Mini Project 3: Propery Values in Ramsey County: A Spatial Analysis

Jay Anderson, Freddy Barragan, and Ben Christensen

28 April, 2022

1 Introduction

Ramsey County, Minnesota, was formed and built upon a foundation of racist, settler-colonial violence; settlers first stole this county's land from Ojibway and Sioux peoples in 1837 and 1851 (Fairchild 1905), then appropriated it into sites of violence upon minoritized peoples over the decades. Structural racism— redlining, racial covenants, and racial violence— has profoundly shaped the demographic composition of modern neighborhoods throughout Ramsey County and fomented present racial disparities.

In the early to mid-twentieth century, the percentage of White-only districts increased and consequently forced African-American communities into 'hazardous' neighborhoods that have had lasting effects on their health, safety, and liberty (Kaplan, n.d.). Although many of these neighborhoods exist today— Northend, Frogtown, and Summit-University— and their demographic compositions have changed dramatically, the lasting effects of structural and environmental racism have been disastrous on racialized communities (Kaplan, n.d.). In light of Ramsey County's past and the current effects of gentrification on minoritized communities, it is critical that we— as statisticians and people— characterize and advocate against the lasting impacts of these discriminatory practices. As such, we aim to characterize the economic geography of housing throughout Ramsey County. We used economic, geographic, and demographic data drawn from the 2015-2019 American Community Survey to model average house prices across census tracts using spatial statistical regression to adjust for the implicit correlation between nearby census tracts.

2 Data

Using the `Tidycensus` R package (Walker and Herman 2021), we pulled demographic, geographic, and economic data across Ramsey County at the census tract level from the 2019 American Community Survey (ACS; N=135). The ACS is a longform survey conducted by the U.S. Census Bureau that aims to continually track the economic, demographic, and social development of neighborhoods across the United States in a 12-month period. As such, this data is an aggregation of responses in each census tract which is necessarily different from other aggregation levels (e.g. census block, census group, SIP code tabulation areas, etc.); across aggregation levels, census tracts are typically preferred by geographers because sample sizes are relatively homogeneous between tracts and provides one of the most granular levels of data. In addition to census data, `Tidycensus` also provides associated shapefiles for each reported census tract.

As our project aims to model mean household property values per tract by broad structural factors we used multiple variables encoding the economic, demographic, and mobility-related characteristics of each neighborhood. We specifically used:

- **Percent Non-White:** In order to account for uniformly small sizes of racial/ethnic groups throughout Ramsey County, we used this broader variable to measure the proportion of all racial/ethnic minorities in a census tract. This is equivalent to $1 - \%\text{White}$.
  - We performed a log transformation on this variable to account for an apparent nonlinear relationship to expected household property values.
- **Average Age:** The average age of residents living in a defined census tract.
- **Median Income:** The median income value of residents living in a defined census tract.
- **Gini Index:** Income inequality measured by the Gini Index per census tract
- **Highway Presence:** Indicator variable describing the presence of at least one highway within a census tract.
- **Average Household Size:** The average household size of all homes in a defined census tract.

Following exploratory analysis, we added an indicator variable for census tract 27123040601; this census tract encompasses North Oaks, Minnesota— a private community whose economic composition is vastly different from surrounding areas. It is also demographically distinct from the tracts in St Paul proper. Not only that, but being a private community ups the prices of all the houses in the community in such a way that our model wouldn't be able to predict without singling out the tract.

3 Methods

Ordinary least squares (OLS) regression is a standard statistical approach that models variable outcomes using a set of explanatory predictors. While OLS is sufficient in various situations, this technique imposes strict assumptions on the independence of data— namely, that data cannot be correlated. This independence assumptions is, however, inappropriate when analyzing spatial data because of the latent effects that space and geography can have on observations. In other words, if we were to model average property values in any given census tract using an OLS model, we would fail to account for the influence that neighboring census tracts would have on the outcome.

