024). This script goes through each of the 44 use type categories, moves retailers to mis-assigned categories as well as removes unwanted retailers. For example, Google mistakenly labels the clothing retailer *Suit Supply* as a *drugstore*. There are many examples of this throughout the dataset. The script also removes less relevant categories such as roofing contractors and real estate agencies. It conflates the two retail data sources together and removes duplicates in favor of the Chain XY dataset. Each category is coded with a 0 or a 1 at the suffix. 0 is if the retailer is local and 1 if the retailer is a national brand name. This is important to consider when building the co-tenancy diversity index score. In general, an accurate retail dataset is crucial to include in the model due to Manhattan’s wide variety and density of retailers and their impact on rent.

4.1.3. Additional Predictors

Demographic information such as population density, median household income, average household spending on retail goods, and daytime population density are sourced from Esri’s 2023 Living Atlas Layer at the block group level. Subway entrance points are from the NYC Metropolitan Transportation Authority (MTA).

5. Applying Methods to Estimate Rent using an Automated SDSS

5.1 Custom ArcPy Geoprocessing Tool to Enrich Real Estate Data

Automation minimizes human error when testing multiple variable combinations and updating data. Implementing this approach, an ArcGIS Pro-based geoprocessing tool was developed to enrich spaces with predictor variables within up to four different trade areas. The *GWR – Create Variables* geoprocessing tool's flexible interface allows the user to select what trade area parameters are appropriate for a study area. In this case, each predictor variable is summarized within a tenth of a mile, a quarter of a mile, five-minutes walking, and ten-minutes walking. Figure 7 below shows the logic conceptually, whereas Figure 8 shows the GUI. This automation is especially needed to test out different versions of the co-tenancy diversity index score, which is tedious and error-prone to calculate manually. This geoprocessing tool is a powerful way to enrich rent data with key variables.

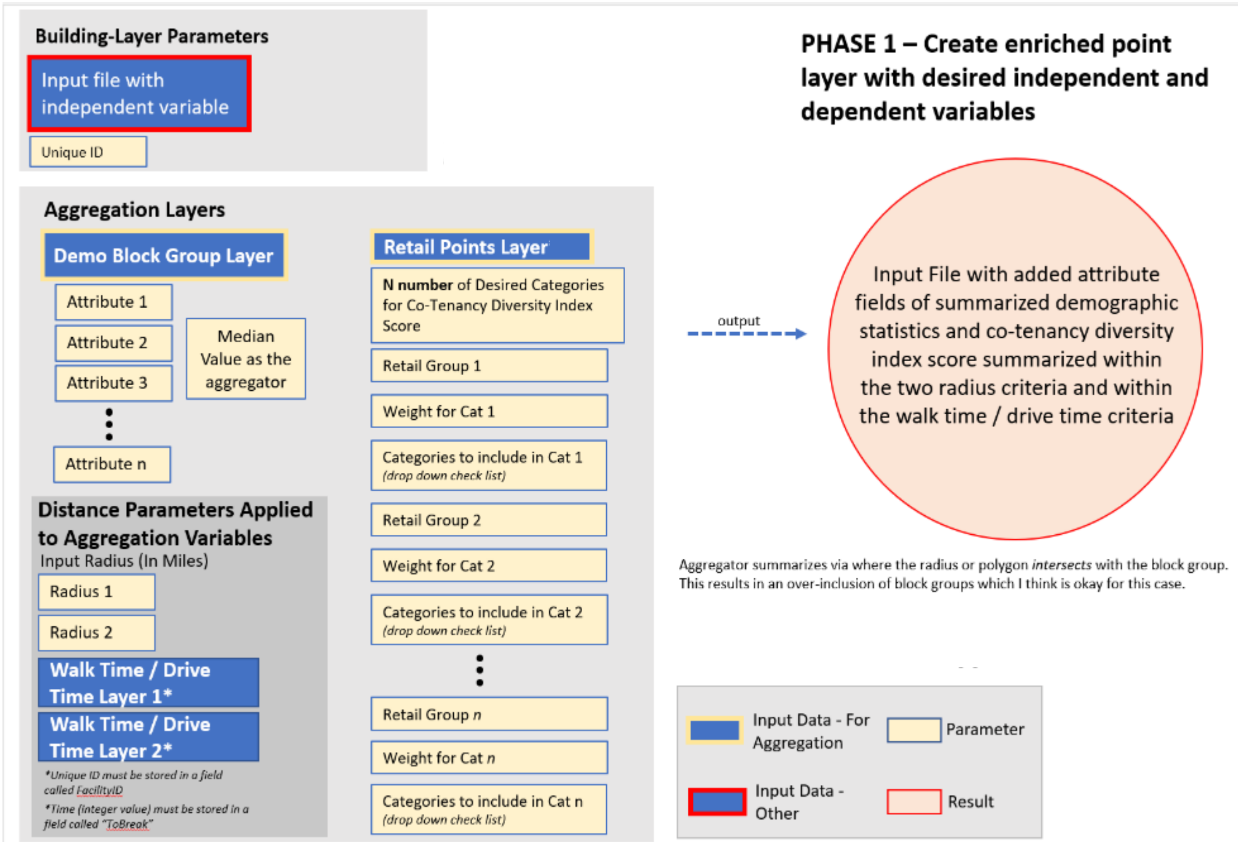
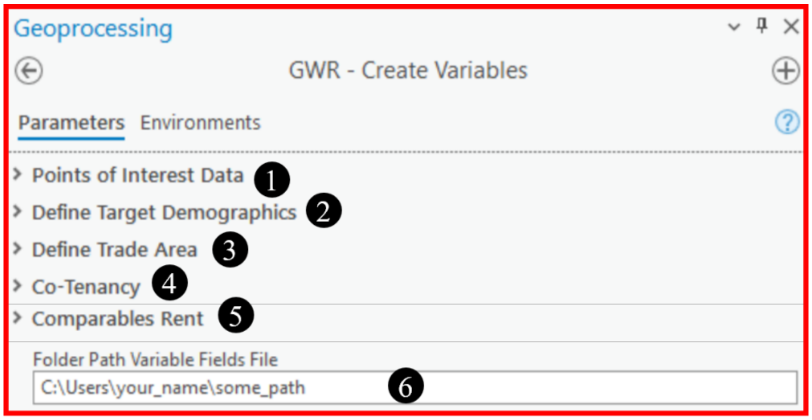
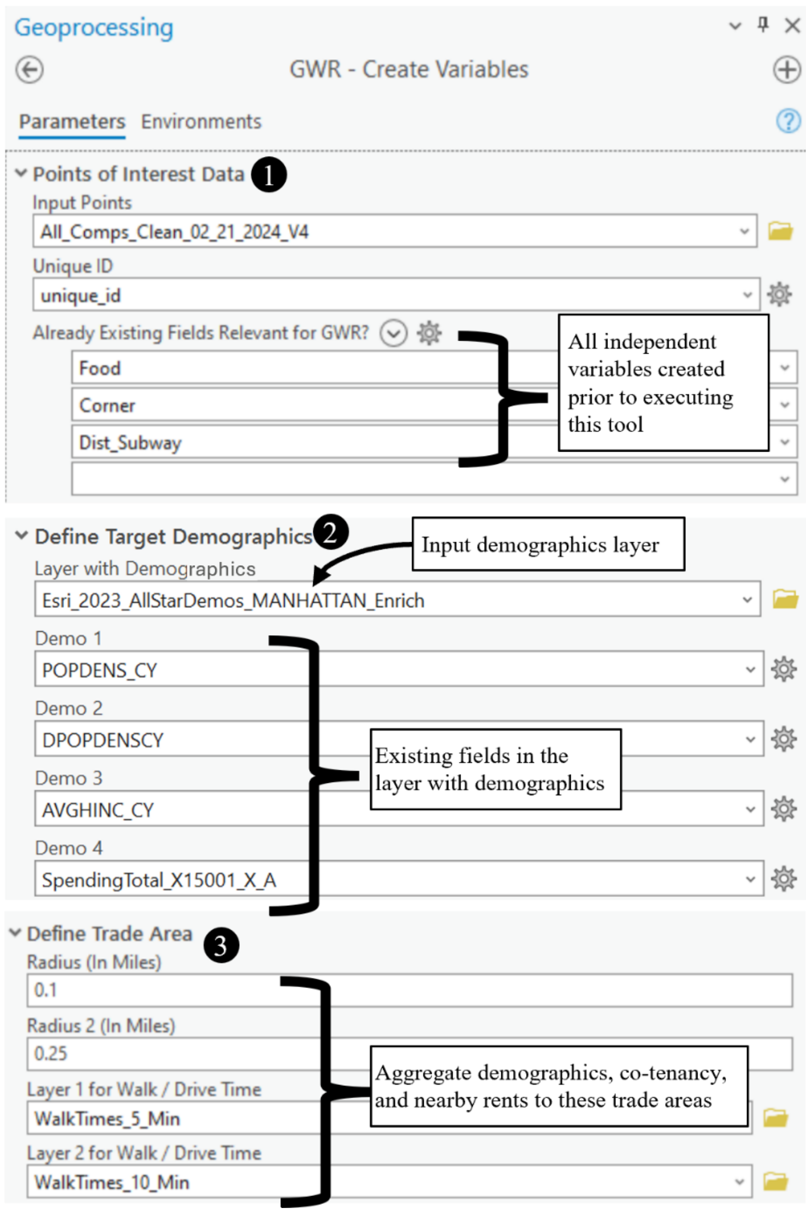


Figure 6 Conceptual model of the ArcPy-based GUI



Geoprocessing
GWR - Create Variables
Parameters
Environments
Points of Interest Data
1
Define Target Demographics
2
Define Trade Area
3
Co-Tenancy
4
Comparables Rent
5
Folder Path Variable Fields File
C:\Users\your_name\some_path
6



Geoprocessing
GWR - Create Variables
Parameters
Environments
Points of Interest Data
1
Input Points
All_Comps_Clean_02_21_2024_V4
Unique ID
unique_id
Already Existing Fields Relevant for GWR?
Food
Corner
Dist_Subway
All independent variables created prior to executing this tool
Define Target Demographics
2
Input demographics layer
Layer with Demographics
Esri_2023_AllStarDemos_MANHATTAN_Enrich
Demo 1
POPDENS_CY
Demo 2
DPOPDENSCY
Demo 3
AVGHINC_CY
Demo 4
SpendingTotal_X15001_X_A
Existing fields in the layer with demographics
Define Trade Area
3
Radius (In Miles)
0.1
Radius 2 (In Miles)
0.25
Layer 1 for Walk / Drive Time
WalkTimes_5_Min
Layer 2 for Walk / Drive Time
WalkTimes_10_Min
Aggregate demographics, co-tenancy, and nearby rents to these trade areas

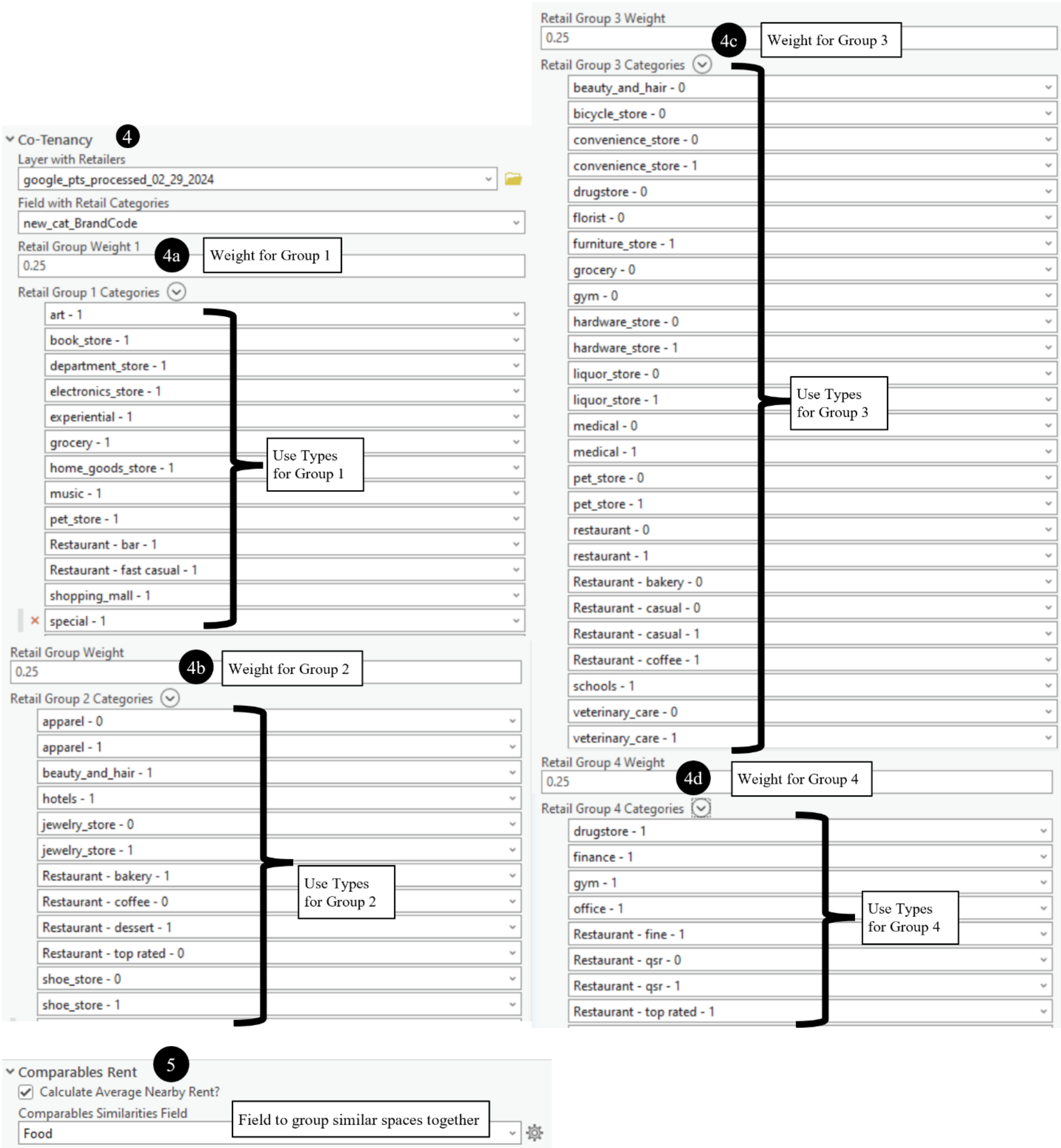


Figure 7 Custom geoprocessing tool interface called GWR - Create Variables. Top (red outline) shows the six categories of parameters, with the succeeding images revealing the full interface.

If all parameters of the geoprocessing tool are filled out, the data is enriched with 24 variables (four demographic, one co-tenancy score, and one average rent of nearby comparables multiplied by four trade areas). This number is in addition to variables created prior to executing the tool such as a previous version of co-tenancy score, distance to the nearest subway entrance, if the space is on the corner or inline, and if the space is a food or dry user. For the comparable spaces section, if the selected space is a food user, the tool only considers other nearby food-occupied spaces when calculating the average rent. Similarity, if the selected space is a dry user, only non-food-occupied spaces are considered. After the tool runs, it generates a csv file to guide the RStudio code on which variables to include in the regression analyses. Table 3 below outlines the variables that result from the tool plus additional ones used for the study, totaling 31 potential predictors.

Table 3 Predictor variables resulting from executing the custom GWR - Create Variables geoprocessing tool.

Num	Name	Description
1	RAD_0_1_mean_POPDENS_CY	Population Density (Pop. / Sq. Mi.) within 0.1 Miles
2	RAD_0_25_mean_POPDENS_CY	Population Density (Pop. / Sq. Mi.) within 0.25 Miles
3	WLK_5_mean_POPDENS_CY	Population Density (Pop / sq mi) within 5-Minutes Walking
4	WLK_10_mean_POPDENS_CY	Population Density (Pop. / Sq. Mi.) within 10-Minutes Walking
5	RAD_0_1_mean_ DPOPDENSCY	Daytime Population Density (Pop. / Sq. Mi.) within 0.1 Miles
6	RAD_0_25_mean_DPOPDENSCY	Daytime Population Density (Pop. / Sq. Mi.) within 0.25 Miles
7	WLK_5_mean_ DPOPDENSCY	Daytime Population Density (Pop. / Sq. Mi.) within 5-Minutes Walking
8	WLK_10_mean_ DPOPDENSCY	Daytime Population Density (Pop. / Sq. Mi.) within 10-Minutes Walking
9	RAD_0_1_mean_SpendingTotal_X15001_X_A	Average Household Spending on Retail Goods within 0.1 Miles
10	RAD_0_25_mean_SpendingTotal_X15001_X_A	Average Household Spending on Retail Goods within 0.25 Miles
11	WLK_5_ mean_SpendingTotal_X15001_X_A	Average Household Spending on Retail Goods within 5-Minutes Walking
12	WLK_10_ mean_SpendingTotal_X15001_X_A	Average Household Spending on Retail Goods within 10-Minutes Walking
13	RAD_0_1_mean_AVGHINC_CY	Average Household Income within 0.1 Miles
14	RAD_0_25_mean_AVGHINC_CY	Average Household Income within 0.25 Miles
15	WLK_5_mean_AVGHINC_CY	Average Household Income within 5-Minutes Walking
16	WLK_10_mean_AVGHINC_CY	Average Household Income within 10-Minutes Walking
17	RAD_0_1_CoTen_V1	Co-Tenancy Diversity Index Score within 0.1 Miles (Perspective 1)
18	RAD_0_25_CoTen_V1	Co-Tenancy Diversity Index Score within 0.25 Miles (Perspective 1)

19	WLK_5_CoTen_V1	Co-Tenancy Diversity Index Score within 5-Minutes Walking (Perspective 1)
20	WLK_10_CoTen_V1	Co-Tenancy Diversity Index Score within 10-Minutes Walking (Perspective 1)
21	RAD_0_1_CoTen_V2	Co-Tenancy Diversity Index Score within 0.1 Miles (Perspective 2)
22	RAD_0_25_CoTen_V2	Co-Tenancy Diversity Index Score within 0.25 Miles (Perspective 2)
23	WLK_5_CoTen_V2	Co-Tenancy Diversity Index Score within 5-Minutes Walking (Perspective 2)
24	WLK_10_CoTen_V2	Co-Tenancy Index Diversity Score within 10-Minutes Walking (Perspective 2)
25	RAD_0_1_Mean_PSF_RENT_ADJ	Average Adjusted 2023 Rent within 0.1 Miles
26	RAD_0_25_Mean_PSF_RENT_ADJ	Average Adjusted 2023 Rent within 0.25 Miles
27	WLK_5_Mean_PSF_RENT_ADJ	Average Adjusted 2023 Rent within 5-Minutes Walking
28	WLK_10_Mean_PSF_RENT_ADJ	Average Adjusted 2023 Rent within 10-Minutes Walking
29	Food	Binary: 1 if space is a food-user, 0 if not.
30	Corner	Binary: 1 if space is within 50 feet of an intersection (corner), 0 if not (inline)
31	Dist_Subway	Distance in feet to the nearest subway entrance

5.2 Model Optimization in RStudio

5.2.1. Install Libraries, Connect to ArcGIS Pro, Initialize Parameters, and Initial Diagnostics

Spatial processing-related R libraries such as *sf*, *sp*, *GWmodel*, *spdep*, *MASS*, and *lmtest*, are used to apply GWR, linear regression, spatial regression, and other statistical analyses. Packages *broom*, *janitor*, *tidyverse*, *ggplot*, *gridExtra*, and *mapview* are used for code syntax streamlining and spatial data visualization. The *arcgisbinding* package must be installed in tandem with correct R-ArcGIS configurations within ArcGIS Pro, shown in Figure 9 below. The R-ArcGIS Bridge allows feature class data from ArcGIS Pro to be read into RStudio (Pobuda, n.d.). The connection between ArcGIS Pro and RStudio is completed using the *arc.check_product* function. An Esri feature class is converted to a simple feature object using the *arc.data2sf* function. More information on the R-ArcGIS Bridge can be found here.

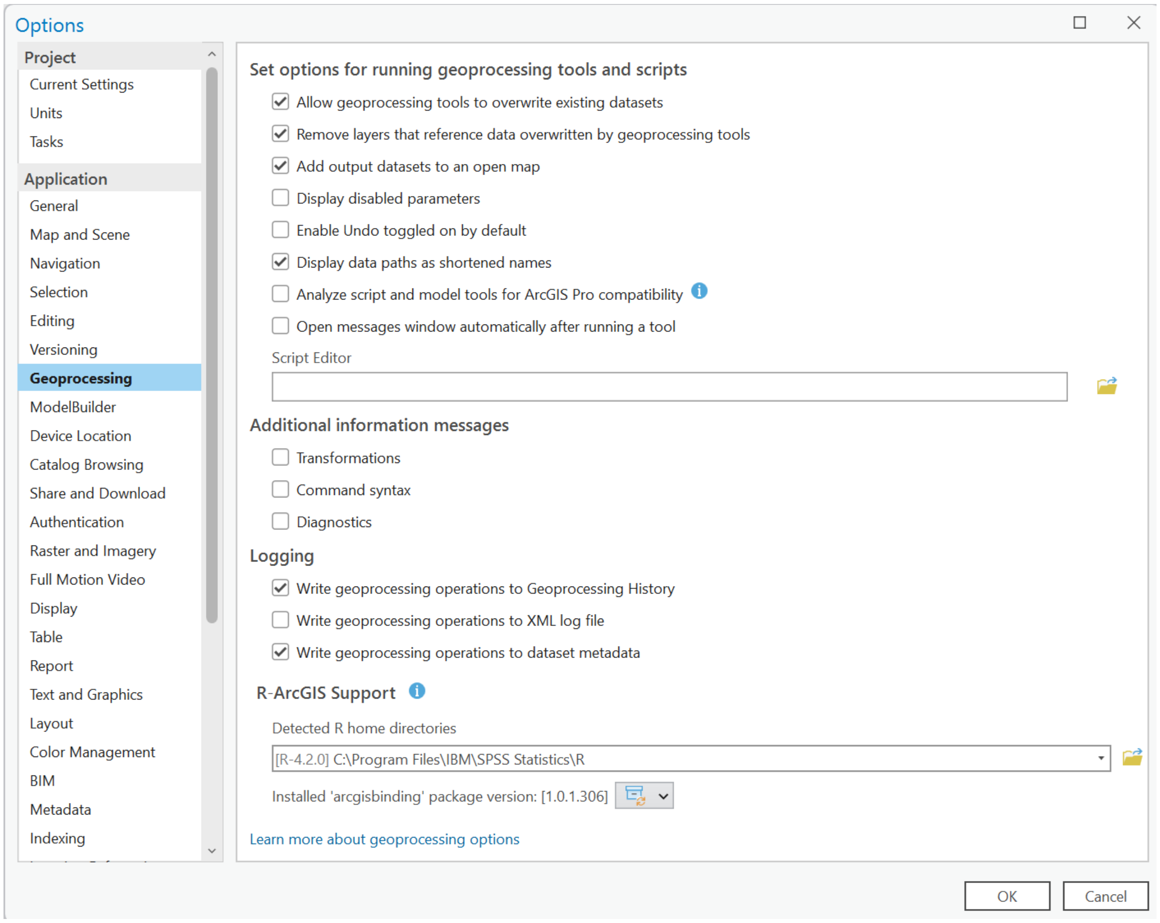


Figure 8 R-ArcGIS Support is configured in an ArcGIS Pro project in the Geoprocessing panel, under Options.

Parameters related to workspace, unique IDs, multicollinearity, GWR, and neighborhood level modeling are configured at the beginning of the code, shown in Table 4.

Table 4 List of configurable parameters.

