SOCI 40217 Initial Results
Assessing Transit Equity via Commute Times

Kathryn (Katy) Koenig

1 Introduction & Background

In the initial analysis below, we explore the impact of transit access and demographic information on commute times for census block groups in Chicago. Long commute times not only effect quality of life but may reveal a lack of job opportunities and disparities in public transit[1], revealing issues of access and equity within the city. Additionally, as public transit ridership decreases and car congestion increases of both personal vehicles and via ridesharing apps[6], despite the negative effects of long commutes[9], travel times may continue to climb.

Spatial econometrics can provide vital insights into this policy issue as there is justification that our data may be more accurately analyzed through a spatial lag and/or spatial error model. For example, a spatial lag model may be justified because traffic patterns in one block group can significantly influence traffic. A spatial error model may better reflect the data as adjacent block groups are more likely to share transit lines. Despite an abundance of statistical research regarding commute times [4, 7], most fail to address spatial issues and the few that do either complete multilevel modeling [3] to address spatial effects or are inclusive regarding model specification [5]. We discuss assumptions of our OLS model to understand which spatial model is applicable in this research in the Spatial Diagnostics section below.

2 Data & Methods

To complete this analysis, we gathered demographic data from the Census Bureau's American Community Survey 5-Year Estimates (ACS) and transportation information from the City of Chicago Data Portal (Chicago Data Portal). The ACS data is computed using a rolling five-year window and is available the year following its collection. Therefore, the ACS data used below reflects demographic information from 2014 to 2018. For this analysis, we use the smallest geographic unit available, census block group, of which there are 3,983 relevant[1] observations in Chicago. The geographic information for geometry of the census block groups was also from the ACS. As all geographical unit boundaries are drawn with the decennial census, the block group boundaries were created in 2010 and are subject to change upon completion of the 2020 census. Information regarding the location of public transit stops is from the Chicago Data Portal. The bus stop dataset was updated in April 2019 and the L stop data was updated in May 2018. All transit lines which run on a "normal" schedule designation were included (i.e. special holiday transit or alternative routes were not included in this analysis). Table 1 provides a detailed description of all variables used in this ordinary least squares regression analysis.

After preprocessing the data to the respective formats described above[2], we then conducted standard OLS using the model below:

$$y = X\beta + \epsilon$$

in which the variables are defined as follows:

- y is an $3,983 \times 1$ vector of observations our dependent variable pct_long_commute

[1] Eight block groups had a population of 0 and were excluded from this analysis.

[2] Notebooks detailing all preprocessing can be found https://github.com/katykoenig/space-transit

- X is our $3,983 \times 12$ matrix where each column represents our independent variables plus our constant term. Our independent variables include all variables in 1 except pct_long_commute plus an interaction term between num_stops and pct_transit
- β is a 12×1 vector of coefficients for our regressors, which we are estimating
- ϵ is an $3,983 \times 1$ vector of error terms

Table 1: Description of Variables

Variable Name	Description	Source
num_stops	Number of Transit Stops within a 1/2 mile radius of center of census tract[3]	Chicago Data Portal
pct_hh_pov	Percentage of Households below the Poverty Line [4]	ACS
pct_working_age	Percentage of Residents between 15 & 64	ACS
pct_employed	Percentage of Employed Residents	ACS
pct_black	Percentage of Residents that Identify as Black or African American	ACS
pct_hisp	Percentage of Residents that Identify as Hispanic or Latinx	ACS
pct_car	Percentage of Residents that Commute via Personal Vehicles (alone or via carpooling)	ACS
pct_public_transit	Percentage of Residents that Commute via Public Transportation	ACS
veh_per_capita	Number of Vehicles per Resident	ACS
density	Number of Residents per Square Mile	ACS
pct_long_commute	Percentage of Residents with Commute Time above 45 Minutes [5]	ACS

After completing standard OLS in which all variables were included (Model A), we then completed a second regression including only the variables which were statistically significant at the five percent level in our first model (Model B). We also ran a third regression model with the same regressors as Model B minus pct_car due to multilcollinearity issues (Model C). Below, we discuss the outcome of these regressions using a frequentist approach as well as analyze the implications of spatial autocorrelation in our variables.