Neighborhood network structures (NNS) are formal mathematical objects that describe the connections between points or objects; they are especially critical in spatial data analysis so that we can encode how observations are correlated between each other in space. Considering a map, if two census tracts share a physical border (i.e. a whole line segment) or touch each other at a single point (i.e. a corner), a Queen NNs would state that these census tracts are 'neighbors' (i.e. they are similar to each other). Other neighborhood structures exist and are generally more conservative in their assessment of neighbor-status. For example, the Rook NNS would only consider census tracts that share a border 'neighbors', while a KNN structure would assign tracts neighbor-status according to their relative distances from one and other. We could further adjust any of these described NNS to adjust for the presence of substantial geographic/social barriers to produce penalized NNS that minimize the occurrence of tracts being naively-classified as neighbors, simply because they share a border.

In this analysis, we used a penalized Queen NNS to model the connections between census tracts, penalizing the connections between tracts that have a highway running between them. While we tested other NNSs— e.g. Rooks, KNNs, and their penalized parallels— this penalized Queen NNS was most appropriate given its generality and the historical role that highways played throughout the Twin Cities: separate and segregate communities. As such, we do not expect the characteristics of communities in census tracts bordered by highways— especially the I-94 Highway— to be very similar.

The simultaneous autoregressive (SAR) model is a spatial regression technique that extends standard regression frameworks to account for underlying spatial relationships between observations in a dataset by using an underlying NNS to impose spatial weights on observations (Whittle 1954; Hooten, Ver Hoef, and Hanks 2014; Wall 2004; Heggeseth 2022). The conditional autoregressive (CAR) model is another spatial regression

technique that also adjusts for underlying spatial correlation between observations, but in a manner that is distinct from SAR models (Besag 1974; Wall 2004; Heggeseth 2022). Specifically, CAR models assume a Markov property or 'memorylessness' to the relationship between neighbors and thus impose weaker spatial correlation structures than SAR models (Wall 2004). The relative benefits of CAR and SAR models are highly contextual and specific to the data and problem at hand.

In this analysis, we used a SAR model to predict average property values within census tracts. We define our proposed model below:

$$\mathbf{Y} = \lambda \boldsymbol{WY} + \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

where

- Y is the vector of mean house prices for all 135 census tracts.
- $\lambda \boldsymbol{WY}$ is a matrix product of some constant λ, our penalized proximity matrix W, and a matrix of housing values for all neighborhoods Y.
- $\boldsymbol{X\beta}$ is the product of the 135×8 covariate matrix— where entries of this matrix are the observed values of the 8 predictors (columns) for the 135 different census tracts (rows) in our data— and a 1×8 coefficient vector.
- $\boldsymbol{\epsilon}$ is the matrix of our tract-specific error terms with $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$

Outside of this analysis, we performed model selection between our model and OLS and CAR analogs of our model, using standard evaluation metrics:

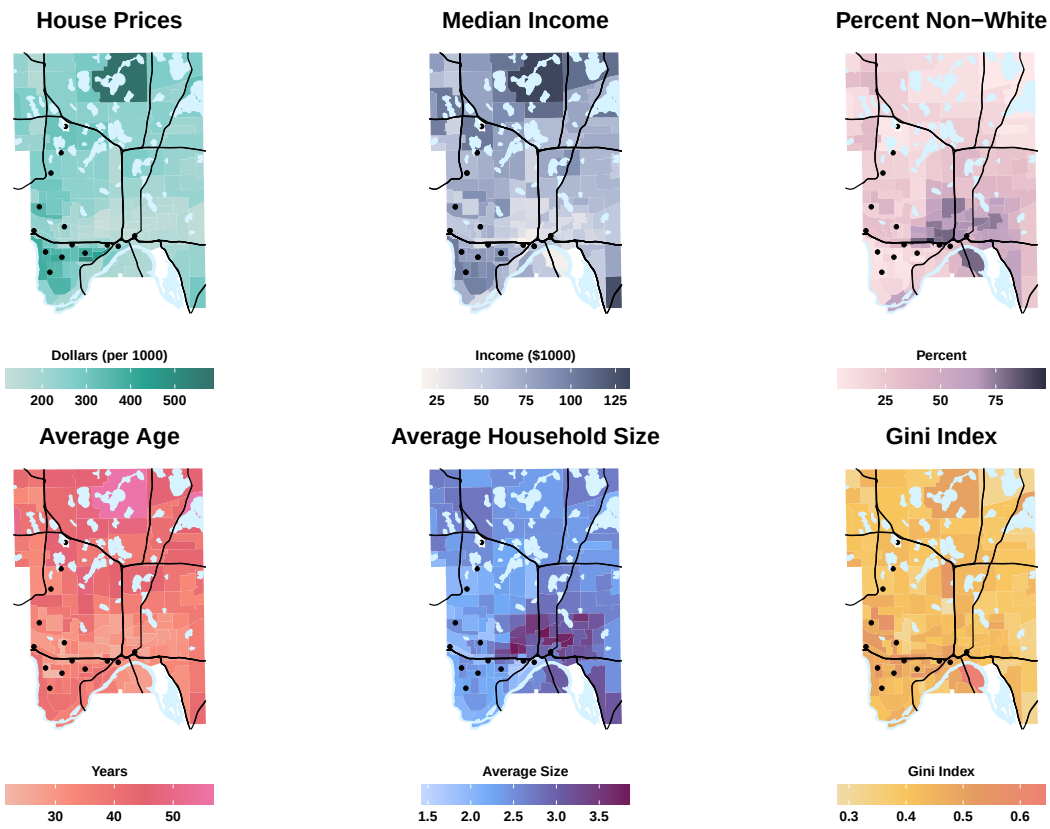
- Residual mapping
- Bayesian Information Criterion (BIC) metrics
- Moran's I Test Statistics of Residuals