Num	Category	Parameter Name	Description	Value
0	Home Geodatabase Path	home_gdb_path	**String**: Path to GDB workspace	“my_path/my_gdb.gdb”
1	Unique ID	unique_id	**String**: Name of column holding the unique identifier in the dataset	“unique_id”
2	Dependent Variable	dep_var	**Numeric Vector**: Dependent variable	bldg_sf$PSF_RENT_ADJ
		dep_var_trans	**Numeric Vector**: Transformed version of the dependent variable	bldg_sf$log_PSF
		dep_var_string	**String**: Name of dependent variable	"PSF_RENT_ADJ"

		dep_var_string_trans	**String**: Name of dependent variable transformed	“log_PSF”
		depend_var_list	**List**: List of names of dependent variable and transformed version	“c("PSF_RENT_ADJ", "log_PSF")”
3	Multicollinearity	r2_threshold	**Double**: Number between 0 and 1 defining the R-squared threshold to consider two pairs of independent variables as multicollinear	0.6
		patterns_to_check	**Dictionary**: User-defined variables not to be modeled together. For example, population density within 5- and 10-minutes walking.	list("CoTen" = c("CoTen"), "Park" = c("Park"), "Special" = c("Special"), "MEDHINC" = c("AVGHINC", "MEDHINC", “TotalSpending”), "AVGHINC" = c("MEDHINC", "AVGHINC", “TotalSpending”), "populationtotals" = c("POPDENS_CY"), "POPDENS_CY" = c("populationtotals", "POPDENS_CY"), "PSF_RENT_ADJ" = c("PSF_RENT_ADJ"), "business" = c("business"), "SpendingTotal" = c("SpendingTotal"))
4	Number of Model Combinations to Try	num_rows	**Integer**: Number of models to try with GWR	25
		num_rows_per_neighborhood	**Integer**: Number of models to try with neighborhood-level regression. This is doubled to test transformed and untransformed variable	30
		max_vars_per_model	**Integer**: Maximum number of variables included in a neighborhood-level model	6
5	Neighborhood Weights Matrix	k_nearest	**Integer**: Number of nearest neighbors considered for weights matrix	4
6	Coordinate System	crs	**Integer**: Global Coordinate System code	4326
		crs_utm	**Integer**: UTM Projected Coordinate System code	32618
7	GWR Parameters	list_args	**List**: List of parameters to input bw is bandwidth, kernel is the distance weights matrix, and adaptive is the type of bandwidth	c("bw", "kernel", "adaptive")
8	Number of R-Squared Neighborhoods Before Using K Means	num_classes	**Integer**: Number of R-squared neighborhoods before using K-means. Less than 10 is suggested because these will get further divided up with K Means	6
9	Number of K Means Clusters based on	num_clust	**Integer**: Number of K-means clusters based on R-squared	12
		max_k	**Integer**: The maximum number of discontinuous R-squared clusters will break up into This ensures that discontinuous neighborhoods with	5

			similar R-squared values do not belong to the same group	
10	Classification Types	classifications_list	**List**: Character list of classification types	c("jenks", "quantile", "equal")
		col_list	**List**: Character list of the names of columns that hold the names of each neighborhood	c("Neigh_Jenks", "Neigh_Quantile", "Neigh_Equal")

The code prints out initial diagnostics such as maps of local indications of spatial autocorrelation (LISA) for all independent and dependent variables, with the dependent variables shown in Figure 9. High-high, high-low, insignificant, low-high, and low-value clusters are indicated in the legend. LISA for the log transformation of rent shows high-high clusters in Times Square and on 5th Avenue on the Upper East Side with low-low clusters in midtown and low-high clusters in SoHo. Globally, the Moran's I for rent and the log transformed rent is 0.557 and 0.479, respectively, with both having a statistically significant p-value of 2.16e-16. A printout of the log transformation is shown in Figure 10 below for illustrative purposes.

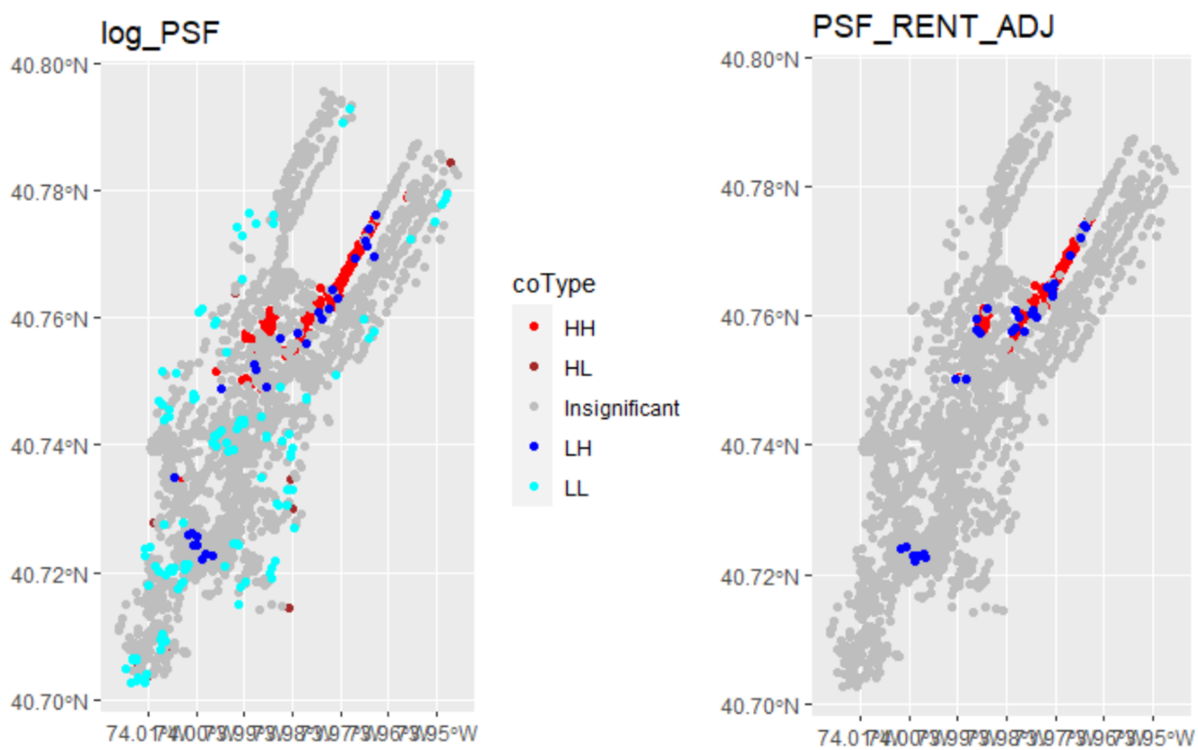


Figure 9 LISA maps for the log transformation of rent (left) and non-transformed rent (right).

```
        Moran I test under randomisation

data:  dep_var_data
weights: nb

Moran I statistic standard deviate = 31.11, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.4421724978     -0.0004651163      0.0002024392
```

Figure 10 Moran's I for the log of rent prices across Manhattan.

5.2.2. Batch Process GWR Models

GWR is streamlined for dozens of multivariate model combinations using 31 different predictor variables while also addressing multicollinearity and significance. Further, similar predictors measured over different trade areas (such as population density within five-minutes and 10-minutes walking or median income within one-tenth of a mile and average income within one-tenth of a mile) are not modeled together. Bisquare distance kernel is used ensuring that observations near the bandwidth's edge have little to no influence on the calculated value calculated for each target observation. In alignment with Oshan's (2019) methodology, this study also adopts scaling predictor variables. Variables are scaled using a min-max normalization value before running GWR, with the basic formula wrapped in a function, referenced in Figure 11 below. This type of scaling is simple and results in no negative numbers, which makes further analysis easier to work with.

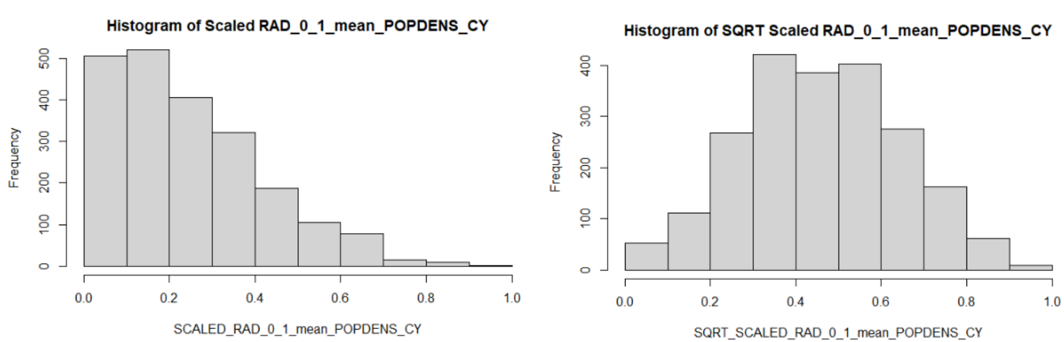
```r
function(column_of_df){

  # Min-Max Scaling
  X <- column_of_df
  X_min <- min(X, na.rm = TRUE)
  X_max <- max(X, na.rm = TRUE)
  X_scaled <- (X - X_min) / (X_max - X_min)

  return(X_scaled)
}
```

Figure 11 Min-max normalization.

Additionally, some variables are transformed with a square root to make its histogram shape closer to a normal distribution. Transforming variables is the only non-automated step that requires analyst review before proceeding. As examples, Figure 12 below shows the histogram for population density and average household income for both scaled (left column) and square root transformed (right column) versions. Normal Q-Q plots are also printed out in the console but omitted here for brevity.



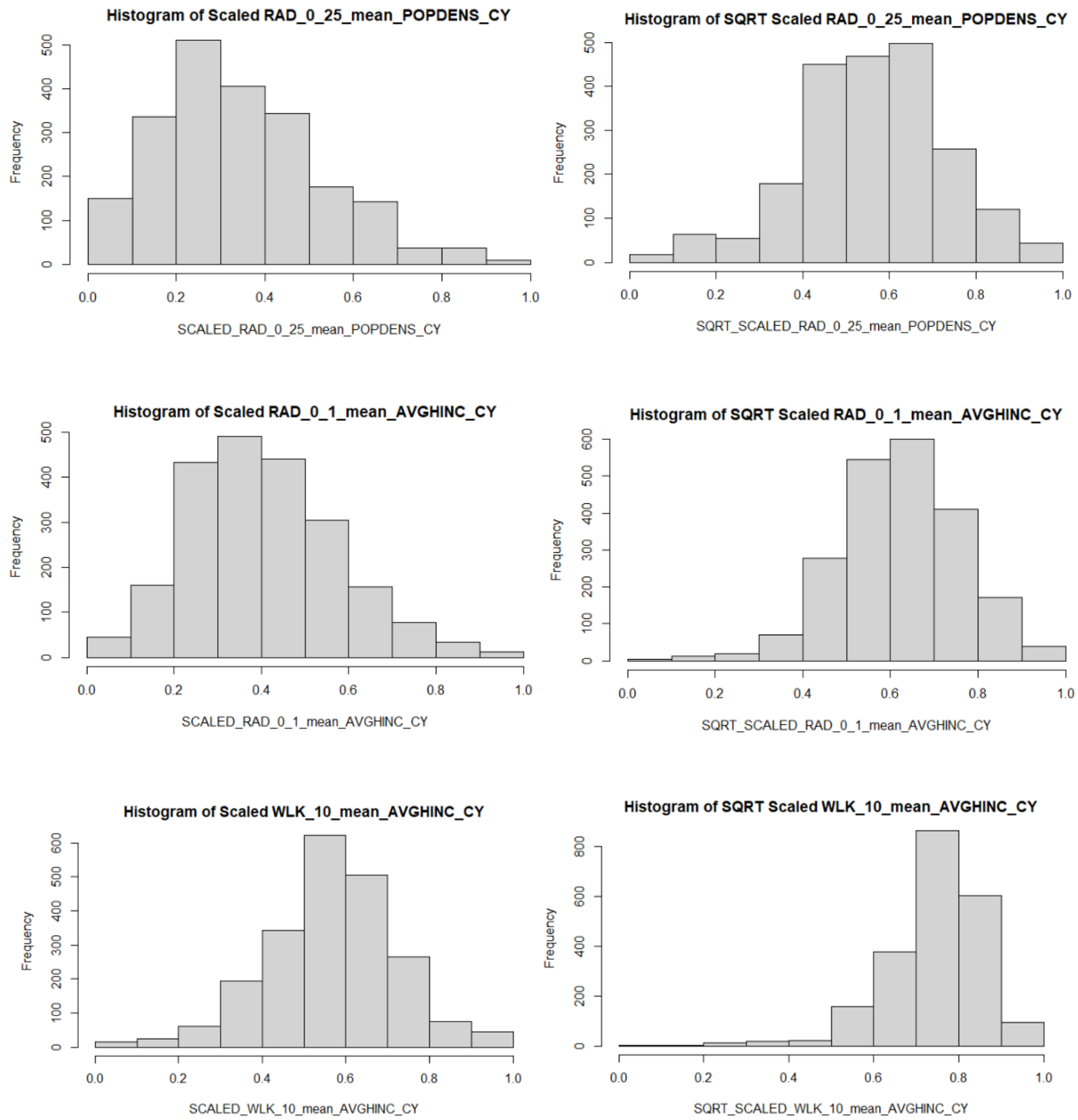


Figure 12 The left column is the scaled variable histogram, and the right column shows the histogram after a square root transformation.

The above figure shows that population density histogram fits a normal curve better after a square root transformation, whereas average household income fits a normal curve better without the transformation. To properly assign a transformation, a list called *transform_list* needs to mirror a list called *vars_strip*. Figure 13 below displays the code and specifically line 410 where transformations are assigned. The first item in *transform_list* corresponds to the transformation that will be applied to the first item in *vars_strip*. A dictionary called *transform_dict* is created with *transform_list* as keys and the mirrored *vars_strip* as values.

Binary variables are excluded from any scaling or transforming. Once this is configured, all succeeding code chunks may be executed.

```
> print(vars_strip)
[[1]]
[1] "POPDENS_CY"

[[2]]
[1] "DPOPDENSCY"

[[3]]
[1] "AVGHINC_CY"

[[4]]
[1] "SpendingTotal_X15001_X_A"

[[5]]
[1] "CoTen_V1"

[[6]]
[1] "PSF_RENT_ADJ"

[[7]]
[1] "CoTen_V2"

[[8]]
[1] "Subway"
```

```
### CREATE TRANSFORM LIST BASED ON HISTOGRAM AND NORMAL Q-Q PLOT RESULTS
### (POP DENS = SQRT, DPOP DENS = SQRT, AVG INC = SCALED,
### RETAIL SPEND = SCALED, CO-TEN V1 = SCALED, MEAN RENT PSF = SQRT, CO-TEN V2 = SCALED, SUBWAY = SQRT)

transform_list <- as.list(c("SQRT","SQRT","Scaled","Scaled","Scaled","SQRT","SCALED","SQRT"))
transform_dict <- setNames(transform_list,vars_strip)
```

Figure 13 “SQRT” or “Scaled” in each item of transform_list (bottom) mirrors the best transformation for each item in vars_strip (top).

Objects and functions related to the global GWR model are saved and referenced in RStudio’s environment. This includes the *valid_combo_dict* which is a dictionary of independent variables as the keys and a list of all other variables that are not collinear as their values. Others include the *gwr_tracker_df* which stores all potential GWR model combinations, the *bldgNB* which stores the spatial weights matrix and *bldgs_as_spatial* which is a *SpatialPointsDataFrame* converted from the original simple feature data frame. The GWModel does not execute on simple features, so converting to a *SpatialPointsDataFrame* is necessary. Considering Manhattan’s wide range of rent values, the dependent variable is logged using a base of 10 and referenced as *log_PSF*.

The *valid_combo_dict* is created by looping through each independent variable and storing it as the key value. Next, a simple bivariate OLS regression is run on that key against all other independent variables. If the OLS model results in an R-squared value of less than 0.6, then that variable is added the list of non-collinear values. As alluded to before, variables with the same root are not modelled even if they are not collinear.

The *gwr_tracker_df* references the *valid_combo_dict* to build out the potential models for GWR. First, *gwr_tracker_df* is initialized with two random non-collinear variables. A third variable is then added based on the common non-collinear variables of the first pair. The fourth is added based on the common of the first three, and so on until there are no more variables to add. While the script allows for the inclusion of up to 20 variables, the maximum utilized for this study is 10.

A function called *gwr_function_flex* iterates through each model in *gwr_tracker_df*, removes any insignificant variables, and ensures all resulting models are unique. Each GWR model has its own unique ID. The script then filters out for models that have large bandwidths over 20% of the data. Because GWR is run using the transformed version of rent (*log_PSF*), the results are stored in the *gwr_res_tracker_df_trans*, seen below in Figure 14. If GWR is run on the original, non-transformed dependent variable, those results would be stored in *gwr_res_tracker_df* (without the *trans* suffix). This allows the analyst to decide on which version of the dependent variable is most appropriate for the study.

	id	bw	kernel	RSS.gw	AIC	AICc	enp	edf	gw.R2	gwR2.adj	BIC
1	7	389	bisquare	119.3013	-650.6110	-569.19798	98.34671	2574.653	0.5283601	0.5103374	-2807.3769
2	16	139	bisquare	109.6552	-595.0318	-126.85118	454.63852	2218.361	0.5664946	0.4776105	-815.8233
3	1	166	bisquare	116.7008	-490.8085	-121.90839	376.39494	2296.605	0.5386407	0.4629947	-1140.4411
4	8	147	bisquare	111.6542	-559.3358	-111.94475	439.28707	2233.713	0.5585917	0.4717445	-866.9169
5	6	156	bisquare	115.7578	-492.0284	-91.41700	403.22371	2269.776	0.5423688	0.4610352	-1000.6184
6	2	222	bisquare	125.2211	-364.4959	-86.37596	298.53422	2374.466	0.5049571	0.4426906	-1441.6889
7	10	166	bisquare	117.3519	-466.3926	-82.81817	388.95807	2284.042	0.5360666	0.4570271	-1050.2588
8	3	393	bisquare	137.9660	-215.1085	-79.28183	158.76737	2514.233	0.4545719	0.4201158	-2048.2188
9	14	111	bisquare	112.9223	-527.8384	-78.29948	444.99870	2228.001	0.5535787	0.4643749	-826.3826
10	18	155	bisquare	115.6813	-484.8893	-70.19656	414.18871	2258.811	0.5426711	0.4587755	-932.1131
11	9	156	bisquare	127.2944	-318.9022	-38.40082	305.55021	2367.450	0.4967605	0.4317834	-1384.3915
12	4	155	bisquare	125.1079	-341.4554	-27.03856	333.66155	2339.338	0.5054046	0.4348299	-1243.2192

	var_1	var_2	var_3	var_4
1	Corner	SCALED_WLK_5_CoTen_V2	SQRT_SCALED_RAD_0_1_Mean_PSF_RENT_ADJ	SQRT_SCALED_WLK_10_mean_DPOPDENSCY
2	Corner	Food	SCALED_RAD_0_25_CoTen_V2	SCALED_WLK_5_mean_AVGHINC_CY
3	Corner	Food	SCALED_WLK_10_CoTen_V2	SCALED_WLK_5_mean_AVGHINC_CY
4	Corner	Food	SCALED_RAD_0_25_CoTen_V1	SCALED_WLK_5_mean_AVGHINC_CY
5	Corner	Food	SCALED_RAD_0_25_CoTen_V1	SCALED_RAD_0_25_mean_SpendingTotal_X15001_X_A
6	Corner	Food	SCALED_RAD_0_25_CoTen_V1	SCALED_WLK_5_mean_AVGHINC_CY
7	Corner	Food	SCALED_RAD_0_25_CoTen_V2	SCALED_WLK_5_mean_SpendingTotal_X15001_X_A
8	Corner	Food	SCALED_WLK_5_CoTen_V1	SQRT_SCALED_Dist_Subway
9	Corner	Food	SQRT_SCALED_Dist_Subway	SQRT_SCALED_WLK_10_mean_POPDENS_CY
10	Corner	Food	SCALED_RAD_0_25_CoTen_V1	SCALED_RAD_0_25_mean_AVGHINC_CY
11	Corner	Food	SQRT_SCALED_Dist_Subway	SQRT_SCALED_RAD_0_1_mean_DPOPDENSCY
12	Corner	Food	SCALED_RAD_0_25_CoTen_V2	SQRT_SCALED_Dist_Subway

	var_5	var_6	var_7	var_8
1	NA	NA	NA	NA
2	SQRT_SCALED_Dist_Subway	SQRT_SCALED_RAD_0_1_mean_DPOPDENSCY	SQRT_SCALED_RAD_0_25_mean_POPDENS_CY	SQRT_SCALED_WLK_10_Mean_PSF_RENT_ADJ
3	SQRT_SCALED_Dist_Subway	SQRT_SCALED_WLK_10_mean_DPOPDENSCY	SQRT_SCALED_WLK_10_mean_POPDENS_CY	SQRT_SCALED_WLK_10_Mean_PSF_RENT_ADJ
4	SQRT_SCALED_Dist_Subway	SQRT_SCALED_RAD_0_1_mean_DPOPDENSCY	SQRT_SCALED_RAD_0_25_mean_POPDENS_CY	SQRT_SCALED_WLK_5_Mean_PSF_RENT_ADJ
5	SCALED_WLK_10_mean_AVGHINC_CY	SQRT_SCALED_Dist_Subway	SQRT_SCALED_RAD_0_25_Mean_PSF_RENT_ADJ	SQRT_SCALED_WLK_5_mean_POPDENS_CY
6	SQRT_SCALED_Dist_Subway	SQRT_SCALED_RAD_0_1_mean_DPOPDENSCY	SQRT_SCALED_RAD_0_25_Mean_PSF_RENT_ADJ	SQRT_SCALED_WLK_10_mean_POPDENS_CY
7	SQRT_SCALED_Dist_Subway	SQRT_SCALED_RAD_0_1_mean_POPDENS_CY	SQRT_SCALED_RAD_0_25_mean_DPOPDENSCY	SQRT_SCALED_RAD_0_25_Mean_PSF_RENT_ADJ
8	SQRT_SCALED_RAD_0_1_mean_POPDENS_CY	SQRT_SCALED_RAD_0_25_Mean_PSF_RENT_ADJ	SQRT_SCALED_WLK_10_mean_DPOPDENSCY	NA
9	SQRT_SCALED_WLK_10_Mean_PSF_RENT_ADJ	SQRT_SCALED_WLK_5_mean_DPOPDENSCY	NA	NA
10	SQRT_SCALED_Dist_Subway	SQRT_SCALED_RAD_0_1_mean_DPOPDENSCY	SQRT_SCALED_WLK_10_Mean_PSF_RENT_ADJ	SQRT_SCALED_WLK_5_mean_POPDENS_CY
11	SQRT_SCALED_RAD_0_25_Mean_PSF_RENT_ADJ	NA	NA	NA
12	SQRT_SCALED_RAD_0_25_mean_DPOPDENSCY	SQRT_SCALED_RAD_0_25_Mean_PSF_RENT_ADJ	NA	NA

Figure 14 Resulting *gwr_res_tracker_df_trans* from running GWR. Moments are at the top, followed by variables one through four, five through six, and seven though eight (bottom).

Each model's results are stored in a CSV file in the working folder, seen in the lines of code below in Figure 15. The *spatial data frame* (SDF) of the *gwr*.res contains the GWR results.

```
            results <- as.data.frame(gwr.res$SDF)

            #write to csv
            gwr_res_print <<- print(gwr.res)

            csv_name <- gwr_trcker_df$id[counter]

            write.csv(results,paste('GWR_Res_',paste0(csv_name,trns),'.csv'), row.names = FALSE)
```

Figure 15 Saving Spatial Data Frame GWR Results to a CSV.

The results contain GWR moments for each point. This includes coefficients, intercepts, error estimates, local R-squared values, and others. The model that ultimately gets chosen has a bandwidth that includes less than or equal to 20% of the data, an R-squared of at least 0.4, and has the lowest AICc. The AICc adds penalties for more parameters in the model, especially if the sample size is low. Even though the sample size here is robust, the AICc is consistently and sufficiently higher than the AIC score of the same model, therefore the AICc is used to determine the best fit. These results get joined to the original simple features data frame into an object called *bldg_best_model_join*. Figure 16 shows randomness of residuals, indicating spatial non-stationarity is removed. Finally, Figure 17 shows the entire logic flow, at a high level.

