Ramsey County Neighborhood Home Values: A Spatial Analysis

Marshall Roll, Eric Wang

November 30, 2023

Abstract

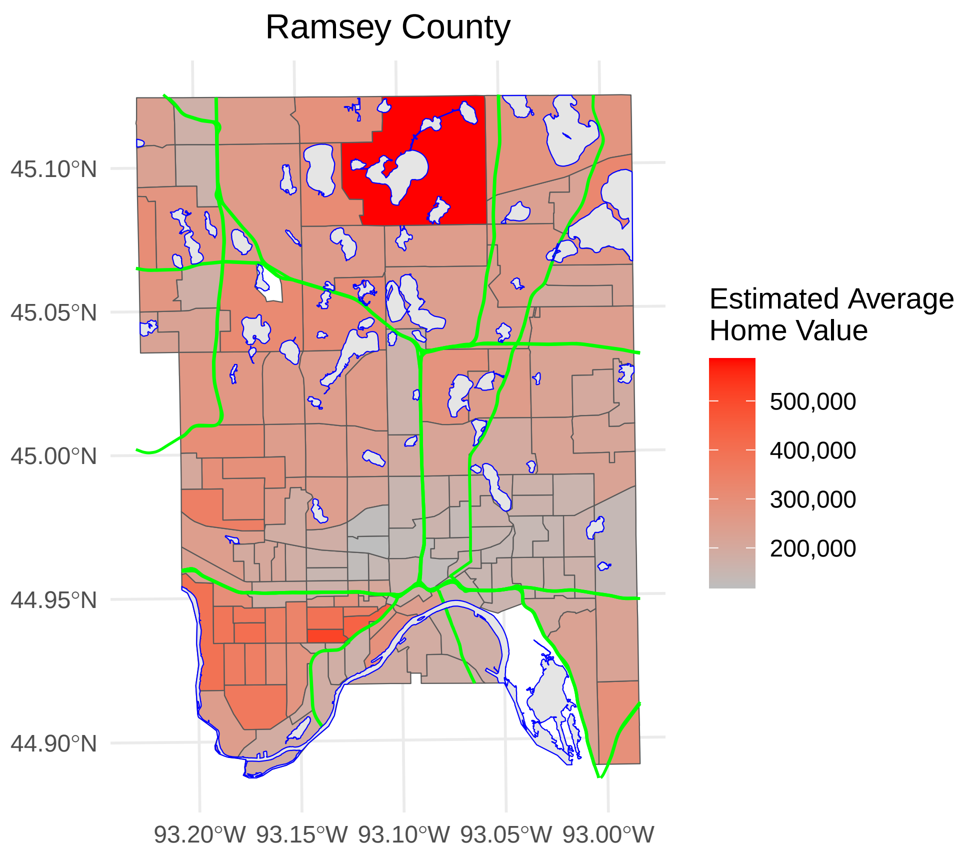
This spatial analysis investigates home values across Ramsey County's neighborhoods, contextualized by historical inequities in housing. Utilizing demographic and housing data from the tidy-census package and spatial econometric models, we examined the impact of various factors on home prices. The study identified that distance-based neighborhood structures significantly influenced home values. Results indicated persistent valuation disparities linked to past discriminatory practices, with certain areas, like Cathedral Hill, deviating from broader trends. Despite these insights, the study acknowledges the limitations inherent in using estimated census data and the potential exclusion of key non-spatial variables. The research underscores the need for continuous policy evaluation to address the long-term effects of historical housing discrimination.

Contents

1 Introduction

The Twin Cities are on land that was originally occupied by the Dakota people, who were forcibly removed as settlements related to fur trapping in St. Anthony and Minneapolis expanded into modern day Ramsey County. St. Paul began to expand in the early 20th century, particularly in the nascent Summit Hill neighborhood, which developed over time into large, very desirable Victorian homes. However, equal opportunity for housing was not available for everyone. In the 1950s, the Rondo neighborhood, primarily composed of Black residents, was razed to make way for the I-94. Further, redlining, the discriminatory practice of designating neighborhoods with a grade A, B, C, or D based on the perceived desirability of each neighborhood. This was largely correlated to each neighborhood's racial and ethnic composition and caused significant and ongoing disinvestment in communities with lower grades. Disparities between such communities continue to exist today, almost 75 years after the practice was banned. (Crowe 2022)

The figure below highlights average neighborhood value by neighborhood in Ramsey County. Census tracts tend to be smaller in size in the southernmost portion of the map, which is closer to downtown St. Paul and larger in the north, which is more suburban. Concentrations of high average home values can be seen in the Summit Hill and Mac-Groveland neighborhoods in southwest Ramsey County. Furthermore, the northern suburbs tend to have higher estimated average home values–the North Oaks neighborhood has the highest, with an average home price of $586,800. By contrast, homes directly north of the I-94, the southern-most east-west highway (the lower-most horizontal highway), tend to have lower average home value than those in other surrounding areas.