Uniformly, our SAR model performed better with respect to these evaluation metrics, however we exclude substantial discussion of the relative performance of models in this analysis. Additionally, because historical information on Ramsey County (Kaplan, n.d.) substantiates a very strong spatial dimension to neighborhood housing prices, SAR models' stronger assumptions about the spatial correlation between census tracts are more appropriate.

4 Results

```
library(ggpubr)

ggarrange(price, income, nonwhite, age, size, gini,  ncol=3, nrow=2)
```



The geography of housing prices indicates the presence of a spatial dimension at play in Ramsey County. From the map, it appears that housing is consistently cheaper in the Midway district and along the I-94 Highway, while housing is generally much more expensive in census tracts below the I-94 Highway in South Western Ramsey County. Average property values were consistently cheaper in the Midway district and along the I-94 Highway, while the proportion of non-White residents was also high in these same areas; together, this indicates a strong negative relationship between average income and the proportion of non-White residents in a neighborhood. Moreover, these same areas showed higher average household size, suggesting a negative relationship between average household size and income. More ostensible are the dramatically high housing prices and average resident age in and surrounding North Oaks, Minnesota in Northern Ramsey County. More subtly, the Gini Index of the tracts appears higher in North Oaks and in the census tracts directly surrounding the I-94 Highway on the east side.

Table 1: OLS Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	11076.916437	4.999965e+04	0.2215399	0.8250279
perc_nonwhite	-18057.453434	7.199995e+03	-2.5079813	0.0134029
age_e	-1993.050839	7.264017e+02	-2.7437311	0.0069541
income_e	2.439717	2.645616e-01	9.2217345	0.0000000
gini	473485.730722	5.978755e+04	7.9194699	0.0000000
north_oaksTRUE	170359.562570	4.044168e+04	4.2124752	0.0000475
AnyHwysTRUE	-5018.454753	6.733859e+03	-0.7452569	0.4574935
household_size_e	-36646.823541	8.425455e+03	-4.3495364	0.0000277

Table 2: Moran's I of OLS Residuals

estimate1	estimate2	estimate3	statistic	p.value	method	alternative
0.2135818	-0.0074627	0.003452	3.762228	0.0001684	Moran I test under randomisation	two.sided

We performed a Moran's I test on our OLS model residuals to identify the presence of spatial clustering and validate the use of a spatial regression model. Using the above table, we found statistically significant evidence (p=0.0001684) suggesting the occurrence of meaningful spatial correlation in our data that cannot be accounted for by the OLS model. As such, we reject the null hypothesis (i.e. data are independent in space) and use our proposed SAR model to adjust for spatial correlation.