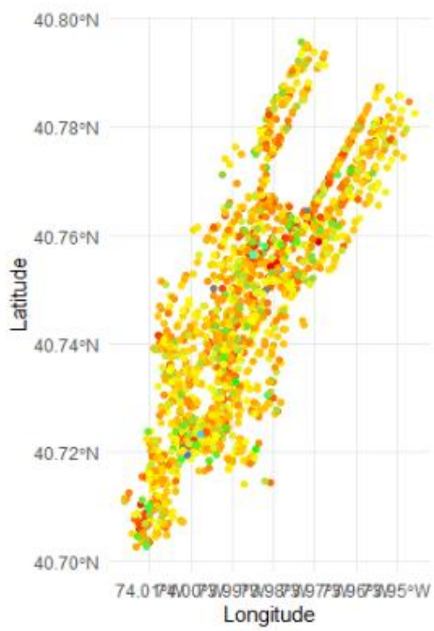


Figure 16 Randomness in residuals indicates spatial autocorrelation is resolved.

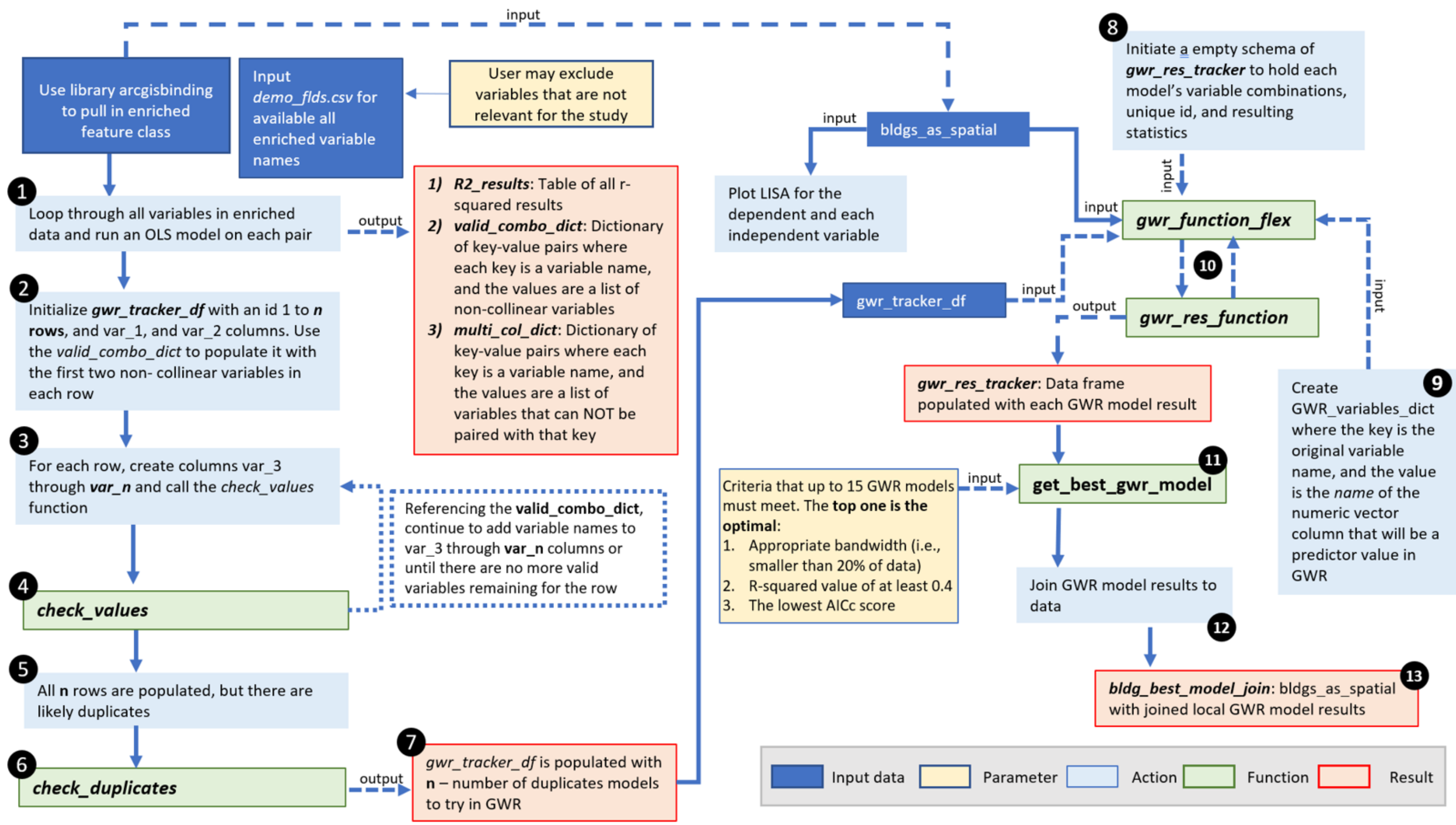


Figure 17 Flowchart of logic to batch process multiple GWR models.

5.2.3. Neighborhood Creation by Combining K-Means with Natural Jenks, Quantile, and Equal Interval Local R-Squared Delineations

Neighborhoods are created using natural Jenks, quantile, and equal interval classifications of the local R-squared values. However, due to the large data size and intense computational power that would be needed, delineation using the fissures method is used instead of natural Jenks. After classifying, the code redistributes points as needed should a neighborhood have less than 150 points. This ensures that there are at least 30 points for 20% data validation, while still maintaining as close to the original delineations as possible. Figure 18 below shows the count of points for each classification type's neighborhoods. The *bldg_best_model_join* is sorted descending based on local R-squared values. For clarity, rows one through 250, which have a minimum R-squared value of 0.680 and a maximum of 0.764 are assigned to neighborhood *Equal 1*.

	Class	Count	Start_Row	Min_R2	Max_R2
1	Equal 1	250	1	0.6802502	0.7635748
2	Equal 2	345	251	0.5967963	0.6798337
3	Equal 3	513	596	0.5129654	0.5955813
4	Equal 4	414	1109	0.4293392	0.5127508
5	Equal 5	794	1523	0.3458218	0.4291871
6	Equal 6	353	2317	0.2622110	0.3457506

	Class	Count	Start_Row	Min_R2	Max_R2
1	Jenk 1	228	1	0.6865033	0.7635748
2	Jenk 2	386	229	0.5893524	0.6854831
3	Jenk 3	652	615	0.4798240	0.5885755
4	Jenk 4	588	1267	0.3964892	0.4786824
5	Jenk 5	555	1855	0.3327167	0.3963641
6	Jenk 6	260	2410	0.2622110	0.3323625

	Class	Count	Start_Row	Min_R2	Max_R2
1	Quantile 1	445	1	0.6509698	0.7635748
2	Quantile 2	445	446	0.5341917	0.6509595
3	Quantile 3	444	891	0.4609386	0.5341337
4	Quantile 4	445	1335	0.4026900	0.4609220
5	Quantile 5	445	1780	0.3538095	0.4026567
6	Quantile 6	445	2225	0.2622110	0.3537999

Figure 18 Equal interval (top), natural Jenks (middle) and quantile (bottom) classifications showing the number of points that fall within each minimum and maximum R-squared value.

A low local R-squared value, such as points within 4 through 6 neighborhoods, indicates that the parameters used in the GWR model are not a good fit to estimate rent in those areas. Further this may indicate that it is not a good fit to group these areas together. Currently, areas with similar local R-squared values are grouped together as one neighborhood, even if they are disjoint. This would mean that disconnected regions like Chelsea and the Upper West Side are classified as a single neighborhood due to their similar R-squared values. Therefore, the script uses K-means machine learning (ML) clustering algorithm based on the existing neighborhood classifications to separate disjoint neighborhoods. This is referenced as *a dual K-means* approach. Figure 19 shows a flow chart of this code logic.

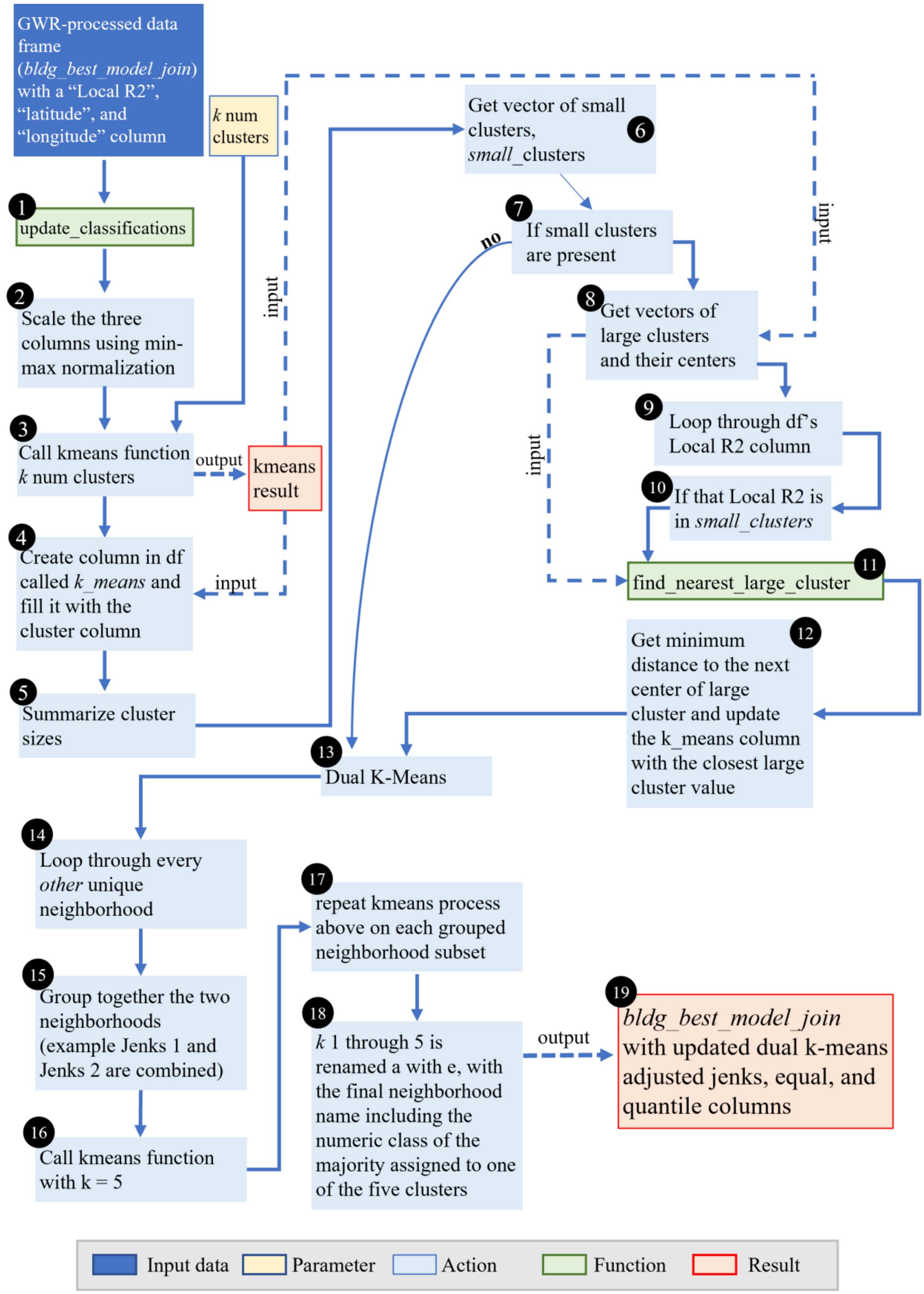


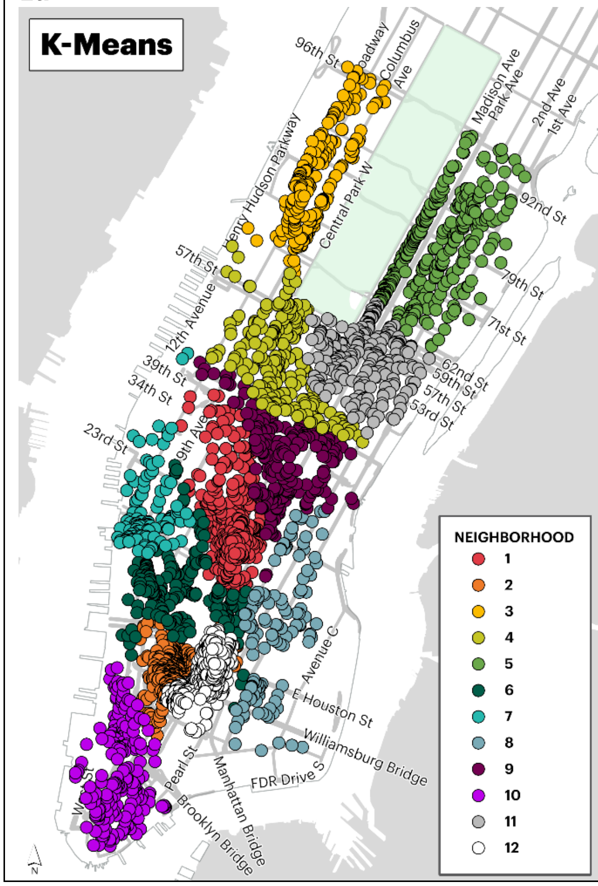
Figure 19 Dual k-means classification flow chart.

Table 5 displays the neighborhood names resulting from the execution of the dual K-means approach. Depending on the distribution of the points, it is possible for numeric neighborhoods to be absorbed into neighboring ones. The dual K-means results in 15 neighborhoods for equal interval, natural Jenks, and quantile. Figure 20 below shows results from classifying neighborhoods using K-means, the original classifications, and by combining the two.

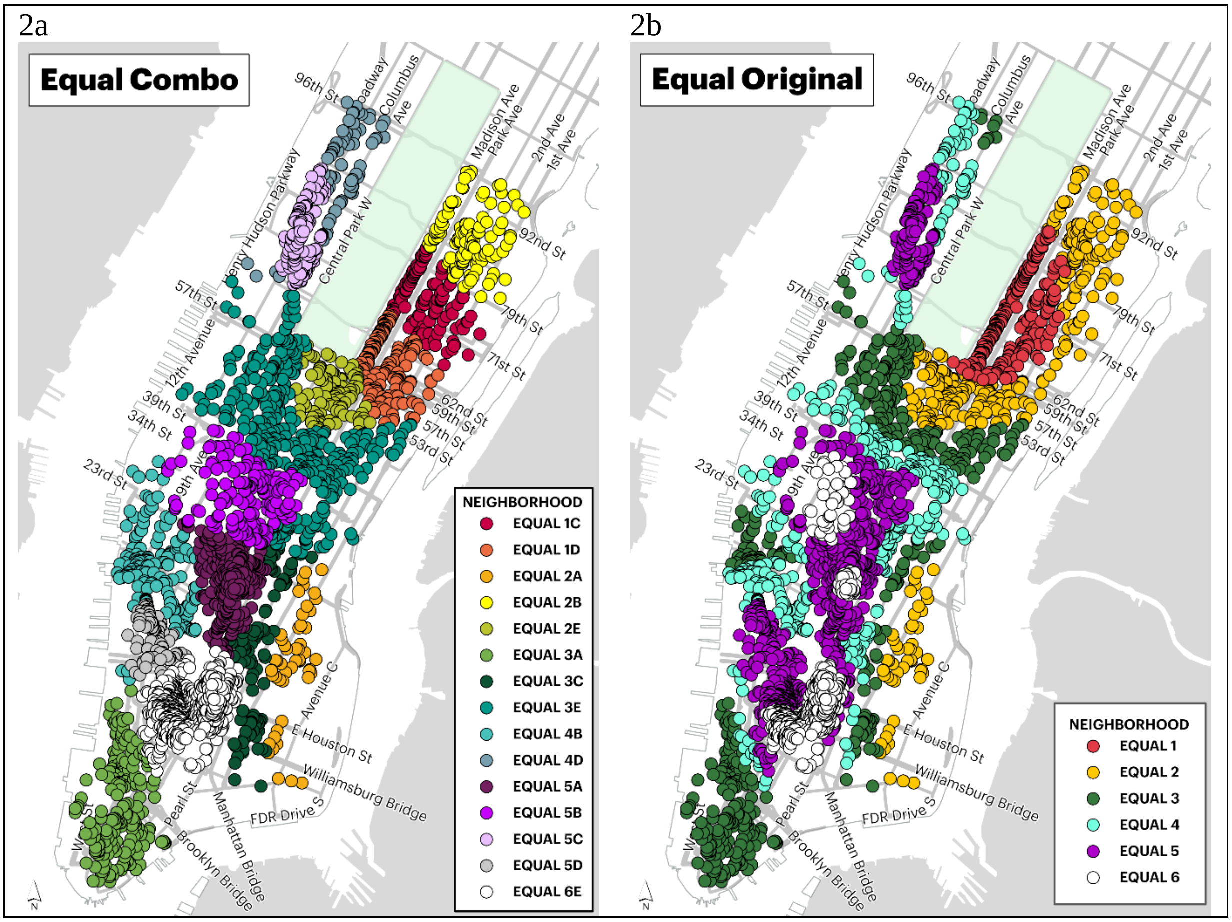
Table 5 Dual K-means neighborhoods for equal interval, natural Jenks, and quantile classifications.

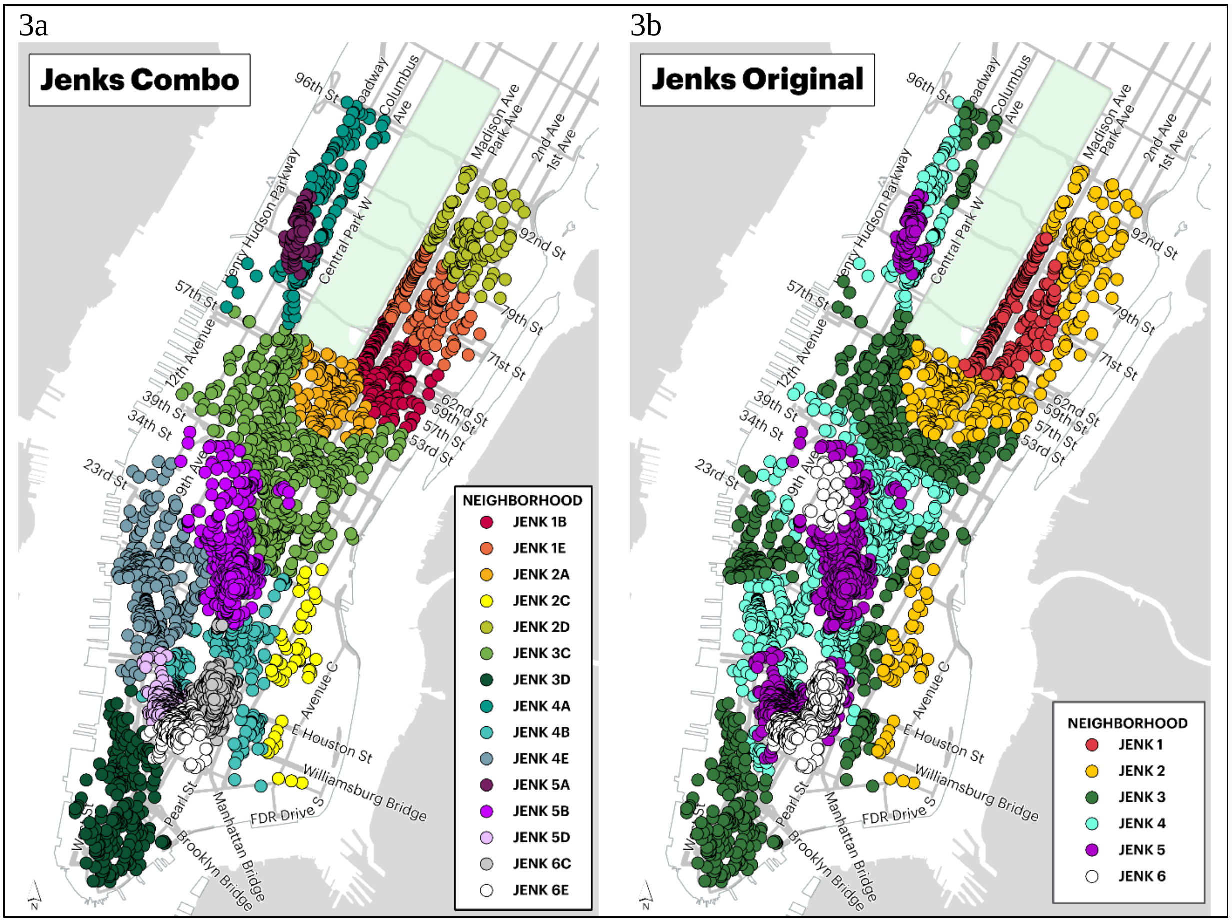
Class	Original Neighborhood	Dual K-Means Neighborhood
Equal Interval	1	Equal 1c, 1d
	2	Equal 2a, 2b, 2e
	3	Equal 3a, 3c, 3e
	4	Equal 4b, 4d
	5	Equal 5a, 5b, 5c, 5d
	6	Equal 6e
Natural Jenks	1	Jenks 1b, 1e
	2	Jenks 2a, 2c, 2d
	3	Jenks 3c, 3d
	4	Jenks 4a, 4b, 4e
	5	Jenks 5a, 5b, 5d
	6	Jenks 6c, 6e
Quantile	1	Quantile 1c, 1d, 1e
	2	Quantile 2a, 2b
	3	Quantile 3c
	4	Quantile 4a, 4b, 4d, 4e
	5	Quantile 5a, 5c, 5d, 5e
	6	Quantile 6b

1a



K-Means
NEIGHBORHOOD
1
2
3
4
5
6
7
8
9
10
11
12
96th St
Broadway
Columbus Ave
Madison Ave
Park Ave
2nd Ave
1st Ave
Henry Hudson Parkway
Central Park W
92nd St
79th St
71st St
62nd St
59th St
57th St
53rd St
57th St
12th Avenue
39th St
34th St
23rd St
9th Ave
Avenue C
E Houston St
Williamsburg Bridge
FDR Drive S
Pearl St
Manhattan Bridge
Brooklyn Bridge


2a
Equal Combo
NEIGHBORHOOD
EQUAL 1C
EQUAL 1D
EQUAL 2A
EQUAL 2B
EQUAL 2E
EQUAL 3A
EQUAL 3C
EQUAL 3E
EQUAL 4B
EQUAL 4D
EQUAL 5A
EQUAL 5B
EQUAL 5C
EQUAL 5D
EQUAL 6E
2b
Equal Original
NEIGHBORHOOD
EQUAL 1
EQUAL 2
EQUAL 3
EQUAL 4
EQUAL 5
EQUAL 6


3a
Jenks Combo
NEIGHBORHOOD
JENK 1B
JENK 1E
JENK 2A
JENK 2C
JENK 2D
JENK 3C
JENK 3D
JENK 4A
JENK 4B
JENK 4E
JENK 5A
JENK 5B
JENK 5D
JENK 6C
JENK 6E
3b
Jenks Original
NEIGHBORHOOD
JENK 1
JENK 2
JENK 3
JENK 4
JENK 5
JENK 6

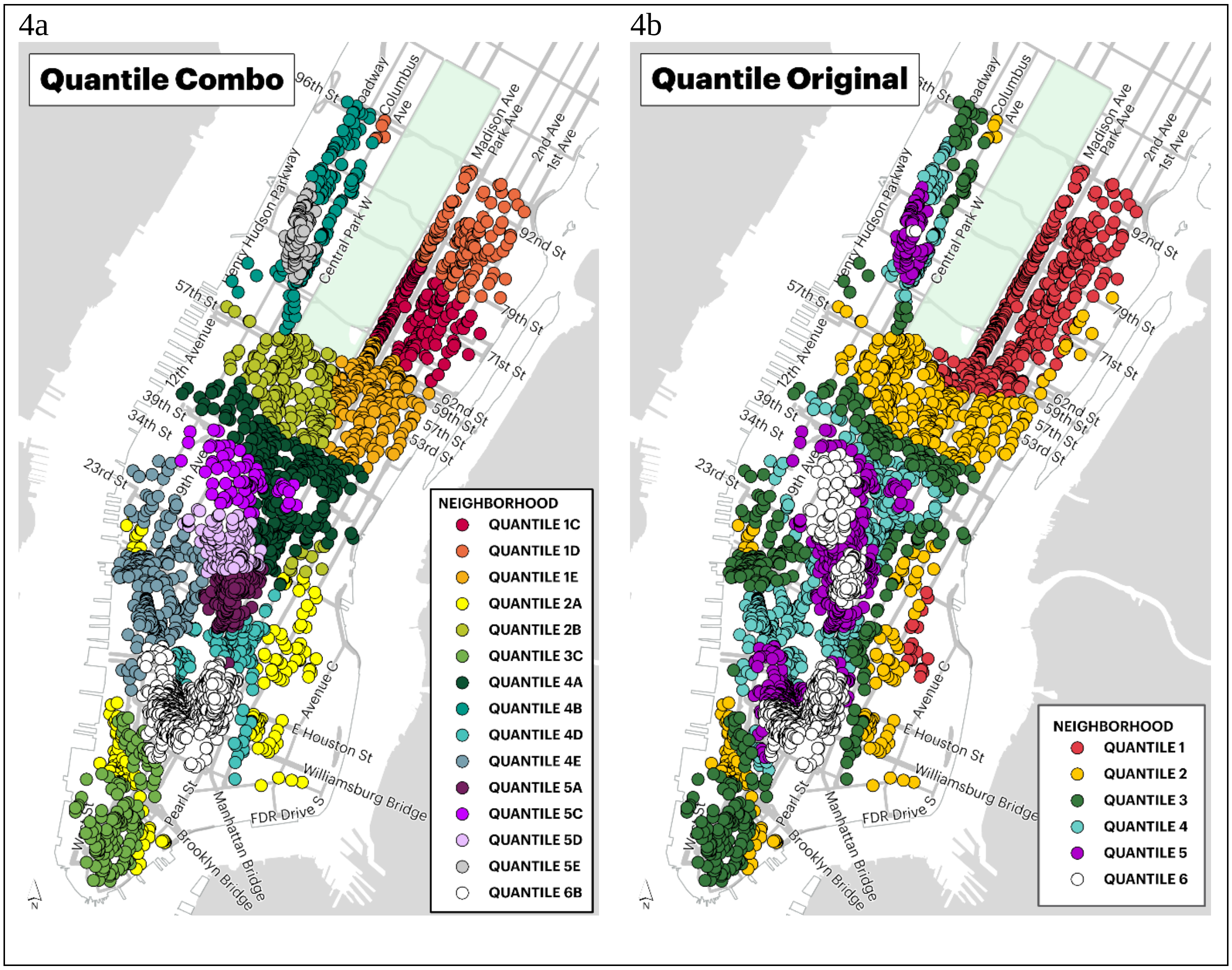


Figure 20 K-means classification (1a), equal interval with dual K-means (2a), equal interval original (2b), natural Jenks with dual K-means (3a), natural Jenks original (3b), quantile with dual K-means (4a) and quantile original (4b).