3 Analysis & Discussion

In Table 2, we provide the outcome of the regressions for the two models described above. As we can see above, in Model A, the percentage of working age residents, the number of personal vehicles per resident and the density of the census block group are not statistically significant and are therefore, not included in Model B. We lose very little explanability in the variability of our data from Model A to Model B (Adj. R Squared loss of -0.0002 between models). Both OLS models account for approximately 26% of the variance in the percent of long commuters.

As we can also see in Table 2, our dependent variables in both models exhibit high multicollinearity. We then examined the correlation between our regressors and provide a heatmap of this analysis in Figure 1. As we can see, percent of car commuters and percent of public transit commuters are very negatively correlated. We therefore ran a third regression, Model C, excluding the pct_car variable from this regression and again provide the results in Table 2. While we lose a some explanability in the variability of datta from Model B to Model C, as evident in the smaller adjusted R squared value for Model B, we will focus our discussion on Model C due to collinearity issues.

[3] A half mile radius was used as this is standard in related literature as this distance is roughly 15 minutes of walking [8]

[4] The poverty threshold for a family of four is US$26,200 [2]. As the ACS bins incomes, I designate any income below US$25,000 to be below the poverty line as it is bin cutoff closest to the national poverty threshold. Therefore, any income above US$25,000 is considered above the poverty line in this analysis.

[5] The 75th percentile for commute times in Chicago falls within the 45 to 59 minute bin as given by the ACS, therefore all commute times greater than 44 minutes were labeled as "long." This accounts for approximately 28% of Chicago commuters

Table 2: Regression Results

Regressor	Model A	Model B	Model C
constant term	-0.0694*	-0.0768**	0.3165**
	(0.0330)	(0.0292)	(0.0127)
num_stops	0.0001**	0.0010**	0.0003**
	(0.0)	(0.0001)	(.00009)
pct_hh_pov	-0.0629**	-0.0603**	-0.0881**
	(0.0185)	(0.0167)	(0.0171)
pct_employed	-0.1486**	-0.1595**	-0.1923**
	(0.0302)	(0.0207)	(0.0212)
pct_black	0.0541**	0.0502**	0.0722**
	(0.0080)	(0.0076)	(0.0077)
pct_hisp	0.0191*	0.0165*	0.0356**
	(0.0093)	(0.0080)	(0.0081)
pct_public_transit	0.8282**	0.8286**	0.4340**
	(0.0330)	(0.0327)	(0.0197)
num_stops × pct_transit	-0.0036**	-0.0036**	-0.0027**
	(0.0003)	(0.0003)	(0.0003)
pct_car	0.3881**	0.3890**	
	(0.0266)	(0.0262)	
pct_working_age	-0.0274		
	(0.0290)		
veh_per_capita	0.0110		
	(0.0270)		
density	0.0003		
	(0.0002)		
R Squared	0.2618	0.2611	0.2200
Adj. R Squared	0.2598	0.2596	0.2187
Multicollinearity Condition No.	61.395	47.360	20.182