This report seeks to analyze the factors underlying neighborhood home value in Ramsey County today and to identify and understand which neighborhoods do not follow broad patterns as have been previously identified. (Crowe 2022) This will help to shed light on the factors that are correlated with neighborhood home value, allowing us to see where historical inequities persist within the Twin Cities. Furthermore, by identifying spatial patterns between neighborhoods, we can determine where average home price may go against established county-wide trends and can lead to a better contextual understanding of these areas.

2 Methods

2.1 Data

Using the tidycensus package in R, we were able to gain demographic information at the census tract level, including average age, income, and household size as well as the proportion of inhabitants renting their property, with a given racial identity, profession by industry, and birthplace region (ie, Minnesota, US, non-US). (Walker and Herman 2023) Using data collected from 2019, many of these variables, including average home price, are estimates calculated by the Census Bureau.

2.2 Variable

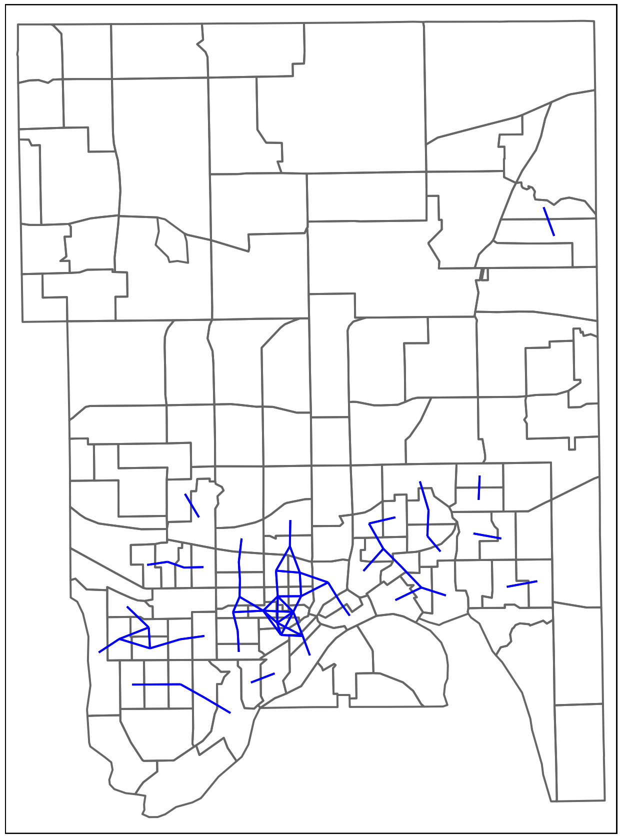
In our spatial data analysis, we used a heatmap to examine correlations between variables and house values and a random forest model to assess variable importance. By integrating both methods, we identified the most impactful variables on house prices, which will serve as principal factors in our subsequent modeling. This dual approach ensures a comprehensive selection of predictors, balancing both correlation and importance.

- Estimated Average Income
- Estimated Average Household Size
- Estimated Age
- Proportion of Residents Born in Another State
- Proportion of Residents Born Outside of USA
- Proportion of White Only Residents
- Proportion of Asian Only Residents
- Area of Census Tract
- Proportion of Residents in Professional Industry
- Proportion of Residents Born in Minnesota

2.3 Neighborhood Structure

In our statistical analysis of spatial data correlation, we investigated four types of neighborhood structures: Queen, Rook, distance-based nearest, and K-nearest neighbors (KNN). Beyond the conventional Queen and Rook contiguity models, we also tried to explore distance-based structures. This decision stemmed from an observed pattern in our dataset concerning house values and their geographical distribution. Specifically, we noticed that regions towards the north exhibited a sparser arrangement than the south ones. In other words, within an equal radius, southern areas tended to have a higher density of neighbors than those in the north. Given the assumption that community sizes are relatively uniform, it's plausible that adjacent northern regions might not exhibit a strong spatial correlation in house prices due to their expansive reach.