5.2.4. Assess Neighborhood Spatial Autocorrelation

Applying the Moran's I test to the OLS residuals for individual neighborhoods assesses the presence of spatial autocorrelation. Ideally, these neighborhoods exhibit lower spatial autocorrelation compared to the global dataset. Nonetheless, the extent of reduction in spatial autocorrelation may vary across neighborhoods.

5.2.5. Pre-Process Variables for Neighborhood

The 45 neighborhoods (15 equal, 15 Jenks, and 15 quantile delineations) require 45 unique regression models, one for each. The next phase of the code finds the best transformation of each base variable through modelling a linear regression against *log_PSF* within each neighborhood. Each neighborhood data subset retains the original scaled variable and its best transformed version. The code evaluates the optimal transformation for each variable by analyzing the normality of residuals through a Quantile-Quantile (QQ) plot, alongside the R-squared value. It calculates the correlation coefficient of points in the QQ plot, where a higher absolute value indicates a more normal distribution of residuals. A transformation that yields a higher R-squared value and a QQ plot correlation coefficient of at least 0.5 is considered better than previous transformations. If no transformation results in normal residuals and a better R-squared value, then only the scaled, non-transformed variable is considered for that neighborhood. Transformations tested include a square root transformation, a log transformation, and box-cox transformations ranging from 0.1 to 3, at one-tenth increments, on each neighborhood data subset, for each classification type. In total there are 32 transformations to test on 32 variables (eight non-binary base variables at four geographies) for 45 different neighborhood data subsets. This is a total of 46,080 tests and is executed in about three minutes. Scatterplots comparing the best transformation and the scaled variable against rent and the log of rent are then saved to a folder in the workspace.

5.2.6. Batch Process OLS and Spatial Regression Neighborhood Models

Just as the *gwr_tracker_df* is used to iterate GWR models, a multivariate regression tracker is similarly created for each neighborhood. This tracker, designated with the prefix mlr and named *mlr_tracker_df_classification_NeighID*, is established to store 60 different models for iterative processing. This parameter is set at the beginning of the code as half of that number. The original parameter for number of models is duplicated because each *mlr_tracker_df_classification_NeighID* holds each model twice, one using the original non-transformed *ADJ_PSF_RENT* as the dependent variable and then another using *log_PSF*. Similar to the *gwr_tracker_df*, each model in the *mlr* tracker gets its own unique ID.

The batch processing executes a pseudo stepwise logic to calibrate the most optimal non-spatial multivariate regression model. First, each model is run as is against the non-transformed adjusted rent. If there are insignificant variables, those are removed, and the model is re-run. Next, there is a check on outliers using a Cook's D 4/n threshold, if there are any, those datapoints are removed, and the data is re-run. Finally, the script checks for heteroskedasticity in the residuals. If heteroskedasticity is present then the model is re-run again with the logged base 10 transformation of the adjusted rent. This process is then repeated for the logged rent variable. It tests the models using an 80/20 calibration/validation methodology. If heteroskedasticity or outliers remain, the final model iteration's results are documented anyways. Coefficients and other moments are added as columns into the neighborhood-specific *mlr_tracker*. The below Figure 21 illustrates this process in a flow chart.

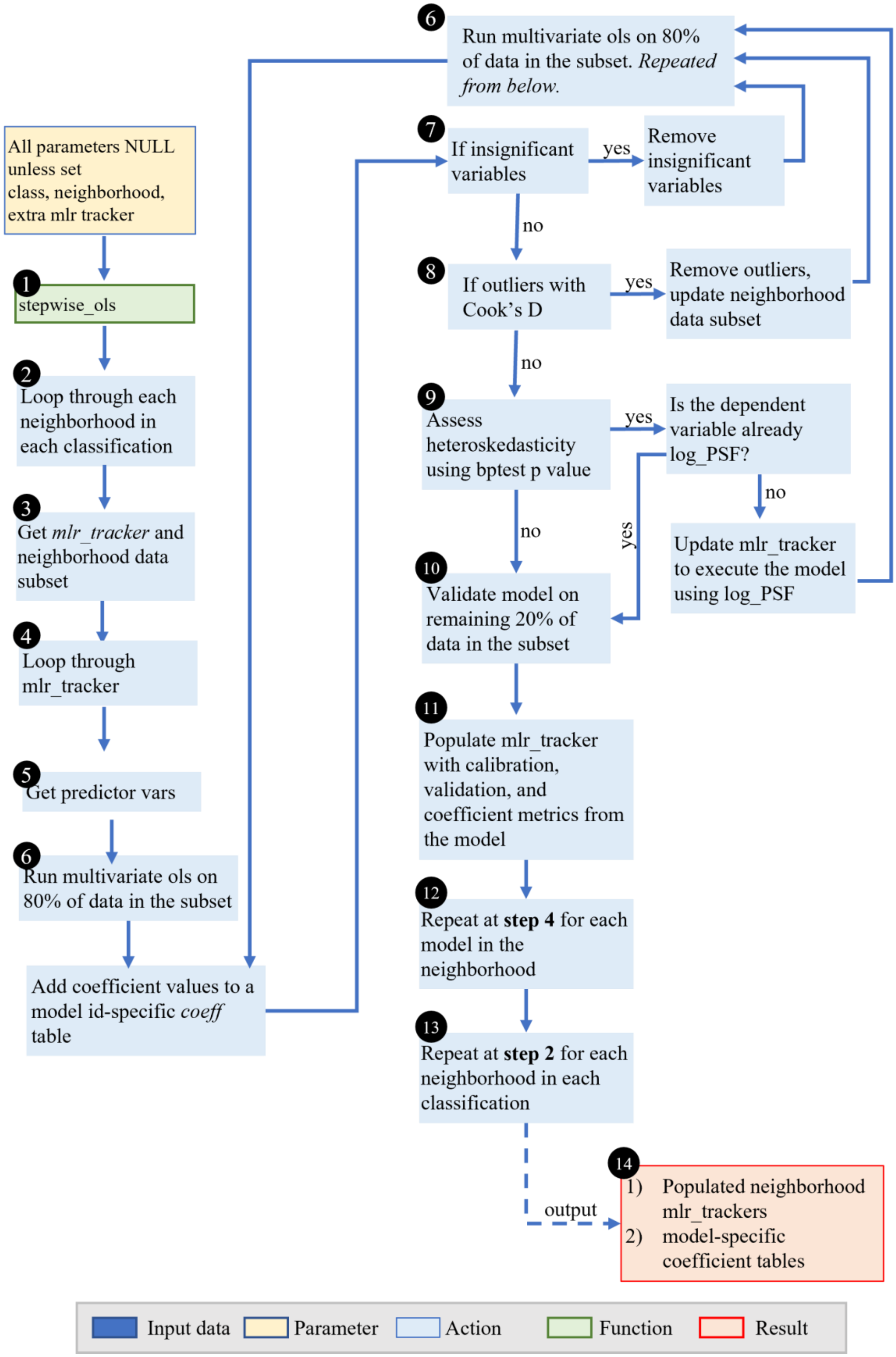


Figure 21 Stepwise OLS flow chart.

Once all iterations for all neighborhoods are run, the best models are aggregated to three classification-specific tables named *best_models_df_classification*. For illustrative purposes, in *best_models_df_jenks*, the best models for neighborhood *Jenk 1b* and *Jenk 2e* are model ID numbers 26, and 29 respectively. Only one model per neighborhood is saved to this aggregated table. If the R-squared value of an optimal model for a given neighborhood is below 0.45, a fresh search for an ideal model is initiated. This includes creating a new *mlr_tracker_df* with an *extra* suffix. For example, if neighborhood *Jenk 1c* has no spatial autocorrelation and an R-squared value of less than 0.5, then the *mlr_tracker_df_jenks_N1c_EXTRA* is generated, populated with another random set of 60 models, and tested. If the best model from the second round of testing has a higher R-squared value than the original one, that row in the *best_models_df_classification* is replaced with the better neighborhood model. This ensures that only neighborhoods needing improved models are subject to a second round of model calibration.

Neighborhoods exhibiting significant spatial autocorrelation with a Moran's I of 0.15 (positive or negative) and a p-value below 0.05, get replaced with a spatial error or spatial lag version of the optimal OLS model. The better of the two spatial regression models is determined by the lower AIC value. If the difference between the two AIC scores is less than three, then the spatial error model is automatically selected. A neighborhood below 0.15 spatial autocorrelation (positive or negative), regardless of significance, remains with the OLS model.

6. Results

6.1 Impacts of Different Co-Tenancy Perspectives

LISA maps reveal distinct narratives for the two co-tenancy perspectives. The first perspective, which focuses on the idea of one type of healthy retail corridor, displays fewer clusters. This is primarily in retail neighborhoods downtown such as SoHo, NoHo, and Union Square as well as the midtown retail corridor surrounding Herald Square. These retail corridors are home to a wide variety of big box and local tenants. In contrast, the second perspective shows smaller, more dispersed clusters, highlighting the relevance of evaluating co-tenancy in relation to a neighborhood's primary theme. In addition to the retail clusters highlighted in perspective one, this perspective shows various retail hotspots throughout Manhattan, across neighborhoods exhibiting all four major themes. This suggests that while a neighborhood should maintain a core identity – regional retail, local trendy retail, residential, or office- it also benefits from integrating key tenants that align with secondary overlapping themes.

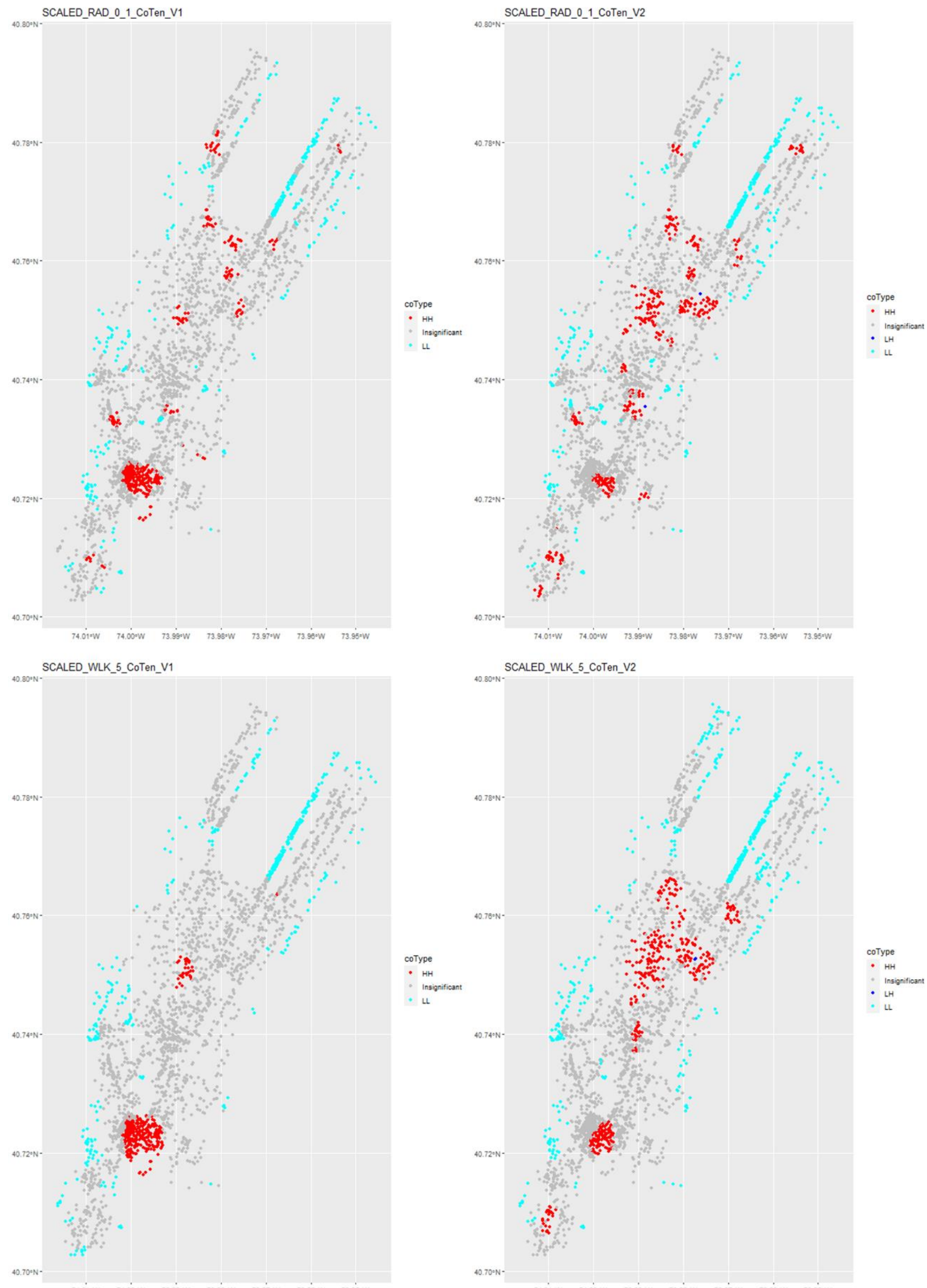


Figure 22 LISA maps for co-tenancy perspectives one (left) and two (right).

6.2 Best Model for GWR

The below Table 6 highlights the optimal GWR model after attempting 25 different models. Figure 23 follows with a histogram of the local R-squared values and residuals. GWR ID number 7 is selected for its lowest AICc value.

Table 6 High level information for the best or user selected GWR model.

Name	Value
ID	7
bandwidth	389
kernel	bisquare
RSS.gw	119.301
AIC	-650.611
AICc	-569.120
enp	98.347
edf	2,574.653
gw.R2	0.528
gwR2.adj	0.510
BIC	-2,807.377
var_1	Corner
var_2	Walk 5 – Co-Tenancy V2
var_3	Rad 0.1 – Mean PSF Rent Adj. (Sqrt)
var_4	Walk 10 – Daytime Population Density (Sqrt)

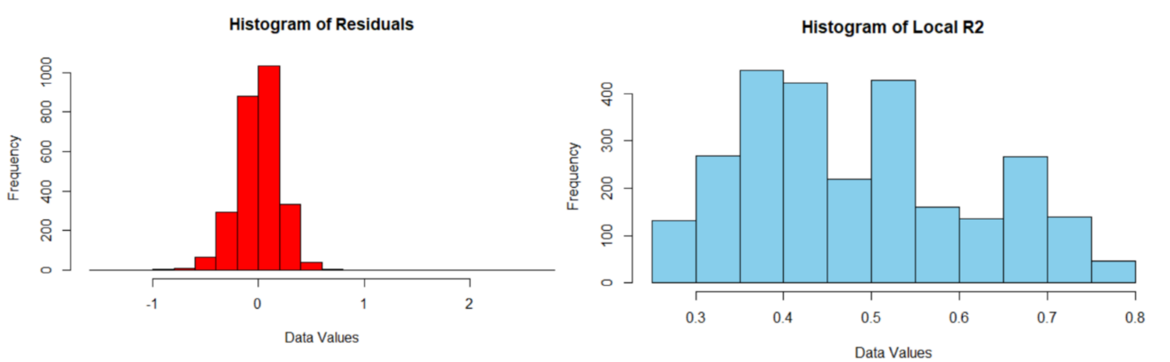


Figure 23 Histogram distribution of GWR residuals and local R-squared.

Table 7, 8, and 9 below show residuals, model diagnostic information, and variable coefficients, respectively, for global regression. Table 10, 11 and 12 show calibration information, diagnostic information, and summary of GWR coefficient estimates, respectively. Mapped GWR coefficients in Figure 24 highlight the spatial non-stationarity across Manhattan.

Table 7 Residuals information for global non-spatial regression.

Min	1Q	Median	3Q	Max
-1.5866	-0.1199	0.0144	0.1332	3.2334

Table 8 Diagnostic information for global non-spatial regression.

Name	Value
Residual Standard Error	0.2186 on 2668 degrees of freedom
Multiple R-squared	0.4958
Adjusted R-squared	0.495
F-statistic	655.8 on 4 and 2668 DF, p-value: < 2.2e-16
RSS	127.549
Sigma(hat)	0.219
AIC	-534.850
AICc	-534.818
BIC	-3,125.158

Table 9 Summary of coefficients for global non-spatial regression.

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	1.728	0.0163	106.137	<2.00E-16***
Walk 5 – Co Ten V2	0.165	0.0293	5.614	2.18E-08***
Rad 0.1 – Mean PSF Rent Adj (Sqrt)	1.648	0.0334	49.325	<2.00E-16***
GWR_Corner	0.0783	0.0097	8.05	1.24E-15***
Walk 10 - Daytime Population Density (Sqrt)	-0.0744	0.0235	-3.169	1.55E-03**
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*				

Table 10 Calibration information for GWR model.

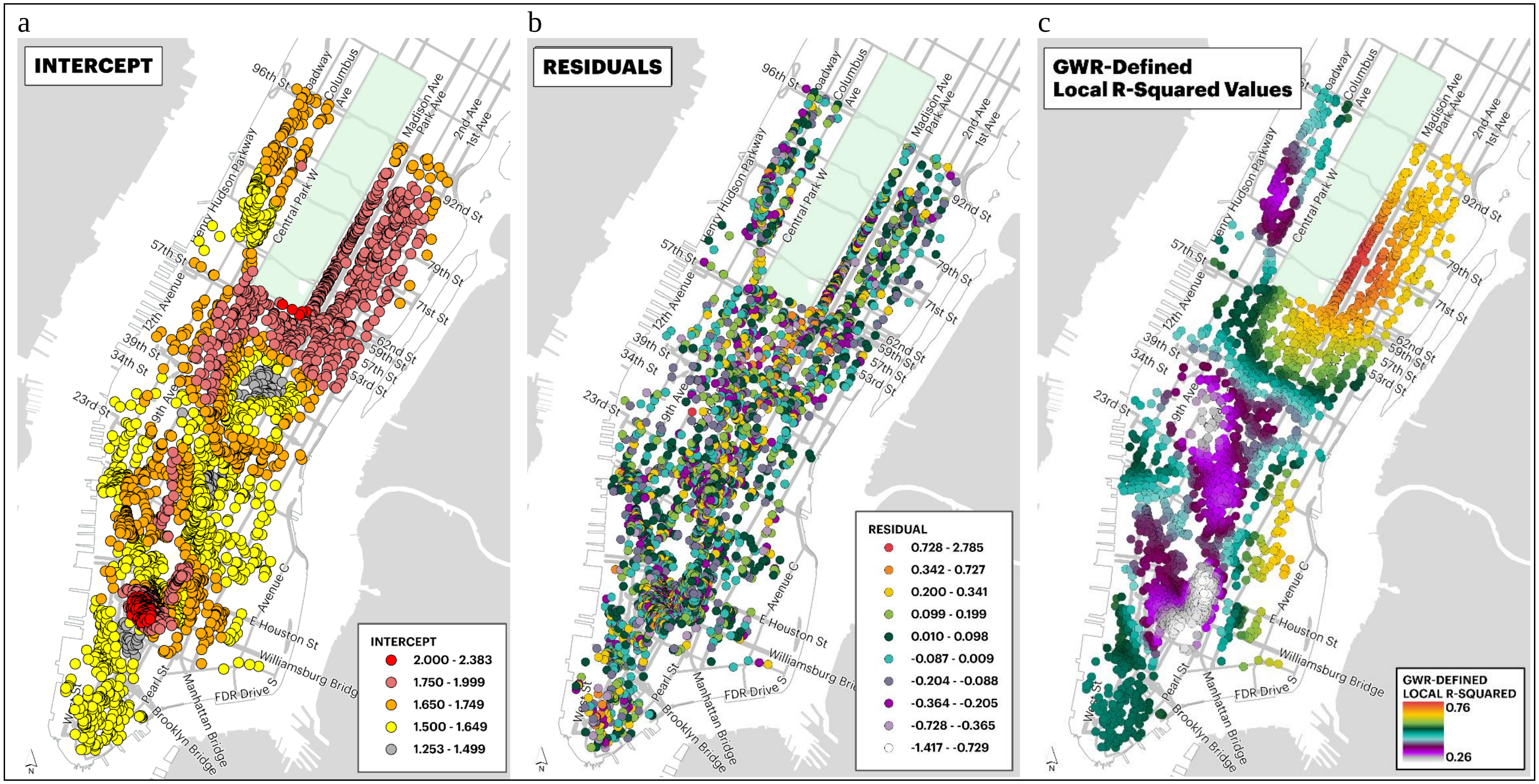
Name	Value
Kernel function	bisquare
Adaptive bandwidth	389 (number of nearest neighbours)
Regression points	the same locations as observations are used.
Distance Metric	Euclidean distance metric is used.