Table 3: SAR Model Summary (Ordered by Significance)

Term	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	53842.719824	4.634395e+04	1.1618070	0.2453139
income_e	2.176899	2.542711e-01	8.5613282	0.0000000
gini	394165.273753	5.554262e+04	7.0966271	0.0000000
north_oaksTRUE	209286.916931	3.386717e+04	6.1796394	0.0000000
household_size_e	-37130.388820	9.154930e+03	-4.0557809	0.0000500
age_e	-1593.596916	7.323211e+02	-2.1760905	0.0295485
perc_nonwhite	-13621.602705	7.715939e+03	-1.7653850	0.0774991
AnyHwysTRUE	-3781.527375	6.638834e+03	-0.5696072	0.5689442

Using our SAR model with $\alpha = 0.05$, we found sufficient evidence to reject the null hypothesis and conclude that median income, income inequality (i.e. gini), and average household size all had significant effects on average property values, when holding relevant covariates constant. Further, we found a statistically significant difference between average property values in North Oaks from all other census tracts.

While the effects of these variables are varied, we found that 1000 dollar increases in median income and 1% increases in income inequality (gini) were generally associated with increases in average property values. Similarly, we found that average house prices in North Oaks were 209286.97 dollars higher than the baseline census tract. Conversely, we found that 1 unit increases in average age and household size were associated with decreases in approximately 1600 and 1400 dollar decreases average property values, respectively.

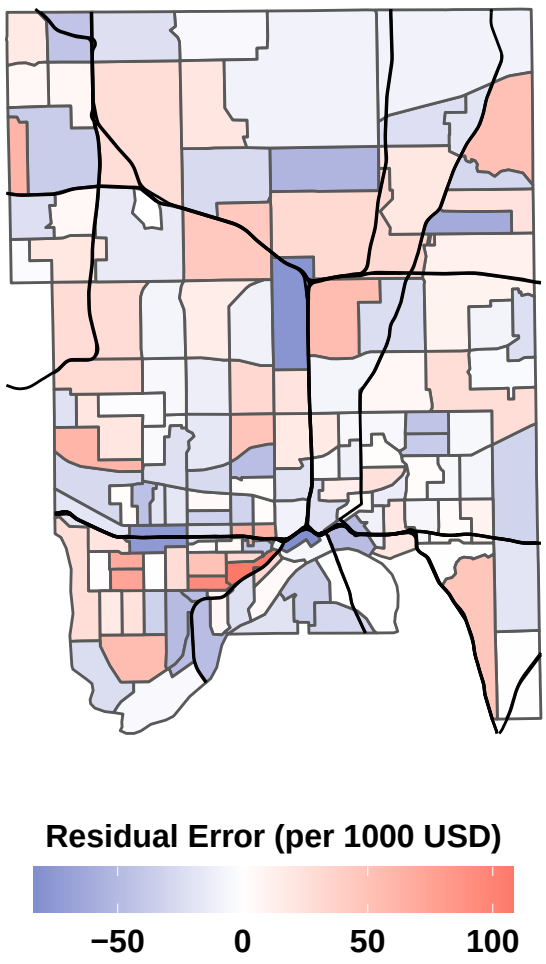


Table 4: Moran's I of SAR Residuals

estimate1	estimate2	estimate3	statistic	p.value	method	alternative
-0.0062265	-0.0074627	0.0034633	0.0210053	0.9832415	Moran I test under randomisation	two.sided

Generally, our residuals of average house prices are randomly distributed across census tracts with no explicit cluster patterns. However, careful inspection of the residuals suggests slight structural differences in the quality of our predictions in the periphery of the I-94. With our currently SAR model, we tend to over-predict housing prices (blue) in census tracts north of the I-94 and under predict housing prices (red) houses south of the I-94.

Importantly, one tract— near the intersection of the I-94 and I-35E— in Southwest Ramsey County was under predicted by over $100,000; this tract is an oddity. Although it has the third highest average property

value, this census tract has a dramatically lower median income than other census tracts with similar average property values.

Using Moran's I test, we can assess whether the systematic errors observed in the map demonstrate sufficient statistical concern. Using the Moran's I test on our SAR residuals, we found insufficient evidence of spatial clustering in our residuals which confirms that our SAR model has improved the OLS framework and adequately adjusted for the latent spatial effects on average property values.

