SOCI 40217 Paper Review
Regressivity in Property Tax Practices

Kathryn (Katy) Koenig

1 Summary

In this analysis, the author seeks to understand the causes of the inaccuracies of property assessments in America as it is an issue of equity. To do so, the author had to create a dependent variable using price of a home at time of assessment and price of home at time of sale. Because the assessed value of a home is generally calculated as 10% of the fair market value, issues regarding units are introduced. To ameliorate this problem, the author aggregates up to census tract and then creates the dependent variable r, or regressivity. The dependent variable is defined as the average sale price to assessed value ratio of the top decile of homes to the average sale price to assessed value ratio of the bottom decile of homes for a given census tract.

The author then uses the following demographic information for each census tract to create regressors: percent white residents, percent residents that are high school graduates, population, poverty percentage[1], percentage of of homes valued at less than $100,000. The author conducts OLS using nine models of varies combinations of the demographic variables. Specifically, Model Specification 2 looks only at the impact of the percent white residents and percent of residents with at least a high school education; Model Specification 3 analyzes only the population and the percent of homes under $100,000. He ultimately decides that Model Specification 1 best represents his data, which I believe is a decision based on R^2 values. This specification details three models: Model 1 including all regressors; Model 2 including all regressors except percent of homes valued at less than $100,000; and Model 3 which includes only percent white residents, percent residents with a high school degree and percent in poverty. For all models, using basic OLS, all coefficients are significant at 1% level, so it may not have been necessary to include Model 2 and Model 3[2].

[1] The author does not specify whether this is the percent of households or percent of individual residents in poverty.

[2] I believe the author included the second and third models due to the difference in R^2 and adjusted R^2: the third model is the only model where the two are equal although this justification is not explicitly stated in the analysis

Most notably, in Model 1, we see that an increase in homes values at less than $100,000 and in percent in poverty are associated with an increase in regressivity. Conversely, an increase in the percent of white residents, in percent residents who have graduated high school and in population are associated with a decrease in regressivity, reflecting that as census tracts grow in population generally, become whiter and have more educated residents, there is less regressive assessments of housing values. It is worth noting that all the coefficients are somewhat small, and the adjusted R^2 value for Model 1 is 0.16 (the highest out of all three models).

2 Spatial Model Justification & Context

This research question is inherently spatial as housing prices are clustered by location. Because high value homes are generally near high value homes and vice versa, the errors in our outcome variables could be correlated, and therefore, there is justification for investigating a spatial error model. Additionally, the methods and timelines for property assessments are also somewhat spatial in that different state and local governments conduct these assessments, so there could be errors in the outcome variable due to localized assessment error. There is also an argument for a spatial lag model because housing values can influence one another, i.e. a home that is an area where other homes have decreased in house value will also lose value, all else equal.

The paper fails to note the timeline for this analysis, and because sources for the data are not given, I am unable to infer this aspect of the data. The paper lacks information regarding the alignment of assessment dates and sales dates which could introduce issues into the data as the assessment could not have been valid with respect to the point of sale. Because this analysis uses census tract data, for which the boundaries are drawn at the decennial census, I would assume the demographic data is also from the census (as opposed to the American Community Survey which is conducted more frequently), which could also be out-of-date at the point of sale as well.

Also, there may be a modifiable areal unit problem as the author is aggregated data points up to the census tract spatial unit. We are not given information regarding the range of data points per census tract so it is hard to further investigate this issue, but I am wondering if a multilevel model or a model with different spatial units would better suit the data if, for example, one census tract had three datapoints while another had 300,000 data points. Furthermore, the analysis notes that only 67,352 census tracts out of 72,539 tracts were used but does not provide a reason for not using all the census tracts.

3 Spatial Techniques Employed & Insights Gained

The author utilizes extensive spatial techniques to understand which model best applies to his data. After creating a first-order queen contiguity matrix[3], he conducts Breush-Pagan and Koenker-Bassett tests to find that his OLS estimates exhibit heteroskedasticity. Because both his Lagrange Multiplier Lag Test and his Lagrange Multiplier Error Test were significant at the one percent level, the author conducted robust versions of these test which were also both significant at the one percent level. He than ran a spatial lag regression, a spatial error regression, the combined spatial lag and spatial error regression in GeodaSpace to find that both the spatial error term λ and the spatial weights matrix coefficient were significant at the one percent level.

While this may provide support for a combined spatial error and spatial lag model as the best for modeling the author's data, I believe further investigation should be done to understand if this joint model is the best representation of the data: the author does address this issue in noting that the spatial unit of the census tract may not be best suited for this investigation. Specifically, he notes that the when using the same models but on a country-wide level the results are similar but "better capture spatial factors." Therefore, what is the justification for census-tract level instead of county-level analysis?

I would posit that there could still be misspecification of the model and that perhaps more regressors are needed in this model. For example, because census tracts are generally drawn to have similar populations, they may vary significantly in size, so I would suggest that density may be a better regressor than the population variable used in this analysis. Furthermore, the multicollinearity condition number is somewhat high (41.555) which may also reflect issues with the regressors.

Overall, the analysis can confirm that there is certainly spatial autocorrelation within the dataset (despite no Moran's I measure being included in the report). While the direction of the relationship between each coefficient and the regressand does not change in the spatial lag, the spatial error and the combined models, the magnitude of all coefficients noticeably decrease from the models run in OLS Model Specification 1, aside from the constant term. This is most likely due to the variability being explained by spatial aspects. For example in the combined spatial lag and spatial error model, the weights matrix has the largest coefficient followed by the coefficient λ, revealing that nearby census tracts have a large effect on the regressivity of a given tract.

[3]While using a first-order queen weights matrix is common practice, I do wish some justification for this choice of weights matrix had been provided.

While many interesting insights can be gleaned from the information provided in the initial analysis, this investigation suffers from a lack of explanation. I believe the analysis would have benefited significantly from additional analysis and discussion of results.