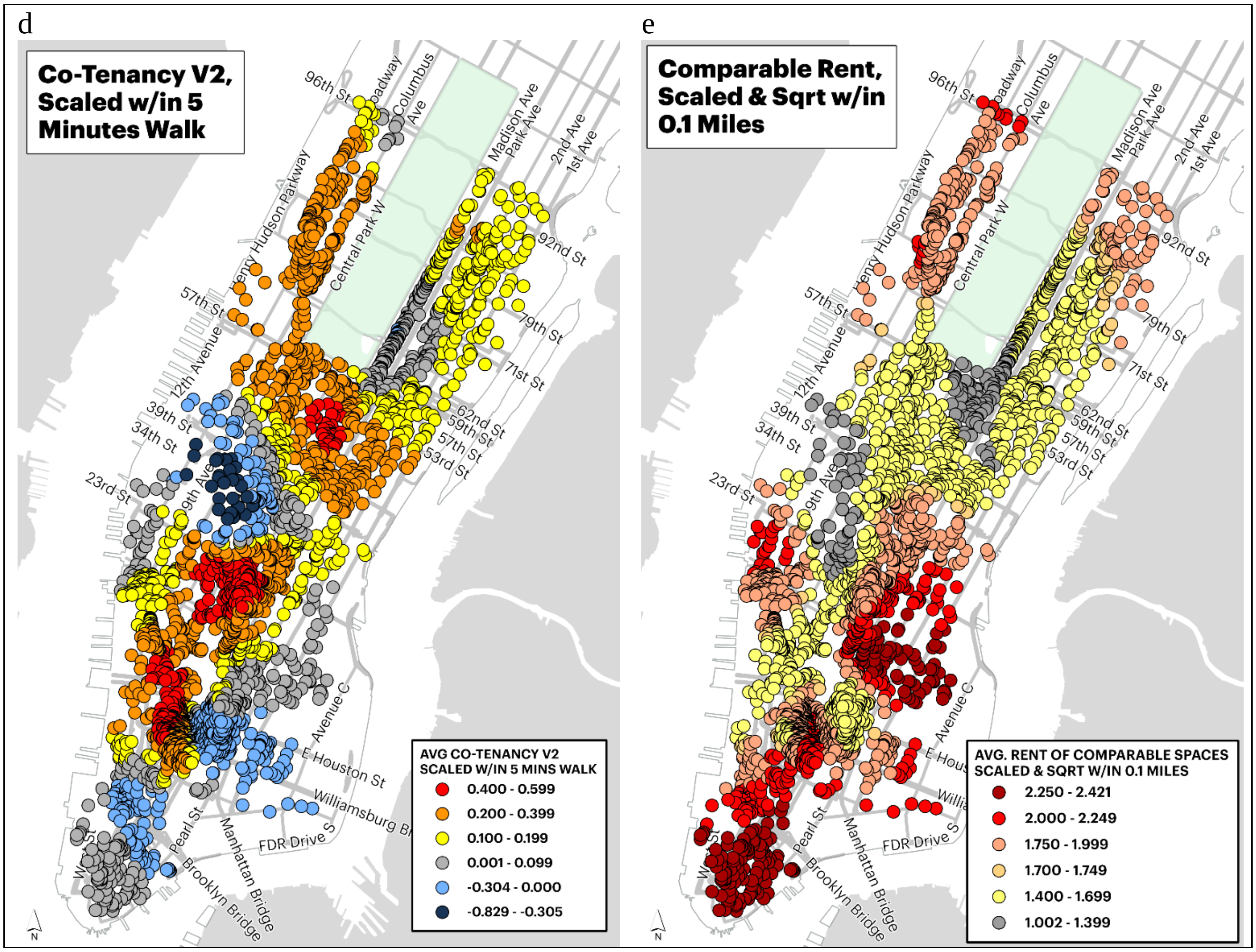
Table 11 Diagnostic information for the GWR model.

Name	Value
Number of Data Points	2,673
Effective number of parameters (2trace(S) - trace(S'S))	98.347
Effective degrees of freedom (n-2trace(S) + trace(S'S))	2,574.653
AIC	-650.611
AICc	-569.198
BIC	-2,807.377
RSS	119.3013
R-square value	0.528
Adjusted R-square value	0.510

Table 12 Summary of GWR coefficients.

Name	Min.	1st Qu.	Median	3rd Qu.	Max.
Intercept	1.253	1.596	1.671	1.784	2.383
Walk 5 – Co Ten V2	-0.829	0.057	0.178	0.297	0.611
Rad 0.1 – Mean PSF Rent Adj (Sqrt)	1.002	1.544	1.764	1.958	2.421
Corner	0.012	0.056	0.077	0.104	0.304
Walk 10 - Daytime Population Density (Sqrt)	-2.373	-0.226	-0.106	0.053	1.423


a
INTERCEPT
INTERCEPT
2.000 - 2.383
1.750 - 1.999
1.650 - 1.749
1.500 - 1.649
1.253 - 1.499
b
RESIDUALS
RESIDUAL
0.728 - 2.785
0.342 - 0.727
0.200 - 0.341
0.099 - 0.199
0.010 - 0.098
-0.087 - 0.009
-0.204 - -0.088
-0.364 - -0.205
-0.728 - -0.365
-1.417 - -0.729
c
GWR-Defined
Local R-Squared Values
GWR-DEFINED
LOCAL R-SQUARED
0.76
0.26


d
Co-Tenancy V2, Scaled w/in 5 Minutes Walk
AVG CO-TENANCY V2 SCALED W/IN 5 MINS WALK
0.400 - 0.599
0.200 - 0.399
0.100 - 0.199
0.001 - 0.099
-0.304 - 0.000
-0.829 - -0.305
e
Comparable Rent, Scaled & Sqrt w/in 0.1 Miles
AVG. RENT OF COMPARABLE SPACES SCALED & SQRT W/IN 0.1 MILES
2.250 - 2.421
2.000 - 2.249
1.750 - 1.999
1.700 - 1.749
1.400 - 1.699
1.002 - 1.399

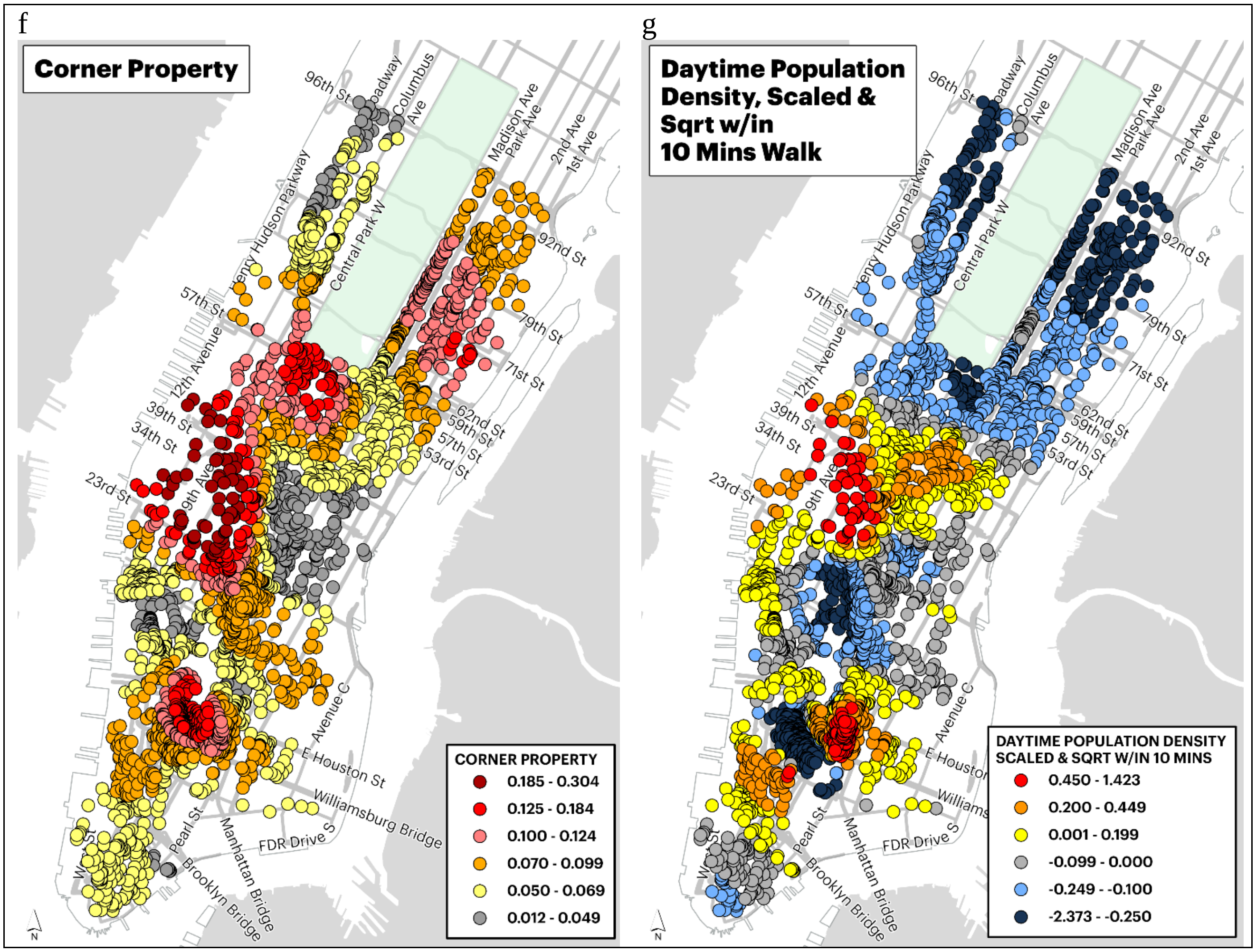


Figure 24 Plotted coefficients for GWR predictor variables.

Finally, clustering points based on the dual K-means approach value should remove or reduce spatial autocorrelation compared to the global dataset. The Moran's I on the residuals globally results in 0.262 which is statistically significant at a 99% confidence interval. Tables 13, 14, and 15 below show the results of the Moran's I test after running OLS using the variables in the optimal GWR model.

Table 13 Moran's I result for natural Jenks delineated neighborhoods.

Neighborhood	Moran's I	Standard Deviate	P Value	Expectation	Variance	Type
Global	0.262	20.457	<2.20E-16	-0.0004	0.0002	99%
J 1b	-0.028	-0.466	0.679	-0.0055	0.0023	Removed
J 1e	-0.050	-0.774	0.780	-0.0073	0.0030	Removed
J 2a	0.109	2.018	0.022	-0.0083	0.0034	95%
J 2c	-0.200	-2.064	0.980	-0.0196	0.0076	Removed
J 2d	-0.015	-0.115	0.546	-0.0083	0.0035	Removed
J 3c	-0.048	-1.455	0.927	-0.0022	0.0010	Removed
J 3d	-0.006	-0.019	0.507	-0.0049	0.0020	Removed
J 4a	-0.037	-0.549	0.708	-0.0072	0.0030	Removed
J 4b	0.158	3.464	0.000	-0.0055	0.0022	99%
J 4e	-0.021	-0.429	0.666	-0.0038	0.0017	Removed
J 5a	-0.214	-2.527	0.994	-0.0154	0.0062	Removed
J 5b	0.080	2.425	0.008	-0.0035	0.0012	99%
J 5d	0.196	4.013	0.000	-0.0060	0.0025	99%
J 6c	0.163	3.336	0.000	-0.0063	0.0026	99%
J 6e	0.177	3.342	0.000	-0.0074	0.0030	99%

Table 14 Moran's I result for quantile delineated neighborhoods.

Neighborhood	Moran's I	Standard Deviate	P Value	Expectation	Variance	Type
Global	0.262	20.457	<2.20E-16	-0.0004	0.0002	99%
Q 1c	-0.0085	-0.061	0.525	-0.00549	0.00234	Removed
Q 1d	0.0172	0.446	0.328	-0.00694	0.00293	99%
Q 1e	0.0180	0.512	0.304	-0.00469	0.00196	Removed
Q 2a	-0.1130	-2.071	0.981	-0.00637	0.00265	99%
Q 2b	-0.1172	-2.358	0.991	-0.00529	0.00225	Removed
Q 3c	0.0058	0.240	0.405	-0.00735	0.00301	99%
Q 4a	-0.0474	-1.098	0.864	-0.00370	0.00159	99%
Q 4b	-0.0469	-0.674	0.750	-0.00806	0.00332	99%
Q 4d	-0.0302	-0.348	0.636	-0.00901	0.00370	99%
Q 4e	-0.0026	0.036	0.486	-0.00412	0.00178	99%
Q 5a	-0.0104	-0.040	0.516	-0.00806	0.00338	95%
Q 5c	0.0569	1.301	0.097	-0.01389	0.00296	Removed
Q 5d	0.1295	2.321	0.010	-0.00862	0.00354	Removed
Q 5e	-0.2038	-2.587	0.995	-0.01333	0.00542	99%
Q 6b	0.2096	7.191	0.000	-0.00201	0.00087	Removed

Table 15 Moran's I result for equal interval delineated neighborhoods.

Neighborhood	Moran's I	Standard Deviate	P Value	Expectation	Variance	Type
Global	0.262	20.457	<2.20E-16	-0.0004	0.0002	99%
E 1c	-0.055	-0.855	0.804	-0.0074	0.0031	Removed
E 1d	-0.028	-0.469	0.681	-0.0056	0.0023	Removed
E 2a	-0.199	-2.028	0.979	-0.0200	0.0078	Removed
E 2b	-0.016	-0.126	0.550	-0.0084	0.0035	Removed
E 2e	0.125	2.228	0.013	-0.0093	0.0037	95%
E 3a	0.006	0.231	0.409	-0.0051	0.0021	Removed
E 3c	-0.162	-2.434	0.993	-0.0102	0.0039	Removed
E 3e	-0.063	-1.783	0.963	-0.0026	0.0011	Removed
E 4b	-0.072	-1.361	0.913	-0.0057	0.0024	Removed
E 4d	-0.093	-1.039	0.851	-0.0143	0.0057	Removed
E 5a	0.056	1.424	0.077	-0.0042	0.0018	Removed
E 5b	0.053	1.432	0.076	-0.0057	0.0017	Removed
E 5c	-0.170	-2.748	0.997	-0.0082	0.0035	Removed
E 5d	-0.006	0.067	0.473	-0.0103	0.0042	Removed
E 6e	0.219	7.585	0.000	-0.0020	0.0008	99%

6.3 Best Model for Each Neighborhood

6.3.1. Natural Jenk-Delineated Neighborhoods

Table 16 below shows variable names of the best models for natural Jenks delineated neighborhoods with Table 17 listing each neighborhood variables' coefficients and an adjusted R-squared value for the 80% calibration subset and for the 20% validation subset. Table 18 shows model fit diagnostics. The number in the parenthesis represents the best model ID for that neighborhood and an asterisk signifies that the model was run using the non-transformed version of adjusted rent as the dependent variable. Spatial lag and spatial error, indicated by SL and SE, are applied to neighborhoods that have a statistically significant Moran's I value greater than 0.15 (positive or negative). Not all neighborhoods yielded a suitable model, indicated by rows shaded in black. For ease of reading, the Jenks natural breaks neighborhood map is printed again below as Figure 25.

Table 16 Natural Jenks neighborhood-model variables.

Name	Var 1	Var 2	Var 3	Var 4
J 1b (26)	Rad 0.1 - Mean PSF Rent Adj	Walk 10 - Population Density		
J 1e (29)	Rad 0.1 - Mean PSF Rent Adj (0.2)	Corner		
J 2a (47)	Walk 5 - Mean PSF Rent Adj (0.1)	Rad 0.25 - CoTen V2	Rad 0.1 - Daytime Population Density	Walk 5 - Avg. HH Retail Spending
J 2c (NA)	[redacted]			
J 2d (36)	Rad 0.25 - Population Density	Rad 0.25 - CoTen V1	Rad 0.1 - Avg. HH Retail Spending	
J 3c (45)	Distance to Subway (0.2)	Rad 0.1 - Mean PSF RENT ADJ		
J 3d (39)	Walk 10 - CoTen V1	Walk 5 - Mean PSF RENT ADJ (2.1)		
J 4a (30)	Rad 0.25 - CoTen V1	Walk 5 - Avg. HH Retail Spending	Walk 10 - Avg HH Income	Rad 0.1 - Population Density
J 4b (29) (SL)	Walk 5 - Mean PSF Rent Adj	Rad 0.1 - Population density (log 10)		
J 4e (40)	Rad 0.25 - Mean PSF Rent Adj	Walk 5 - Population Density (log10)	Walk 5 - CoTen V2 (1.1)	Corner
J 5a (NA)	[redacted]			
J 5b (36)	Walk 10 - CoTen V1	Rad 0.1 - Mean PSF RENT ADJ (1.2)	Rad 0.1 - Avg. HH Income	Walk 10 - Daytime Population Density
J 5d (26) (SE)	Walk 10 - Daytime Population Density	Walk 10 - CoTen V2 (3.0)		
J 6c (40) (SE)	Walk 10 - Mean PSF Rent Adj (2.7)	Rad 0.25 - Avg. HH Income	Corner	
J 6e (40) (SE)	Rad 0.25 - Mean PSF Rent Adj (3.0)	Walk 5 – Daytime Population Density		

Table 17 Natural Jenk neighborhood-level coefficients.

Neighborhood	Intercept	Var 1	Var 2	Var 3	Var 4	Adj. R2 (80%)	Adj. R2 (20%)
J 1b (26)	2.593	0.874	-0.747			0.714	0.428
J 1e (29)	0.486	2.786	0.147			0.701	0.815
J 2a (47)	-1.837	3.827	1.534	-1.239	0.903	0.454	0.262
J 2c (NA)	[redacted]						
J 2d (36)	1.952	-0.464	1.066	0.578		0.487	0.184
J 3c (45)	2.301	-0.295	2.139			0.563	0.479
J 3d (39)	2.030	0.351	12.180			0.165	0.073
J 4a (30)	0.618	1.079	2.403	-0.995	0.762	0.491	0.290
J 4b (29) (SL)	1.976	0.768	-0.227			0.483	0.328
J 4e (40)	1.667	1.768	-0.415	0.576	0.088	0.489	0.220
J 5a (NA)	[redacted]						
J 5b (36)	2.113	0.603	3.071	-0.435	-0.523	0.460	0.157
J 5d (26) (SE)	2.524	1.581	1.285	0.024		0.326	0.237
J 6c (40) (SE)	2.060	1.293	0.318	0.146		0.283	0.090
J 6e (40) (SE)	1.702	3.378	4.192			0.266	0.358

Table 18 Natural Jenk neighborhood-level model fit diagnostics.

Neighborhood	RMSE	Mean Abs. Err	Median Abs. Err	Hetero-skedasticity	Bias	Bias STD
J 1b (26)	0.181	0.147	0.135	No	1.08E-15	4.2E-14
J 1e (29)	0.160	0.129	0.110	Yes	-3.49E-16	-1.4E-14
J 2a (47)	0.302	0.228	0.167	Yes	1.79E-16	6.7E-15
J 2c (NA)	[redacted]					
J 2d (36)	0.147	0.122	0.105	No	-2.10E-16	-9.0E-15
J 3c (45)	0.172	0.137	0.116	No	-7.88E-18	-3.3E-16
J 3d (39)	0.183	0.148	0.128	No	1.69E-15	7.5E-14
J 4a (30)	0.156	0.121	0.096	No	-2.07E-16	-9.0E-15
J 4b (29) (SL)	0.139	0.114	0.112	Yes	-1.57E-15	-6.9E-14
J 4e (40)	0.164	0.134	0.116	No	5.84E-16	2.5E-14
J 5a (NA)	[redacted]					
J 5b (36)	0.183	0.149	0.130	No	-1.62E-15	-6.9E-14
J 5d (26) (SE)	0.197	0.159	0.139	No	1.26E-15	5.1E-14
J 6c (40) (SE)	0.149	0.120	0.112	No	2.49E-16	1.1E-14
J 6e (40) (SE)	0.230	0.187	0.160	No	-1.27E-15	-5.2E-14

Out of the 15 models, two have an R-squared value above 0.7, one above 0.5, six above 0.4, and the remaining below. The top six models were all OLS models as GWR successfully removed all spatial autocorrelation. Neighborhood 1b, which overlaps with Midtown East and the Upper East Side from 59^{th} Street to 65^{th} Street is modeled as a function of rent of nearby similar spaces within one-tenth of a mile and population density within a 10-minutes' walk. Its R-squared value is 0.71. In this neighborhood, the rent of nearby spaces has a positive effect and population density has a negative effect, both with a similar magnitude of coefficient. Neighborhood 1e, which sits above 1b until approximately 80^{th} Street, is modeled as a function of rent of nearby similar spaces within one-tenth of a mile and if the space is on the corner. In this neighborhood, both predictors affect rent positively, however, average rent of comparable spaces has around 26 times the magnitude of influence on the independent variable compared to if the space is on the corner or not.

Referencing Figure 25 below, neighborhoods 5b and 3c are split along Broadway, with 3c on the northeast side and 5b on the southwest side. Broadway is a main throughway of Manhattan that stretches from the very southern point to the top of the island. Neighborhood 3c, which encompasses a majority of Midtown and the southwest portion of Times Square, is modeled with the rent of nearby comparable spaces within one-tenth of a mile (positive coefficient) and the distance to the nearest subway (negative coefficient). This neighborhood has an R-squared value of 0.56. Neighborhood 5b, which covers Nomad, Flatiron, Union Square, parts of Chelsea and Greenwich Village, has a mixed office, retail, and residential theme and is modeled as function of co-tenancy perspective one within a 10-minutes' walk, the rent of nearby comparable spaces within one-tenth of a mile, average household income within one-tenth of a mile, and daytime population density within a 10-minutes' walk. Co-tenancy has a moderate positive influence on rent. Comparable spaces have the largest positive influence. Both average household income and daytime population density have a moderate negative influence on rent. This neighborhood has an R-squared value of 0.46.

Neighborhood 2c and 5a produced no models with at least two significant variables and are therefore left blank. Neighborhood 2c includes the southeastern edge of the dataset encompassing from the intersection of East 23rd Street and 2nd Avenue moving south towards Delancey Street and Avenue D. 5a includes Broadway from 65th Street until approximately 85th Street. This could be because of variable misspecification where other factors may be significant that were unable to be considered for this study. Another possibility is this GWR model is not effective at grouping together points in this region. Overall, 789 points in six neighborhoods cannot be effectively modeled (lower than 0.4 R-squared) or cannot be modeled at all by GWR groupings, representing 29.6% of the dataset.