5 Conclusions

Using a simultaneous autoregressive model, we characterized the variability of average property values across Ramsey County, MN and its relationships to the economic, demographic, and geographic features of Ramsey's communities. Generally, we found that the economic features of a given community were meaningfully related to average property values— this feels intuitive. However, the role of wealth in predicting house value is troubling, given the deep wealth inequities that exist throughout Minnesota and the role racial capitalism has played in the accumulation of wealth, land, and resources for White Americans.

The high concentration of wealth in planned and/or mostly-white communities throughout Minnesota will be pivotal in widening the wealth gap. Future— mostly, white— generations will inherit homes in these 'desirable' neighborhoods as non-White communities are thrust into precarity by the fists of gentrification, Capitalist labor conditions, and a dwindling social security net, thus creating generational feedback loops that allow wealthy communities to inherit land and perpetuate observed racial disparities.

While one might expect property values to increase according to average household size in order to accommodate larger families, our data suggests an inverse effect where average property values decrease with average household size. Together, this relationship indicates that larger families may be relegated to live in areas with lower property values— many of which are situated were once situated in 'undesirable' and 'hazardous' areas— due to the economic strains of raising a family. This result is especially troubling when we consider how average household size is associated with increased proportions of non-White citizens, meaning that non-White communities tend to live in potentially worse conditions. Income inequality had a positive relationship to property values, even after adjusting for median income. There were pronounced increased income inequality scores that demonstrate the presence of large numerical gaps in income within wealthy communities where where significantly richer individuals that inflate the Gini index in wealthy tracts also bring on higher average home values.

Although not statistically significant once accounting for economic factors, it seems that race (e.g. non-White percent) may play a factor in the geography of property values, given the geography of residual errors. Because our model errors tend to be concentrated in census tracts on the periphery of the I-94 its plausible that our weights weren't sufficiently stringent to account for latent historical factors (e.g. redlining). If that is the case, then this suggests that historical redlining may still have lasting impacts on housing values even after adjustment for current economic, structural, and demographic variables. While incorporating historic redlining classifications provides a tangible step in improving future models, it speaks to the lasting negative effects of redlining and the construction of I-94.

A critical limitation to our work is the roles that gentrification and immigration have played on morphing the economic and demographic landscape of Ramsey county. Because house prices in historically marginalized neighborhoods have been forcibly raised or indirectly raised by wealthy White people— namely, landlords, housing developers, and corporations— the exodus of low-income and non-Whtie people from Ramsey county may have decreased the sensitivity to the true effects of race and racism on property values. Our analysis is also limited by our omission of the temporality in housing prices and demography. As such, we must emphasize that our findings are only capturing a specific moment in time and not the precursors or histories (i.e. historic redlining) that have reshaped housing prices. Further, because our data is an aggregate summary

of multiple homes in each census tract, our aggregate measures are implicitly tied to the variability in sampling and the sampling errors in the 2019 American Community Survey of Ramsey county.

6 Acknowledgements

We'd like to thank Brianna Heggeseth for helping us along throughout this project. Specifically, we thank her sincere effort in cleaning and preparing this `tidycensus` data and for being so accommodating throughout the writing process.

7 References

Besag, Julian. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society: Series B (Methodological)* 36 (2): 192–225.

Fairchild, Henry Shields. 1905. *Sketches of the Early History of Real Estate in St. Paul.* Minnesota historical Society.

Heggeseth, Brianna. 2022. "7.4 Areal Data." In *Correlated Data Notes.*

Hooten, Mevin B, Jay M Ver Hoef, and Ephraim M Hanks. 2014. "Simultaneous Autoregressive (SAR) Model." *Wiley StatsRef: Statistics Reference Online*, 1–10.

Kaplan, Margaret. n.d. "Redlining in St. Paul - Interfaithaction.org." The Housing Justice Center. https://interfaithaction.org/wp-content/uploads/2019/10/Redlining.pdf.

Walker, Kyle, and Matt Herman. 2021. *Tidycensus: Load US Census Boundary and Attribute Data as 'Tidyverse' and 'Sf'-Ready Data Frames.* https://CRAN.R-project.org/package=tidycensus.

Wall, Melanie M. 2004. "A Close Look at the Spatial Structure Implied by the CAR and SAR Models." *Journal of Statistical Planning and Inference* 121 (2): 311–24.

Whittle, Peter. 1954. "On Stationary Processes in the Plane." *Biometrika*, 434–49.