To test our hypothesis, we integrated two distance-based neighborhood structures. For the Distance Nearest Neighbors approach, we chose a relatively small threshold of 3500 meters. Below this distance, most areas lacked neighbors, with only a few densely packed regions showing neighboring effects. This boundary was set as part of an extreme conjecture to assess spatial independence at a finer scale. Conversely, for the KNN method, we adopted a more conservative approach by selecting four as the number of neighbors for a given area. This number was deemed sufficient to encapsulate local spatial interdependencies without overextending the influence of distant regions, thus providing a balanced perspective on spatial correlation within our dataset.



2.4 Model

In our spatial model selection section, we started by fitting an Ordinary Least Squares (OLS) model as our baseline. We expanded our analysis to include combinations of Spatial Autoregressive (SAR) and Conditional Autoregressive (CAR) models with four neighborhood structures. We compared their Bayesian Information Criterion (BIC) values to facilitate efficient comparison across many models. The lower the BIC, the better the model. The BIC for the OLS model was recorded at 3328.939, which we used as a benchmark. The following table presents the BIC values for all other model combinations, with lower values than the OLS model highlighted in bold.

3 Results

3.1 BIC

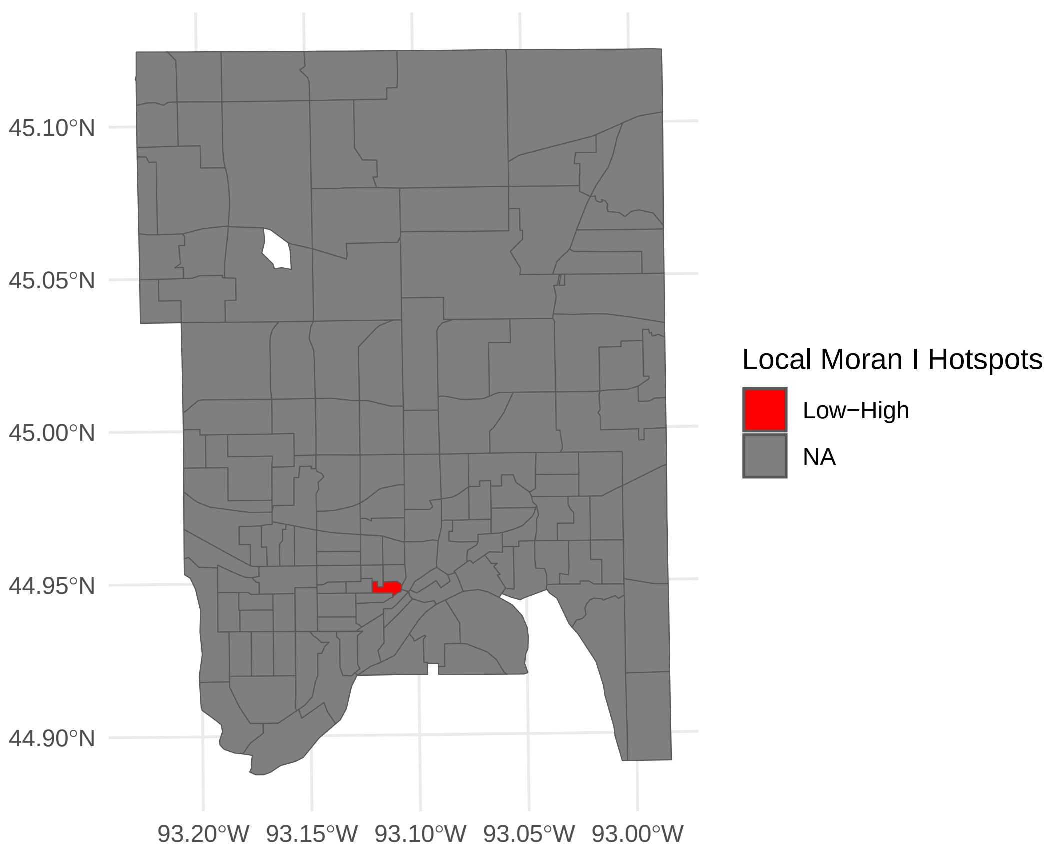
Our results indicated that, overall, the SAR models were marginally outperformed by the CAR models. Furthermore, models with distance-based neighborhood structures consistently showed significantly lower BIC values than those using the Queen and Rook structures, corroborating our initial hypothesis. As the table shows, the Distance Nearest + SAR combination had the most favorable performance, surpassing the KNN + SAR, and was significantly better than any SAR/CAR models combined with Queen or Rook structures. Based on these findings, we concluded that the SAR model, paired with the Distance Nearest neighborhood structure, should be the configuration of choice for our spatial data analysis.