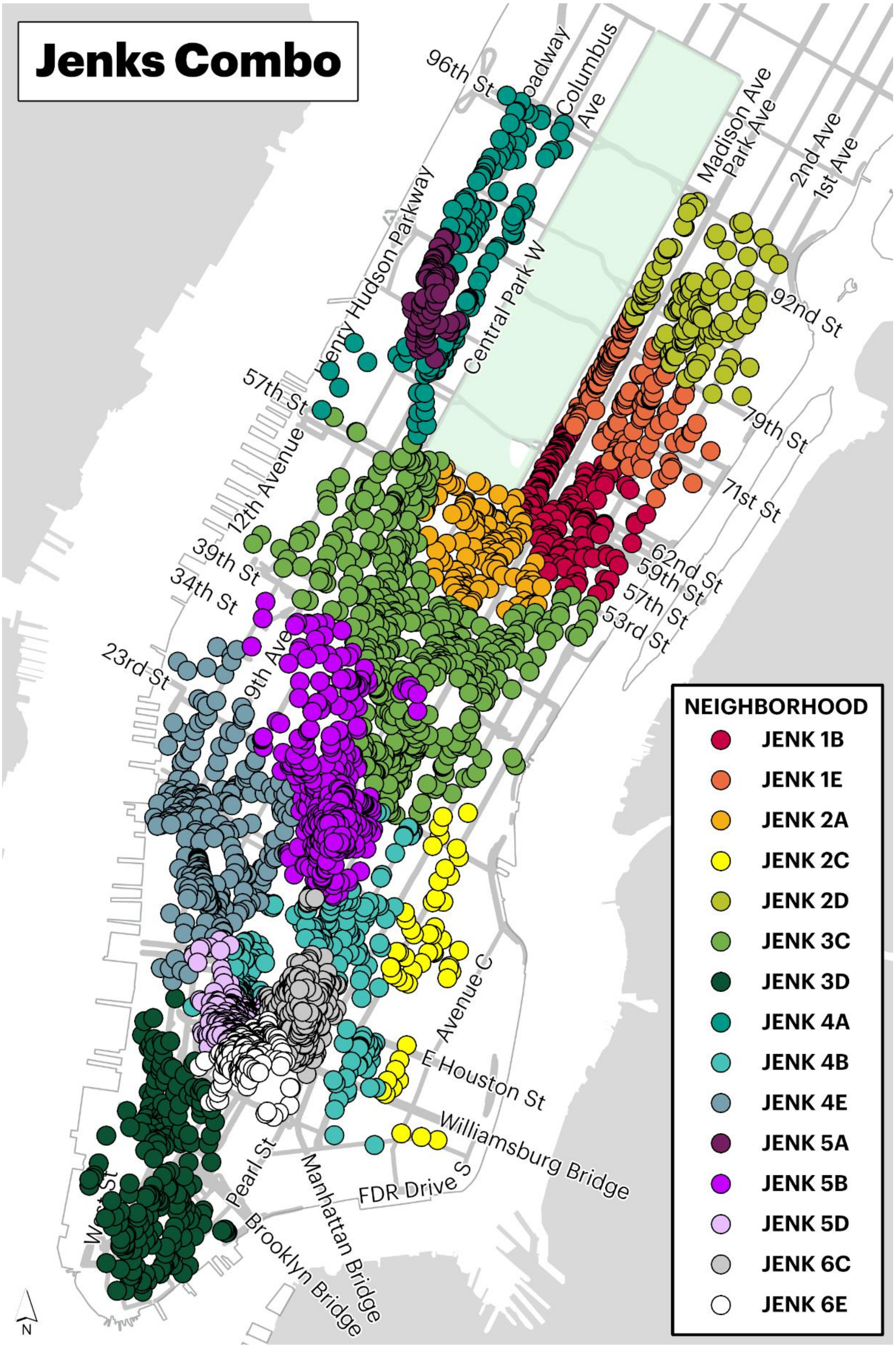


Figure 25 Dual k-means natural Jenks delineated neighborhoods.

6.3.2. Quantile-Delineated Neighborhoods

Table 19 shows the variable names of the best models for quantile delineated neighborhoods with Table 20 listing each coefficient and an adjusted R-squared value for the 80% calibration subset and for the 20% validation subset. Table 21 shows model fit diagnostics. The number in the parenthesis represents the best model ID for that neighborhood and an asterisk signifies that the model was run using the non-transformed version of adjusted rent as the dependent variable. Spatial lag and spatial error, indicated by SL and SE are applied to neighborhoods that have a statistically significant Moran’s I value greater than 0.15 (positive or negative). For ease of reading, the quantile neighborhood map is printed again below as Figure 26.

Table 19 Quantile neighborhood-model variables.

Name	Var 1	Var 2	Var 3	Var 4	Var 5
Q 1c (29)	Rad 0.1 - Mean PSF Rent Adj (0.1)	Corner			
Q 1d (48)	Walk 5 - Mean PSF Rent Adj	Corner	Rad 0. 1 - CoTen V2 (3.0)		
Q 1e (41)	Walk 5 - Avg HH Income	Rad 0.25- Population Density (0.5)	Rad 0.25 - Mean PSF Rent Adj		
Q 2a (38)	Corner	Rad 0.1 - Mean PSF Rent Adj			
Q 2b (43)	Rad 0.1 - Mean PSF Rent Adj	Walk 10 - CoTen V1 (3.0)	Walk 10 - Daytime Population Density		
Q 3c (15)*	Rad 0.25 - Daytime Population Density	Walk 10 - Co Ten V2			
Q 4a (39)	Rad 0.1 - Mean PSF Rent Adj	Rad 0.1 - CoTen V2 (0.1)			
Q 4b (40)	Walk 10 - Avg HH Income	Walk 10 - Mean PSF Rent Adj (0.3)	Walk 5 - Population Density	Walk 10 - CoTen V2	
Q 4d (1)	Rad 0.1 - Mean PSF Rent Adj (log 10)	Corner			
Q 4e (33)	Corner	Rad 0.1 – Mean PSF Rent Adj			
Q 5a (34)	Corner	Rad 0.1 – Mean PSF Rent Adj			
Q 5c (4)*	Rad 0.1 - Mean PSF Rent Adj (2.2)	Walk 5 - CoTen V2 (1.8)	Walk 5 - Avg. HH Retail Spending	Walk 5 - Population Density	Corner
Q 5d (43)	Walk 10 - CoTen V2 (3.0)	Rad 0.1 - Population Density	Rad 0.25 - Daytime Population Density		
Q 5e (39)	Walk 5 - CoTen V2	Walk 5 - Avg. HH Retail Spending (log 10)			
Q 6b (50) SL	Rad 0.1 - Mean PSF Rent Adj	Corner			

Table 20 Quantile neighborhood-level coefficients.

Neighborhood	Intercept	Var 1	Var 2	Var 3	Var 4	Var 5	Adj. R2 (80%)	Adj. R2 (20%)
Q 1c (29)	-1.444	4.681	0.097				0.772	0.673
Q 1d (48)	1.965	2.115	0.098	0.707			0.466	0.371
Q 1e (41)	2.262	0.462	-0.590	0.774			0.674	0.384
Q 2a (38)	1.893	0.065	4.343				0.432	0.121
Q 2b (43)	2.098	1.790	1.192	-0.334			0.663	0.637
Q 3c (15)*	2.692	101.848	338.043				0.123	0.157
Q 4a (39)	0.969	2.478	1.191				0.453	0.537
Q 4b (40)	0.664	-0.780	1.337	1.474	1.943		0.419	0.246
Q 4d (1)	2.796	0.574	0.078				0.355	0.047
Q 4e (33)	1.994	0.069	2.950				0.591	0.376
Q 5a (34)	1.830	0.103	4.520				0.433	0.065
Q 5c (4)*	137.705	10,707.656	-320.540	564.573	-665.939	-60.084	0.914	0.793
Q 5d (43)	2.593	0.901	-1.184	-0.876			0.344	0.130
Q 5e (39)	2.947	1.021	5.451				0.198	-0.011
Q 6b (50) SL	2.010	1.760	0.081				0.390	0.257

Table 21 Quantile neighborhood-level model fit diagnostics.

Neighborhood	RMSE	Mean Abs. Err	Median Abs. Err	Hetero-skedasticity	Bias	Bias STD
Q 1c (29)	0.154	0.126	0.115	Yes	-2.61E-15	-1.00E-13
Q 1d (48)	0.149	0.123	0.114	No	-4.44E-16	-1.90E-14
Q 1e (41)	0.215	0.171	0.153	Yes	-7.59E-16	-3.02E-14
Q 2a (38)	0.123	0.098	0.092	No	1.24E-15	5.67E-14
Q 2b (43)	0.192	0.153	0.130	No	7.08E-17	2.88E-15
Q 3c (15)*	98.173	76.289	70.159	No	-1.18E-13	-6.30E-14
Q 4a (39)	0.168	0.133	0.113	No	8.29E-16	3.54E-14
Q 4b (40)	0.152	0.123	0.107	No	5.06E-16	2.15E-14
Q 4d (1)	0.131	0.104	0.088	No	1.46E-16	6.62E-15
Q 4e (33)	0.147	0.117	0.111	No	-5.18E-16	-2.20E-14
Q 5a (34)	0.167	0.130	0.112	No	8.17E-16	3.44E-14
Q 5c (4)*	65.789	55.008	52.896	No	-2.97E-14	-1.03E-14
Q 5d (43)	0.169	0.138	0.143	No	3.16E-16	1.36E-14
Q 5e (39)	0.172	0.140	0.119	No	5.98E-16	2.48E-14
Q 6b (50) SL	0.194	0.156	0.136	Yes	1.97E-16	8.06E-15

Except for neighborhood 6b, all neighborhoods do not have statistically significant spatial autocorrelation above a Moran's I value of 0.15. One model has an R-squared value over 0.9, one over 0.7, two above 0.6, one above 0.5, five above 0.4, and the remaining below. Quantile 5c has the highest R-squared value of 0.91 and is modeled using five variables against the non-transformed version of rent. Referencing Figure 26, neighborhood 5c has the highest R-squared value and is comprised of Midtown South approximately from 9th Avenue to Madison Avenue from 38th Street to 9th Street. This is a majority office-centric neighborhood with some residential and is modeled as a function of rent of nearby spaces within one-tenth of a mile, co-tenancy perspective two within one-tenth of a mile, average annual household spending on retail goods within a five-minutes' walk, population density within a five-minutes' walk, and if the property is on the corner. Rent of comparable spaces has the strongest positive influence on rent, co-tenancy has a moderately negative effect, retail spending has a moderately positive, population density a has the strongest negative effect on rent, and finally, being on the corner has a mildly negative effect on rent.

Neighborhood 4b, which runs up Columbus Avenue of the Upper West Side, expanding west to Broadway from 79th Street to 96th Street, is modeled as a function of average household income within a 10-minutes' walk, rent of nearby spaces within a 10-minutes' walk, population density within a 5-minutes' walk, and co-tenancy perspective 2 within a 10-minutes' walk. Income has a moderately negative effect, rent of nearby spaces, population density, and co-tenancy all have about the same magnitude of positive effect on rent, with population density having the highest positive influence. This is one of the few neighborhoods where rent of comparable spaces does not have the highest influence. This neighborhood has an R-squared value of 0.419.

Finally, neighborhood 2a is the most disjoint, separated into three clusters including the eastern edge of the dataset in downtown Manhattan, the west end of Chelsea, and crossing diagonally from the northwest corner of Tribeca towards the east end of the Financial District. This cluster is modeled as a function of nearby comparable rent within one-tenth of a mile and if it is a corner space. Although this neighborhood does not have the worst-modeled neighborhood, it may have been a better fit to separate these three disjoint regions. Overall, 941 points in five neighborhoods have an R-squared value of 0.4 or lower and cannot be effectively modeled by quantile-delineated neighborhoods, representing about 35% of the dataset.

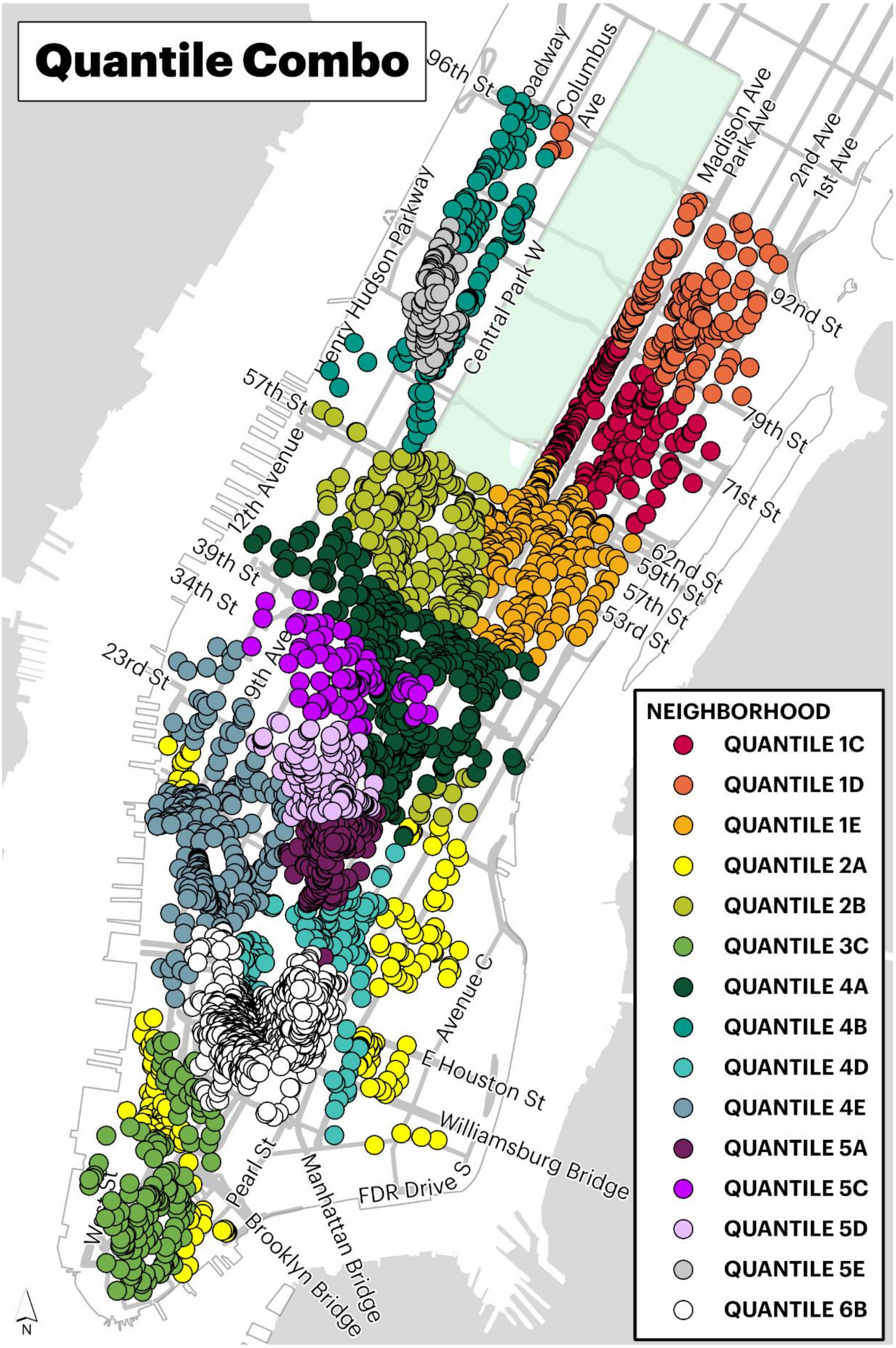


Figure 26 Dual k-means quantile delineated neighborhoods.

6.3.3. Equal Interval-Delineated Neighborhoods

Table 22 below shows the variable names of the best models for quantile delineated neighborhoods with Table 23 listing each coefficient and an adjusted R-squared value for the 80% calibration subset and for the 20% validation subset. Table 24 which shows additional model fit diagnostics. The number in the parenthesis represents the best model ID for that neighborhood and an asterisk means that the model was run using the non-transformed version of adjusted rent as the dependent variable. Spatial lag and spatial error, indicated by SL and SE, are applied to neighborhoods that have a statistically significant Moran's I value greater than 0.15 (positive or negative). Not all neighborhoods yielded a suitable model, indicated by rows shaded in black.

Table 22 Equal interval neighborhood-model variables

Name	Var 1	Var 2	Var 3	Var 4
E 1c (48)	Corner	Rad 0.1 - Mean PSF Rent Adj (0.2)		
E 1d (46)	Walk 10 - Population Density	Walk 10 - Daytime Population density (log 10)	Rad 0.25 - Mean PSF Rent Adj	
E 2a (NA)	[redacted]			
E 2b (29)	Walk 10 - Avg HH Income (3.0)	Walk 10 - Population Density		
E 2e (45)	Walk 5 - Avg. HH Retail Spending	Walk 10 - Daytime Population Density (3.0)	Walk 5 - Mean PSF Rent Adj	Rad 0.1 - Avg HH Income (log 10)
E 3a (NA)	[redacted]			
E 3c (43)	Food	Rad 0.1 - Mean PSF Rent Adj (log 10)		
E 3e (31)	Rad 0.1 - Population Density	Rad 0.25 - Mean PSF Rent Adj	Rad 0.1 - CoTen V2	
E 4b (40)	Walk 5 - CoTen V2 (1.4)	Rad 0.1 - Mean PSF Rent Adj		
E 4d (23)*	Walk 5 - CoTen V1 (log 10)	Rad 0.1 - Avg HH Income	Rad 0.1 - Population Density	
E 5a (35)	Rad 0.1 - Mean PSF Rent Adj	Rad 0.1 - Avg. HH Retail Spending (0.9)	Walk 5 - CoTen V1	
E 5b (23)*	Rad 0.25 - Population Density	Walk 10 -Avg. HH Retail Spending	Walk 5 - Mean PSF Rent Adj	Walk 10 - CoTen V1 (1.6)
E 5c (43)	Rad 0.1 - Mean PSF Rent Adj	Rad 0.1 - Co Ten V1		
E 5d (7)*	Walk 5 - Mean PSF Rent Adj	Rad 0.1 - Daytime Population Density		
E 6e (26) SL	Rad 0.1 - Mean PSF Rent Adj (0.1)	Corner		

Table 23 Equal interval neighborhood-level coefficients.

Neighborhood	Intercept	Var 1	Var 2	Var 3	Var 4	Adj. R2 (80%)	Adj. R2 (20%)
E 1c (48)	0.530	0.141	2.740			0.714	0.623
E 1d (46)	2.557	-0.852	-0.190	0.475		0.674	0.731
E 2a (NA)							
E 2b (29)	2.533	1.696	-0.769			0.466	0.156
E 2e (45)	0.060	2.648	0.635	0.740	-0.773	0.508	0.234
E 3a (NA)							
E 3c (43)	3.071	0.067	0.822			0.337	0.121
E 3e (31)	2.129	-0.244	1.351	0.153		0.495	0.150
E 4b (40)	1.911	0.430	3.406			0.632	0.329
E 4d (23)*	284.849	234.942	382.282	-345.779		0.370	0.141
E 5a (35)	1.908	2.764	-0.349	0.752		0.415	0.361
E 5b (23)*	43.173	-275.112	400.542	1,184.032	-243.828	0.484	0.401
E 5c (43)	1.802	3.558	0.353			0.310	0.354
E 5d (7)*	-67.881	1,040.050	2956.985			0.341	0.230
E 6e (26) SL	-1.265	3.063	0.088			0.368	0.327

Table 24 Equal interval neighborhood-level model fit diagnostics.

Neighborhood	RMSE	Mean Abs. Err	Median Abs. Err	Hetero-skedasticity	Bias	Bias STD
E 1c (48)	0.147	0.120	0.103	No	-1.60E-15	-6.29E-14
E 1d (46)	0.192	0.152	0.118	No	8.00E-17	3.05E-15
E 2a (NA)						
E 2b (29)	0.145	0.119	0.109	No	0	0
E 2e (45)	0.288	0.230	0.178	No	8.80E-16	3.32E-14
E 3a (NA)						
E 3c (43)	0.116	0.091	0.076	No	-1.40E-15	-6.41E-14
E 3e (31)	0.175	0.137	0.109	No	-1.70E-15	-7.09E-14
E 4b (40)	0.125	0.102	0.089	No	4.09E-16	1.77E-14
E 4d (23)*	47.658	39.853	38.875	No	3.75E-14	2.18E-14
E 5a (35)	0.169	0.136	0.114	No	2.81E-17	1.20E-15
E 5b (23)*	90.587	71.615	58.699	No	-1.27E-13	-5.35E-14
E 5c (43)	0.164	0.129	0.106	No	-1.77E-16	-7.31E-15
E 5d (7)*	90.285	75.097	67.244	No	-9.76E-14	-3.67E-14
E 6e (26) SL	0.196	0.159	0.142	Yes	2.60E-15	1.08E-13

One model has an R-squared value over 0.7, two above 0.6, one above 0.5, four above the 0.4 range, and the remaining below. Neighborhood 6e is the only neighborhood that uses spatial lag to account for remaining spatial effects. Equal 1c has the highest R-squared value of 0.72 and is approximately comprised of the Upper East Side from 66th Street to 79th Street. This neighborhood is best fit with an OLS model comprising if it is on the corner and nearby rent of comparable spaces within one-tenth of a mile. Echoing many other neighborhoods, nearby comparables have a much higher influence on rent. Neighborhood 4b covers Chelsea from 23rd Street through the West Village to Perry Street as well as the border of Greenwich Village up 6th Avenue until 14th Street. With an R-squared value of 0.63, this neighborhood is a function of co-tenancy perspective two within a five-minutes' walk and nearby comparable rent prices within one tenth of a mile. The influence of comparable spaces on rent is roughly eight times stronger than that of co-tenancy.

Neighborhood 5a, located to the east of neighborhood 4b, does not have the highest R-squared value, however its resulting model contains retail-focused variables for understanding rent dynamics. Its model includes rent of nearby comparable spaces within one-tenth of a mile, average annual household spending on retail goods within one-tenth of a mile, and co-tenancy perspective one within a five-minutes' walk. Nearby comparables have the highest positive influence, followed by spending on retail goods with a minor negative influence, and co-tenancy with a moderate positive influence on rent. This neighborhood includes retail hubs Nomad, Flatiron, Union Square and the northeast portion of Greenwich Village. Therefore, it is supportive of the argument that retail-focused variables influence rent in retail neighborhoods.

Neighborhoods 2a and 3c produced no models with at least two significant variables and are therefore left blank. These include the southeastern edge of the dataset in lower Manhattan. This could be because of variable misspecification or that this GWR model is not effective at grouping together points in this region. Further, there is less data available in that area. Using equal interval clusters, 1,150 spaces in seven neighborhoods cannot be accurately estimated because of low R-squared values or because of no resulting model. This represents 43.1% of the dataset.

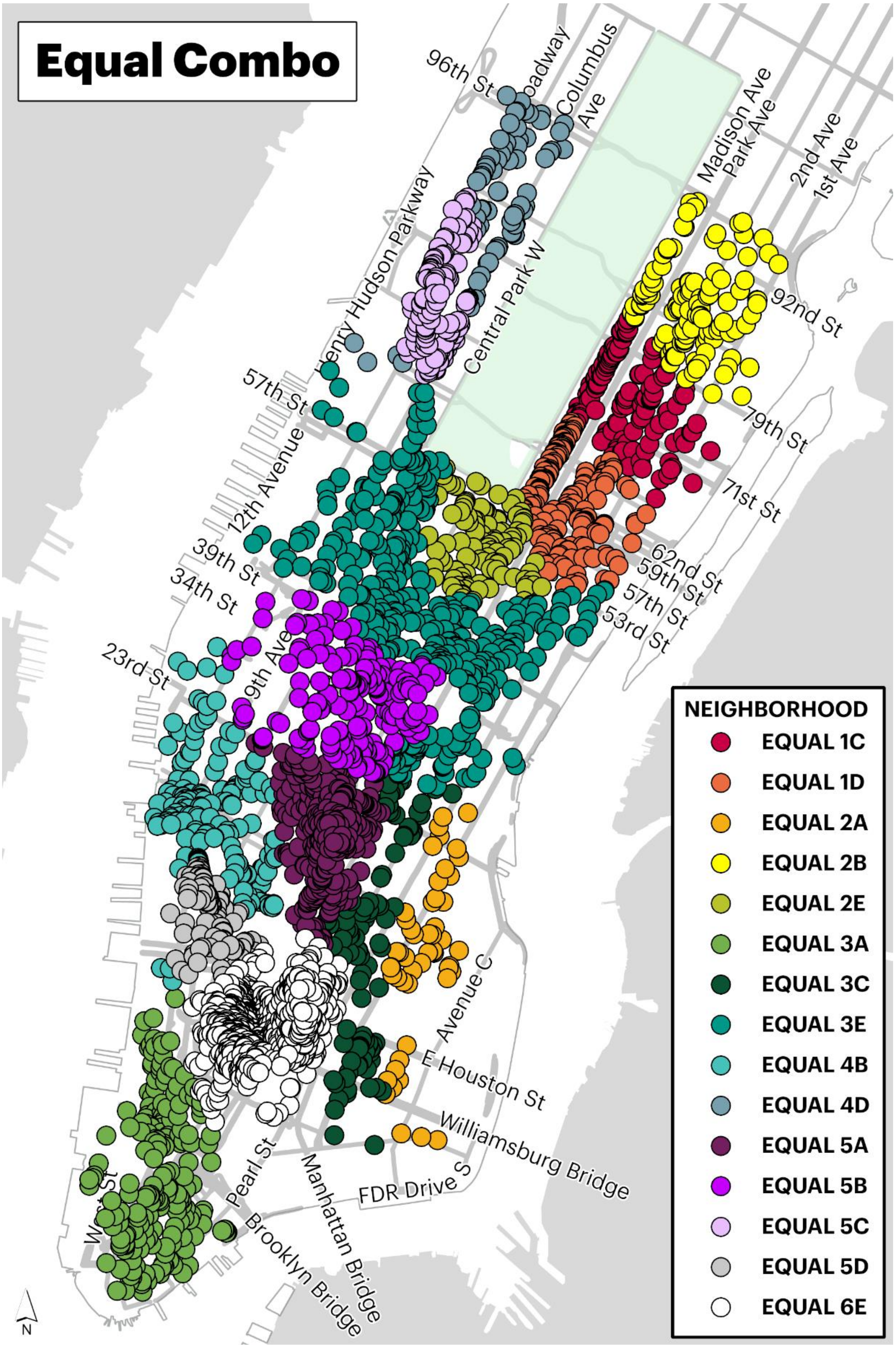


Figure 27 Dual k-means equal interval delineated neighborhoods.

6.3.4. Comparing Across Jenks, Quantile, and Equal Interval

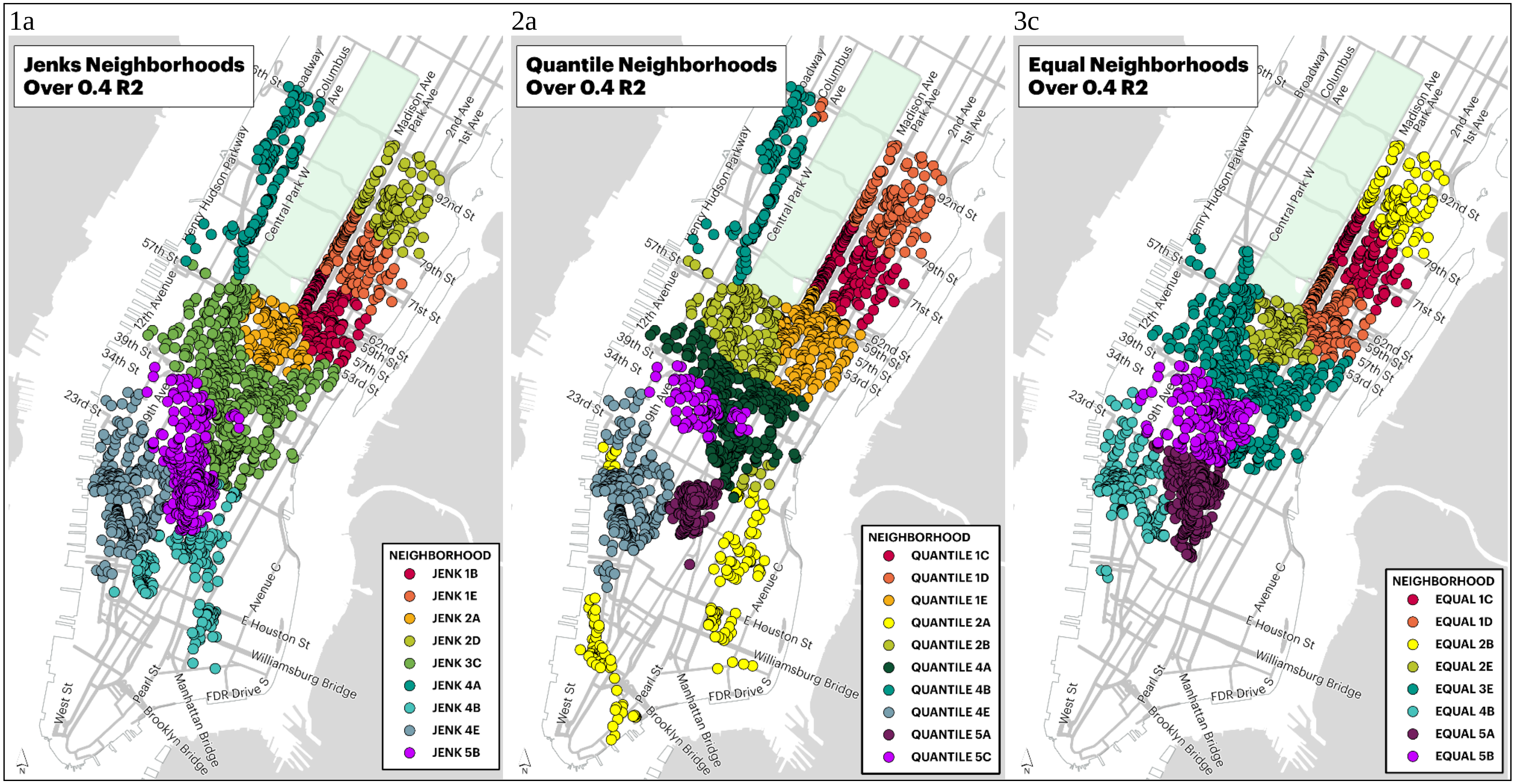
Neighborhoods that are effective ways of clustering to estimate rent take on similar shapes across Jenks, quantile and equal interval delineations. This holds true for the opposite as well. However, there are a few interesting patterns to highlight that break this pattern. Looking at the intersection of Chelsea, Nomad, and the Flatiron District centered around 23rd Street corridor, the quantile cluster 5d produces a model with an R-squared value of 0.344. Referencing Figure 28 below, that neighborhood is split east-to-west down the middle along 23rd Street using equal interval neighborhoods which yields models with 0.48 R-squared on the north side (5b) and 0.42 on the south side (5a). All models that required a spatial lag or spatial error model have an R-squared value lower than 0.4. As referenced in the Moran's I results in Tables 13,14 and 15, using GWR's local R-squared values is an effective way to remove or reduce spatial autocorrelation. Finally, Table 25 shows the number of spaces that are in at least one neighborhood with a model of at least an 0.4 R-squared value.

Table 25 Number of spaces that can be estimated in at least one delineation class.

Type	Percent of N	No. Spaces
in at least 1	75.1%	2,005
in at least 2	65.7%	1,753
in all three	51.3%	1,369

Models with no results include Jenks 2c, Jenks 5a, equal interval 2a, and equal interval 3c. Three out of four of these neighborhoods are on the southeastern edge of the dataset, stretching diagonally across from roughly 23rd Street and Park to Avenue C and Delancey Street. This cuts through Gramercy, East Village, the Lower East Side, and parts of Stuyvesant Town. These three neighborhoods overlap with the quantile 2a neighborhood which has an R-squared value of 0.43. However, as stated, this is a disjoint neighborhood and should have been split up.

Additionally, Jenks 5a, which runs up Broadway from 65th Street through 81st Street does not have a suitable model. This area corresponds to neighborhoods delineated by quantile and equal interval exhibit R-squared values below 0.4.


1a
Jenks Neighborhoods
Over 0.4 R2
NEIGHBORHOOD
JENK 1B
JENK 1E
JENK 2A
JENK 2D
JENK 3C
JENK 4A
JENK 4B
JENK 4E
JENK 5B
2a
Quantile Neighborhoods
Over 0.4 R2
NEIGHBORHOOD
QUANTILE 1C
QUANTILE 1D
QUANTILE 1E
QUANTILE 2A
QUANTILE 2B
QUANTILE 4A
QUANTILE 4B
QUANTILE 4E
QUANTILE 5A
QUANTILE 5C
3c
Equal Neighborhoods
Over 0.4 R2
NEIGHBORHOOD
EQUAL 1C
EQUAL 1D
EQUAL 2B
EQUAL 2E
EQUAL 3E
EQUAL 4B
EQUAL 5A
EQUAL 5B

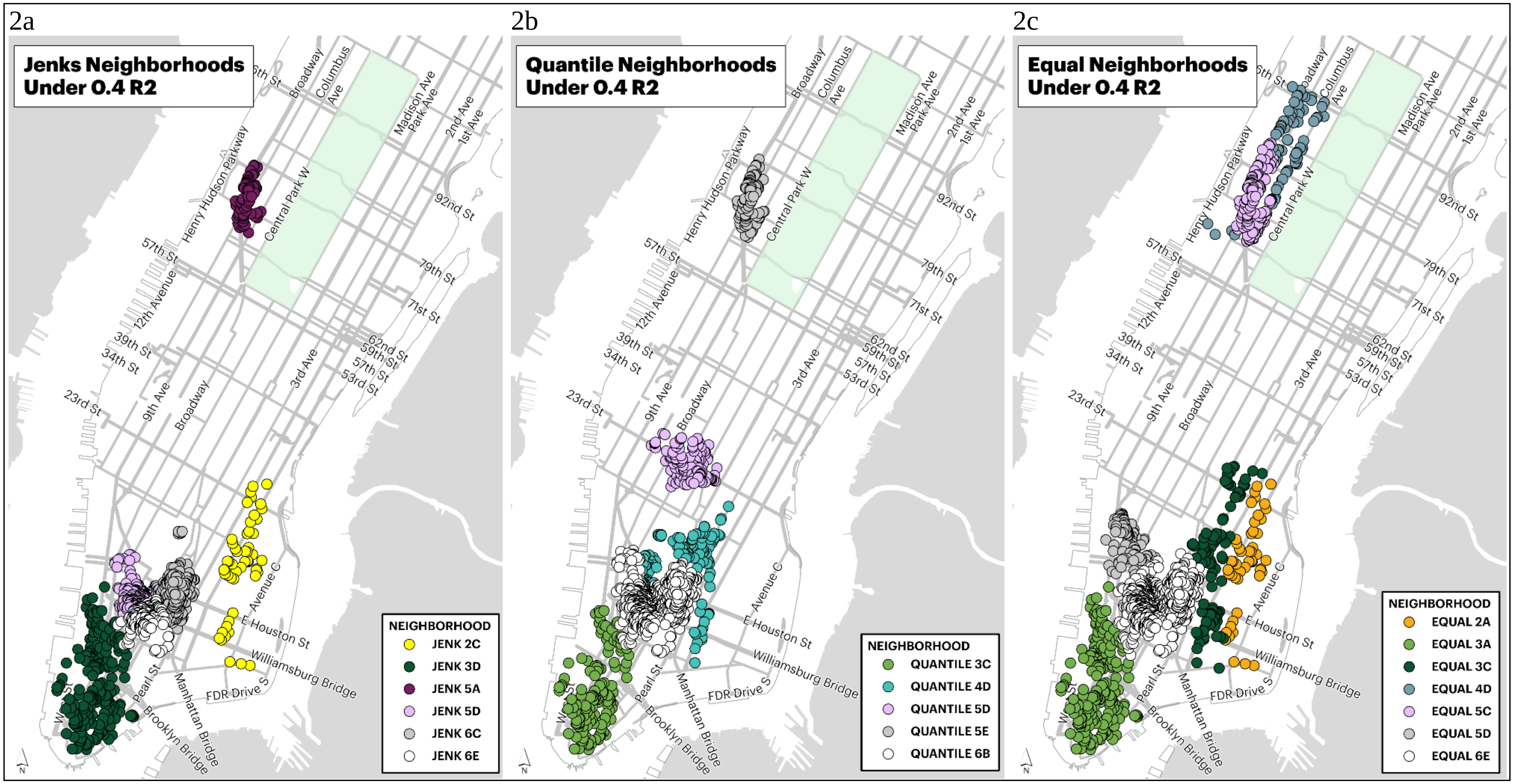


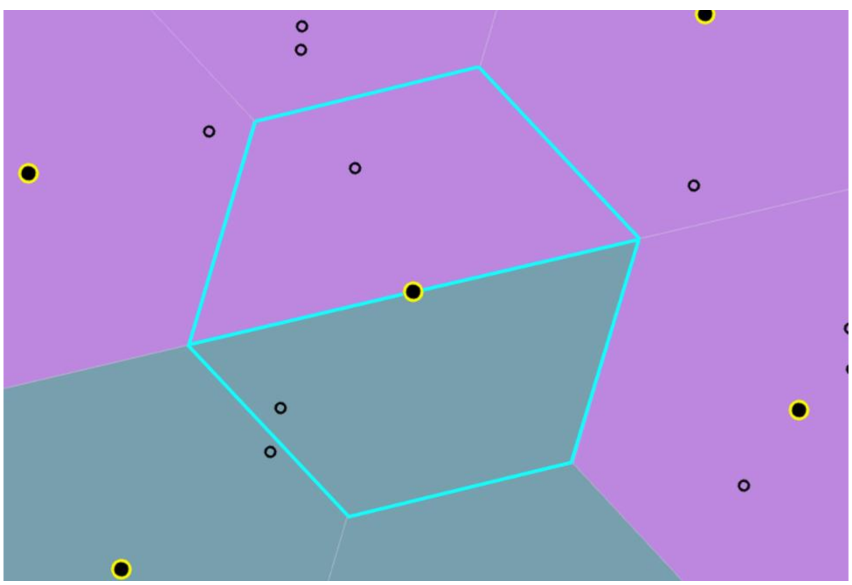
Figure 28 Comparing similarities across neighborhood delineations. Neighborhoods with above an R-squared value of 0.4 on top and below an R-squared value of 0.4 on the bottom.

6.4 RStudio Processing Speed

Using RStudio version 2022.02.7, the code processed 10 model combinations in eight minutes and 20 model combinations in 18 minutes. However, upon upgrading to the latest RStudio release at the time of this writing, version 2023.12.01-402, the processing time for 20 models increased to about 30 minutes. Furthermore, in the older RStudio version, the stepwise OLS iterative function, when applied to 13 neighborhoods categorized into three classes, took around four minutes to process 100 model combinations each. This equates to 3,900 preliminary tests, excluding additional runs needed for addressing insignificant variables, outliers, and heteroskedasticity concerns. However, the OLS iterative function generates 3,900 objects containing the coefficient results for each model, which causes a substantial slowdown in the software's performance for subsequent analyses. After the RStudio upgrade, the time required for the same OLS process escalates to one hour and 30 minutes. Despite the longer processing time, a significant improvement in responsiveness of RStudio provided a much more user-friendly experience. The significant differences observed in processing times post-upgrade warrant further investigation.

6.5 Estimate Rent using an Esri Experience Builder Interactive Map

The Esri Experience Builder online web application hosts all neighborhood-level information in a hexagon-tessellated polygon layer as well as predictor variables in a block group layer. Any predictor within any trade area is enriched using the centroid of the block group (i.e., population density within 0.25 miles and co-tenancy perspective two within a five-minutes' walk are calculated using the centroid). Using the embedded Survey 123 platform, the user can drop a pin and calculate a rent estimation pulling coefficients from the tessellated layer and predictor observation values from the block group layer. Also aggregated within each cell of the tessellated layer is the mean value of the local point GWR coefficients. The web application uses quantile neighborhoods as an example where each cell is assigned to a neighborhood based on the majority neighborhood of the points it encompasses. Tessellated cells with no majority or no points at all are assigned to a neighborhood in two ways. In the first situation, tessellated cells with no majority (i.e. 50% one neighborhood and 50% another neighborhood) are split directionally based on the distribution of points. In the second situation, tessellated cells with no points are split based on which points are closest to the cell's respective edges and corners, connected at the empty cell's centroid. Figure 29 below illustrates these two examples. In total, 9.37 square miles of cells are assigned to any given neighborhood and 7.19 square miles assigned to a neighborhood with a model that has at least an 0.4 R-squared.



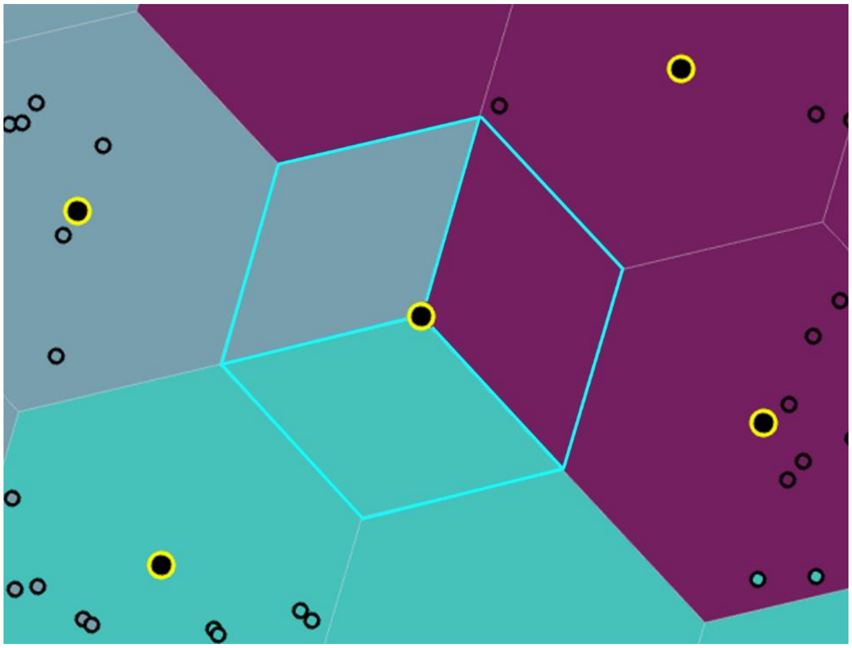


Figure 29 Handling neighborhood assignment of tessellated cells for those with no majority points (top) and no points at all (bottom).

Each hexagon, or subdivided section of a hexagon, in the tessellated map represents a neighborhood and contains model information pertinent to that area. The tessellated neighborhood layer is referenced in Figure 30, with the pop-up for cell Y-22 highlighted in Figure 31. After dropping a point for a space built out for food on West 55th Street, a pop-up displays the actively listed rent, model-estimated rent, and additional information, shown in Figure 32. The model for neighborhood 2b calculated this space's rent as $238.79 per square foot, compared to its listed rent of $190. Only the rent estimation in 2b for food users is relevant. Other neighborhood rent estimations shown in the pop-up are not related to this property. The full interface of the web application is shown in Figure 33.

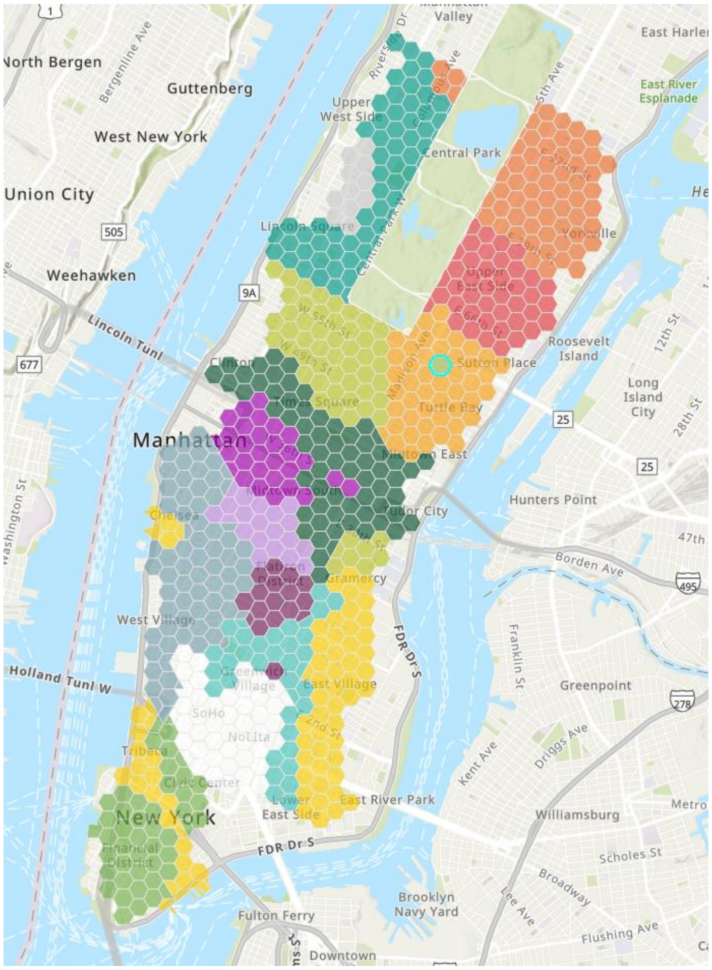
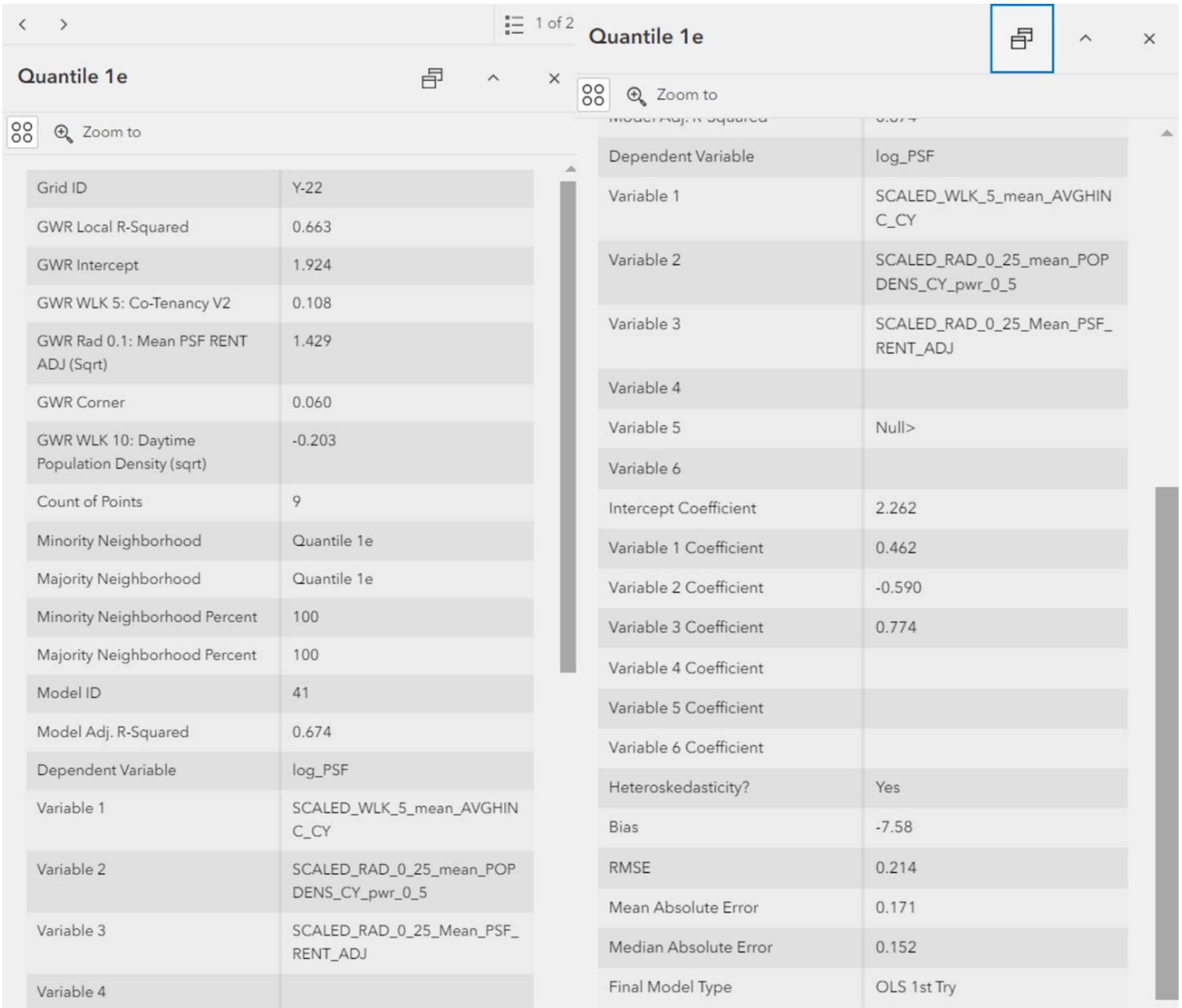


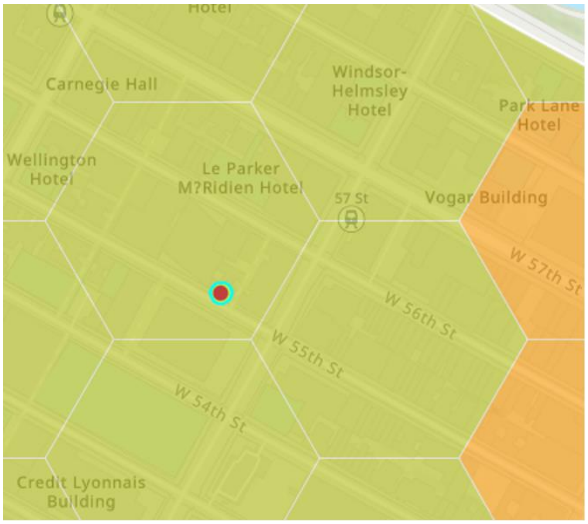
Figure 30 Tessellated hexagons symbolized by neighborhood.