	SAR	CAR
Queen	3329.533	3331.831
Rook	3329.928	3331.845
Distance Nearest ($d = 3500m$)	**3306.04**	**3317.024**
KNN ($k = 4$)	**3322.16**	3331.924

3.2 Moran's

Moran's I is a measure of spatial autocorrelation characterized by a correlation in a signal among nearby locations in space. It ranges from -1 (indicating perfect dispersion) to +1 (perfect correlation), with 0 suggesting a random spatial pattern.

Local Moran's I, often visualized through a Hotspot map, identifies clusters of similar values and is used to detect spatial outliers, revealing areas of significant local spatial autocorrelation. Based on the Local Moran's I Hotspots graph, all the areas except the Cathedral Hill area of Ramsey County are predominantly shaded in dark grey, indicating p-values greater than 0.005, which suggests that the spatial clustering of similar values in these areas is not statistically significant. This implies that for our model residuals, aside from Cathedral Hill, there are no other strong spatial autocorrelation patterns of house values within the county at this significance level. Notably, the Cathedral Hill area stands out with a Low-High trend, indicating that it has a lower attribute value but is surrounded by higher-value areas. This could suggest an enclave of lower values in a generally higher-valued context, which may warrant further investigation into the socio-economic or physical factors contributing to this disparity.



Global Moran's I, in contrast, gives a single summary measure of spatial autocorrelation, assessing whether the pattern expressed is clustered, dispersed, or random across the entire study area. Based on the provided results from the Global Moran's I test, the spatial autocorrelation of the residuals from our spatial model is not statistically significant. The Moran I statistic is 0.03064679, which is close to the expectation of -0.01639344 under the null hypothesis of random distribution. This suggests that the pattern of the residuals is more random than clustered. The p-value of 0.6884 is much higher than the common significance levels (e.g., 0.05 or 0.01), indicating a high probability of observing such a Moran's I value under the null hypothesis of random distribution. Therefore, we fail to reject the null hypothesis, implying no evidence of spatial autocorrelation in the residuals. In other words, the result from the Global Moran's I test suggests that the residuals of our spatial model are randomly distributed in space, with no significant clustering or dispersion patterns. This indicates that our spatial model is well-specified and that spatially structured variables might not be omitted.

```
##
##  Moran I test under randomisation
##
## data:  ramsey_data$sar_resid
## weights: W  n reduced by no-neighbour observations
##
##
## Moran I statistic standard deviate = 0.40101, p-value = 0.6884
## alternative hypothesis: two.sided
## sample estimates:
## Moran I statistic       Expectation          Variance
##        0.03064680       -0.01639344        0.01376053
```

4 Conclusions

Our spatial analysis of neighborhood home values in Ramsey County reveals that historical inequities in housing persist, influenced by the legacy of practices like redlining and uneven urban development. The efficacy of distance-based models over traditional contiguity suggests the nuanced interplay of socio-spatial factors in real estate valuation. The standout case of Cathedral Hill, defying broader trends, calls for a deeper, localized investigation to understand its unique dynamics.

However, the study's reliance on census data and predefined distance thresholds for spatial correlation introduces potential limitations. The absence of significant spatial autocorrelation in model residuals indicates a well-specified model but also points to the potential omission of influential non-spatial factors. Our future work should incorporate more current and detailed data and explore additional variables to unravel the complexities of real estate valuation further and inform policies to rectify longstanding disparities.

Acknowledgements

Thank you to Professor Brianna Heggeseth for facilitating this project and teaching us these concepts.

References

Crowe, Timothy. 2022. "Effects of Redlining in the Twin Cities."

Walker, Kyle, and Matt Herman. 2023. *Tidycensus: Load US Census Boundary and Attribute Data as 'Tidyverse' and 'Sf'-Ready Data Frames*. https://walker-data.com/tidycensus/.