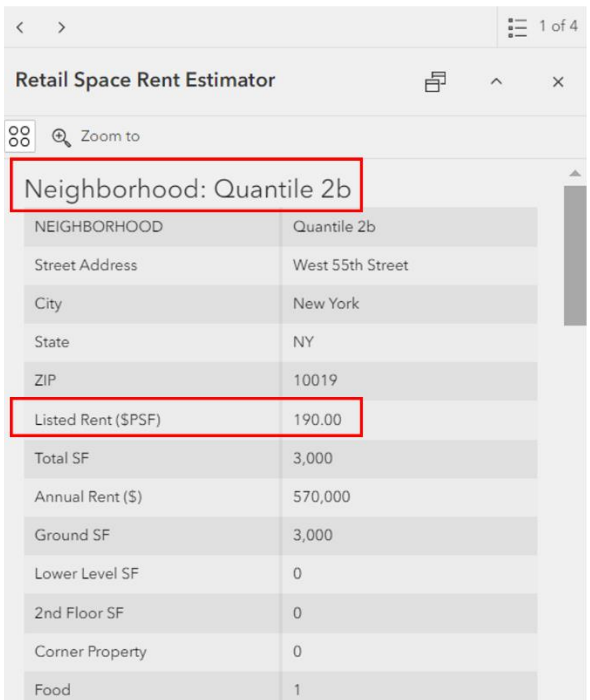


Grid ID	Y-22
GWR Local R-Squared	0.663
GWR Intercept	1.924
GWR WLK 5: Co-Tenancy V2	0.108
GWR Rad 0.1: Mean PSF RENT ADJ (Sqrt)	1.429
GWR Corner	0.060
GWR WLK 10: Daytime Population Density (sqrt)	-0.203
Count of Points	9
Minority Neighborhood	Quantile 1e
Majority Neighborhood	Quantile 1e
Minority Neighborhood Percent	100
Majority Neighborhood Percent	100
Model ID	41
Model Adj. R-Squared	0.674
Dependent Variable	log_PSF
Variable 1	SCALED_WLK_5_mean_AVGHINC_CY
Variable 2	SCALED_RAD_0_25_mean_POPDENS_CY_pwr_0_5
Variable 3	SCALED_RAD_0_25_Mean_PSF_RENT_ADJ
Variable 4	

Dependent Variable	log_PSF
Variable 1	SCALED_WLK_5_mean_AVGHINC_CY
Variable 2	SCALED_RAD_0_25_mean_POPDENS_CY_pwr_0_5
Variable 3	SCALED_RAD_0_25_Mean_PSF_RENT_ADJ
Variable 4	
Variable 5	Null>
Variable 6	
Intercept Coefficient	2.262
Variable 1 Coefficient	0.462
Variable 2 Coefficient	-0.590
Variable 3 Coefficient	0.774
Variable 4 Coefficient	
Variable 5 Coefficient	
Variable 6 Coefficient	
Heteroskedasticity?	Yes
Bias	-7.58
RMSE	0.214
Mean Absolute Error	0.171
Median Absolute Error	0.152
Final Model Type	OLS 1st Try

Figure 31 Tessellated hexagon pop-ups.





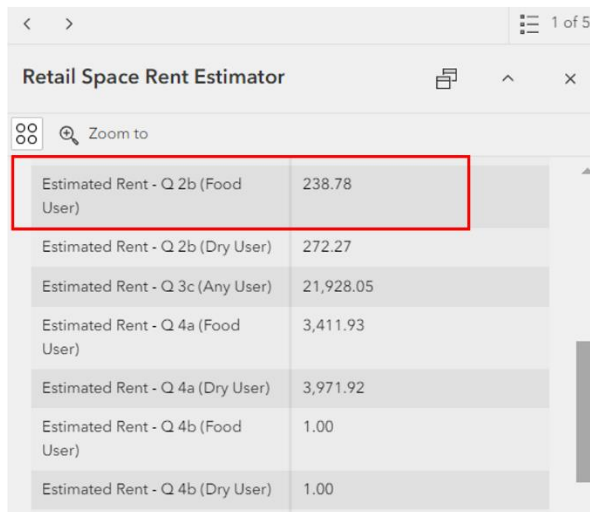
1 of 4

Retail Space Rent Estimator

Zoom to

Neighborhood: Quantile 2b

NEIGHBORHOOD	Quantile 2b
Street Address	West 55th Street
City	New York
State	NY
ZIP	10019
Listed Rent ($PSF)	190.00
Total SF	3,000
Annual Rent ($)	570,000
Ground SF	3,000
Lower Level SF	0
2nd Floor SF	0
Corner Property	0
Food	1



1 of 5

Retail Space Rent Estimator

Zoom to

Estimated Rent - Q 2b (Food User)	238.78
Estimated Rent - Q 2b (Dry User)	272.27
Estimated Rent - Q 3c (Any User)	21,928.05
Estimated Rent - Q 4a (Food User)	3,411.93
Estimated Rent - Q 4a (Dry User)	3,971.92
Estimated Rent - Q 4b (Food User)	1.00
Estimated Rent - Q 4b (Dry User)	1.00

Figure 32 Results for a space's estimated rent as a food-user within neighborhood 2b.

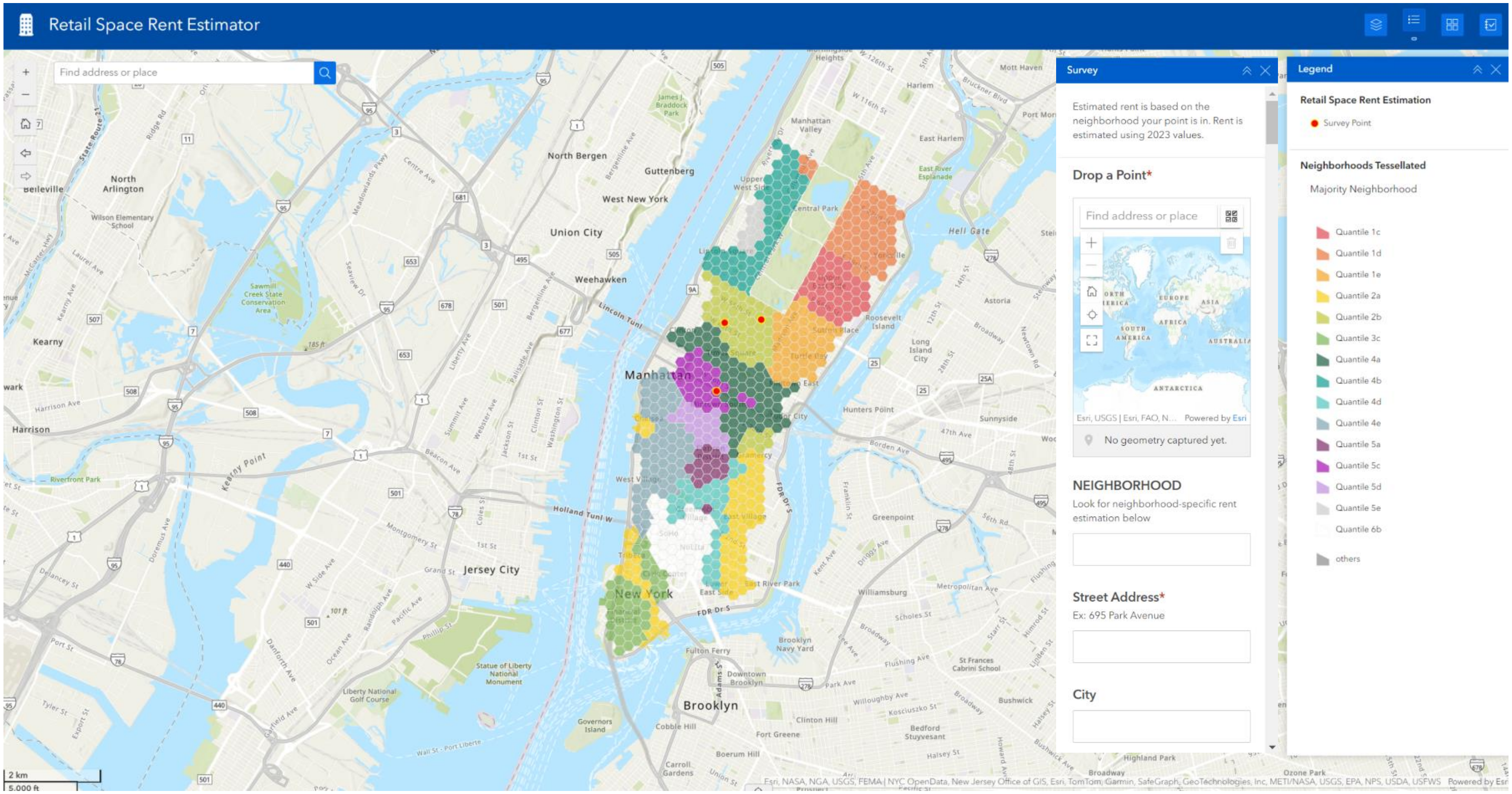


Figure 33 Retail Space Rent Estimator WebApp using Esri Experience Builder and Survey 123.

7. Conclusion and Discussion

7.1 Future Research

Future improvements include obtaining more relevant predictor variables. Although the company I work for has invested in an expanding database, key variables such as square feet of store frontage, if the space has an upgraded HVAC system, and if the space has available outdoor space is not readily accessible in a consolidated format at the time of writing this paper. Further, the size of the space is not considered, with the assumption that price per square feet of nearby comparable spaces accounts for this variable. The inclusion of these variables may alleviate some of the misspecifications and offer better GWR and neighborhood-level models across the board.

Another improvement is to create models for K-means neighborhood delineations. Using K-means in addition to the dual K-means method is a more robust way to create neighborhoods and perhaps get better results. Results from K-means would be an insightful baseline comparison against the performance of the dual K-means neighborhoods.

A more effective method to demonstrate the validity of this study is by developing a custom widget for integration with Esri's Experience Builder. Users can place a pin on the map to calculate neighborhood-specific predictors, such as population density within one tenth of a mile or average household income within 5 minutes, at the point level. Currently, these observations are only generalized at the census block's centroid. Depending on the best fit model determined by the SDSS, the widget would then apply OLS, spatial error, or spatial lag model to estimate rent. As of the time of this writing, Esri's Survey 123 can only handle basic arithmetic operations such as addition, subtraction, multiplication, and division. All box-cox transformations are precalculated as attributes of the block group layer. An additional way to demonstrate validity would be recording and presenting validation metrics including mean bias, root mean square error, and mean and median absolute error. The SDSS has added this functionality for future analysis.

Finally, while the primary focus of this paper has been on retail rent valuation in Manhattan, the application of the SDSS is far-reaching. It could effectively analyze and predict outcomes in other domains such as urban planning, environmental monitoring, public health, transportation, and economic development. The adaptability of the SDSS makes it an impactful tool, with the potential to enhance decision-making across multiple fields.

7.2 Limitations

A significant limitation of this SDSS is its inability to consistently replicate results. Previous test runs have produced varying degrees of fit for the GWR models. Additionally, the current configuration of the code does not support resuming analysis from the point of interruption in a previous session. This system is designed as an advanced spatial tool that allows analysts to adjust parameters and results as necessary one session at a time.

Incorporating broader social factors into rent estimation could enhance its accuracy. Real estate is an imperfect market with measurable and non-measurable factors that are not considered within the scope of this study. For example, the micro-economics such as the supply and demand of spaces in a neighborhood influence rent price. As space becomes scarce and competition for prime locations grows, landlords have the leverage to increase rents. Further, if consumer confidence is high, as measured by the University of Michigan’s index of consumer sentiment, and interest rates are low, there is likely more cash flow for the retailer and an indicator for the landlord to charge more for their space than they would in less ideal conditions.

Although sufficient for this study, Google's *Places Search* API to pull in retailers is not a complete and perfect method to retrieve all retailers throughout Manhattan. To do this, 18 search points were spread out throughout Manhattan. The *Places Search* API uses those 18 points to pull in all categories of retailers that surround each one. However, a limitation with this API is that it pulls the first 20 Google-defined relevant and most proximate retailers for each of the 44 categories. A total of 15,840 local and brand name retailers were pulled. Ideally, the search points should be spread out in a way that parallels the density of retailers throughout Manhattan. The third-party Chain XY dataset is sufficient to supplement the Google retail dataset.

7.3 Conclusion and Discussion

The diverse retail landscape of Manhattan necessitates a flexible approach in valuing rent. This paper highlights the usefulness of leveraging automated spatial statistics to estimate the rent of retail spaces. The process, from data collection to model generation to visualizing results, is tedious and error prone. Therefore, implementing an automated SDSS is an effective way to create dynamic neighborhood-level rent prediction models. The methodology outlined in this paper determines and quantifies which predictors are significant in estimating rent across Manhattan.

The way that the neighborhoods cluster together is an indication of how regression may or may not work out. Referencing the LISA map in Figure 6 of the log transformation of rent, clusters of high-high spatial autocorrelation in the Upper East Side and Times Square areas are within neighborhoods of high performing models. These include quantile 1c, 1d and 1e on the Upper East Side and 2b in the Midtown and Times Square regions. The top three neighborhoods for natural Jenks, 1b, 1e, and 3c are all in these areas, as well as the top two neighborhoods for equal interval 1c and 1d. Further, the results are of course dependent on the spatial distribution of the observations in the dataset. Using standard PLUTO neighborhood boundaries plus the addition of Little Italy and Union Square (see Figure 1), Table 26 below shows the density of observations (number of spaces per square mile), used for this study. Out of the 30 neighborhoods listed below, the five densest neighborhoods are smaller in size, ranking number 17, 28, 22, 25, and 27. These include SoHo, Nolita, Flatiron, NoHo, and the Meatpacking District. Given this study's focus on real estate, clusters of data points primarily along major retail corridors of Manhattan are expected.

Table 26 Density of spaces per neighborhood.

Neighborhood	Density	Area (Sq. Mi.)	Rank Size
SoHo	1,608.7	0.20	17
Nolita	1,241.3	0.06	28
Flatiron District	829.8	0.13	22
NoHo	720.6	0.07	25
Meatpacking	573.6	0.06	27
Nomad	572.0	0.06	26
Union Square	452.5	0.03	30
Greenwich Village	337.0	0.29	13
Theater District / Times Square	330.6	0.22	15
Midtown	304.1	1.44	4
West Village	282.8	0.45	10
Tribeca	272.8	0.36	11
Upper East Side	219.3	1.75	2
Financial District	206.5	0.49	9
Chelsea	174.1	0.78	6
Murray Hill	145.5	0.27	14
Gramercy	127.4	0.20	16
East Village	118.0	0.70	8
Upper West Side	105.8	1.95	1
Hell's Kitchen	93.4	0.96	5
Kips Bay	85.5	0.32	12
Hudson Square	74.0	0.15	21
Lower East Side	73.1	0.71	7
Civic Center	62.5	0.13	23
Chinatown	56.7	0.19	20
Battery Park City	30.6	0.20	19
Stuyvesant Town	29.8	0.20	18
Little Italy	26.7	0.04	29
Hudson Yards	13.5	0.07	24
East Harlem	2.7	1.50	3

This paper highlights the complex nature of retail rent valuation in Manhattan, emphasizing the need for adaptable, neighborhood-specific models. Future research should focus on incorporating more predictors, improving the spatial resolution of data, improving efficiency in the RStudio code, and creating more of a customized interface within the Esri-based web application. Implementing an SDSS can enhance strategic real estate decision-making.

Appendix A: Retail Points Data and Co-Tenancy

Co-tenancy survey template is seen below in Figure 34, with Table 27 showing the retail categories to choose from.

	A	B
1	**NAME:**	**YOUR NAME HERE**
2	**GOAL: How can retailers be categorized and weighted by importantance to indicate a healthy and diverse retail corridor?**	
3		
4	***INSTRUCTIONS***	
5	*Create **up to eight** broad-level categories by grouping together individual categories located in column A below*	
6	*The higher the weight, the more imporatnt that retail group is to the the health of a retail corridor*	
7	*1 suffix means a national brand*	
8	*0 suffix means a local retailer*	
9	*Each retail category can only go into **one** bucket*	
10	*The **WEIGHT TOTAL** will automatically add up as you decide the weight of each group*	
11	***SUGGESTION**: Cut and Paste each eactegory into the bucket so you keep track of what you've already placed*	
12	*Click the **RETAILERS** tab to see which retailers are in each category*	
13	*Click the **EXAMPLE CATEGORIZATION** tab for a completed example*	

Figure 34 Co-tenancy diversity index score survey.

Table 27 Use types included in co-tenancy survey.

Category	Category	Category
apparel - 0	grocery - 0	restaurant - 1
apparel - 1	grocery - 1	Restaurant - bakery - 0
art - 1	gym - 0	Restaurant - bakery - 1
beauty_and_hair - 0	gym - 1	Restaurant - bar - 1
beauty_and_hair - 1	hardware_store - 0	Restaurant - casual - 0
bicycle_store - 0	hardware_store - 1	Restaurant - casual - 1
book_store - 0	home_goods_store - 0	Restaurant - coffee - 0
book_store - 1	home_goods_store - 1	Restaurant - coffee - 1
convenience_store - 0	hotels - 1	Restaurant - dessert - 1
convenience_store - 1	jewelry_store - 0	Restaurant - fast casual - 1
department_store - 0	jewelry_store - 1	Restaurant - fine - 1
department_store - 1	liquor_store - 0	Restaurant - juice - 1
drugstore - 0	liquor_store - 1	Restaurant - qsr - 0
drugstore - 1	medical - 0	Restaurant - qsr - 1
electronics_store - 0	medical - 1	Restaurant - top rated - 0
electronics_store - 1	music - 0	Restaurant - top rated - 1
experiential - 0	music - 1	schools - 1
experiential - 1	office - 1	shoe_store - 0
finance - 1	other_retailer - 0	shoe_store - 1
florist - 0	other_retailer - 1	shopping_mall - 1
furniture_store - 0	pet_store - 0	special - 1
furniture_store - 1	pet_store - 1	veterinary_care - 0
	restaurant - 0	veterinary_care - 1

References

Anselin, L. (1998). GIS research infrastructure for spatial analysis of real estate markets. *Journal of Housing Research*, 9(1), 113-133. https://doi.org/10.1080/10835547.1998.12091900

Basu, S., & Thibodeau, T. G. (1998). Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics*, 17(1), 68-85. https://doi.org/10.1023/A:1007703229507

Bureau of Labor Statistics. (2024). Consumer Price Index. Retrieved from https://www.bls.gov/cpi/

Cellmer, R. (2013). Use of spatial autocorrelation to build regression models of transaction prices. *Real Estate Management and Valuation*, 21. https://doi.org/10.2478/remav-2013-0038

Cliff, A. D., & Ord, K. (1970). Spatial autocorrelation: A review of existing and new measures with applications. *Economic Geography*, 46, 269. https://doi.org/10.2307/143144

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. Wiley.

Fotheringham, A. S., & Li, Z. (2023). Measuring the unmeasurable: Models of geographical context. *Annals of the American Association of Geographers*, 113(10), 2269–2286. https://doi.org/10.1080/24694452.2023.2227690

Furtado, B. A. (2011). Neighbourhoods in urban economics: Incorporating cognitively perceived urban space in economic models. *Urban Studies*, *48*(13), 2827–2847. https://doi.org/10.1177/0042098010391288

Getis, A. (2009). Spatial weights matrices. *Geographical Analysis, 41*(4), 404-410. https://doi.org/10.1111/j.1538-4632.2009.00768.x

Gillen, K., Thibodeau, T.G., & Wachter, S. (2001). Anisotropic autocorrelation in house prices. *Journal of Real Estate Finance and Economics*, 23(1), 5-30.

Golab, C. (1982). The geography of the neighbourhood, In R. Bayor (Ed.), *Neighbourhoods in Urban America* (pp. 70–85). Washington, DC: Kennikat Press.

Google. (n.d.). *Place Search*. Retrieved January 15, 2024, from https://developers.google.com/maps/documentation/places/web-service/search

Ismail, S. (2006). Spatial autocorrelation and real estate studies: A literature review. Universiti Teknologi Malaysia International Repository.

Kiely, T. J., & Bastian, N. D. (2019, November 6). The spatially-conscious machine learning model. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13, 31-39.

Koschinsky, J., Lozano-Gracia, N., & Piras, G. (2012). The welfare benefit of a home's location: An empirical comparison of spatial and non-spatial model estimates. *Journal of Geographical Systems*, 14, 1 – 38. https://doi.org/10.1007/s10109-011-0148-6

Migden-Miller, A. (2024, April 15). *Google Pts V2.ipynb* [Computer software]. Git Hub. https://github.com/andiemigden/SDSS-Phase-One-Retail-Rent-Estimation/blob/main/Google%20Points%20Cleanup/Google%20Pts%20Cleanup%20V2.ipynb

Miller, H. J., & Goodchild, M. F. (2014). *Data driven geography*. The Ohio State University and University of California, Santa Barbara.

New York City Department of City Planning. (2023). *PLUTO and MapPLUTO (Version 23.3)*. New York City Department of City Planning. Retrieved April 8, 2024, from https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page

Oshan, T. M., Li, Z., Kang, W., Wolf, L. J., & Fotheringham, A. S. (2019). MGWR: A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS International Journal of Geo-Information*, 8(6), 269. https://doi.org/10.3390/ijgi8060269

Pace, R. K., LeSage, J., & Zhu, S. (2009). Impact of Cliff and Ord on the housing and real estate literature. *Geographical Analysis*, 41(4), 418-424. https://doi.org/10.1111/j.1538-4632.2009.00770.x

Pace, R. K., Barry, R., & Sirmans, C. F. (1998). Spatial statistics and real estate. *Journal of Real Estate Finance and Economics*, 17(1), 5-13.

Pobuda, M. (n.d.). Using the R-ArcGIS Bridge: The arcgisbinding Package. Esri. Retrieved April 8, 2024, from http://r.esri.com/assets/arcgisbinding-vignette.html

Sun, S. (2022, March). Chapter 4: Spatial regression in R. In Data analysis and visualization with R: Spatial. Department of Geography and Environmental Science, Hunter College. http://www.geo.hunter.cuny.edu/~ssun/R-Spatial/

Tomal, M., & Helbich, M. (2022). The private rental housing market before and during the COVID-19 pandemic: A submarket analysis in Cracow, Poland. *Environment and Planning B: Urban Analytics and CityScience*, 49(6), 1646-1662. https://doi.org/10.1177/23998083211062907

Valente, J., Wu, S., Gelfand, A., & Sirmans, C. F. (2005). Apartment rent prediction using spatial modeling. *Journal of Real Estate Research*, 27(1), 105-136. https://doi.org/10.1080/10835547.2005.12091148

Zhang, R., Du, Q., Geng, J., Liu, B., & Huang, Y. (2015). An improved spatial error model for the mass appraisal of commercial real estate based on spatial analysis: Shenzhen as a case study. *Habitat International*, 46, 196-205. https://doi.org/10.1016/j.habitatint.2014.12.001