* significant at the 5 percent level
* significant at the 1 percent level

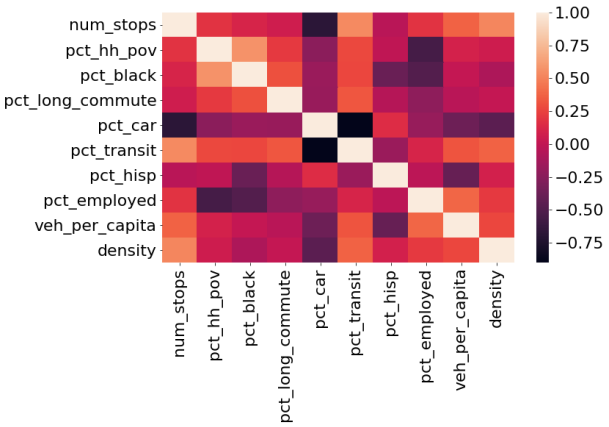

Figure 1: Heatmap of Correlation of Variables

While all of our coefficients are statistically significant at the one percent level, some have low coefficients, reflecting little effect on the percent of commuters in the census block with long commutes. For example, when a census block group gains an additional transit stop (num_stops), we expect to see an increase in the percent of long commuters of 0.03, ceteris paribus. This positive effect may demonstrate that increasing the presence of public transit does not necessarily mean fewer residents with long commutes. Moreover, the coefficient between our interaction term of number of transit stops and percent public transit commuters is negative but quite small, revealing that when public transit access increases in areas in which public transit commuters also increase, we would expect slightly fewer commuters with travel times above 44 minutes.

We also see that for each additional percentage of public transit commuters results in a 0.4340 percent increase in the percentage of long commuters, all else equal. In Model B, we also see that the coefficient for percent of car commuters is also positive, which may reflect that residents that commute via other means of transit, e.g. walking, biking, may have significantly shorter commutes that those that use their car or public transit to work. It is worth noting the coefficient for percent of public transit commuters has the largest magnitude in all three of our regressions, reflecting a stark difference in commute times depending on mode of transport.

Somewhat surprisingly, we see that as the percent of employed residents in a block group increases, the percent of commuters with commute times greater than 44 minutes decreases. While one may expect that more workers means more traffic and therefore, longer commute times on average, in Model B, we see that when a block group's percent of employed residents increases by one percent, we expect to see a decrease in long commuters of roughly 0.19 percent. The decrease in long commute times may reflect changes in commute modes, e.g. buying a vehicle with the income from this employment or the addition of a bus stop designed for commuters.

Our regression also demonstrates issues of equity regarding commute times, race and ethnicity. As the percent of black residents in a block group increases by one percent, we expect to see an increase in long commuters of 0.0722 percent. Similarly, as the percent of Hispanic/Latinx residents increases by one percent, we expect to see an increase in long commuters of approximately 0.0356 percent. Interestingly, as the percentage of residents below the poverty threshold increases by one percent, we expect to see a decreases 0.0881 in long travel time commuters.

Both of our regressions have Jarque-Bera test statistic of almost 0.0, therefore we do not reject the null hypothesis of normality. Therefore, all the properties of maximum likelihood apply to our data. Our dataset is quite large with 3,983 observations, so we are also able to employ the Central Limit Theorem so our data approximates to a normal distribution and the asymptotic properties of generalized method of moments (GMM) will apply.

4 Spatial Diagnostics

To understand spatial autocorrelation issues, we first computed a first order queens spatial weights matrix as it is the most commonly used type of weight matrix. As we can see in the histogram of the neighbors from this weight matrix provided in Figure 2, the range of neighbors varies significantly, with our minimum value and maximum value of one to 23, respectively, with mean of six neighbors and median of 6.5 neighbors. The distribution of data is positively skewed. Coupled with the large difference in the number of neighbors across spatial units varies significantly may induce heteroskedasticity into our models.

In our diagnostics for heteroskedasticity, we have p-values of 0.0 for both our Breusch-Pagan and Koenker-Bassett tests, and we reject the null hypothesis of heteroskedasticity: our current OLS model exhibits homoskedasticity. Therefore, the errors given in Table 2 are incorrect as the standard errors given for each variable depend on the assumption of homoskedasticity.

Moran's I p-value for all three models are 0.0, showing that our data does not have spatial randomness but instead exhibits spatial dependence. We then checked for spatial lag and spatial error in our model using Lagrange Multiplier lag and Lagrange Multiplier spatial error tests, respectively, and found that both where significant at the five percent level for all three models. Accordingly, we ran the robust versions of our corresponding Lagrange Multiplier tests and again found that for both spatial error and spatial lag tests, we again reject the null hypothesis for our models.

The robust Lagrange Multiplier lag test shows that the error term of our current OLS models are correlated with the regressors and therefore, all coefficients estimated for our models above are biased. The robust Lagrange Multiplier error test reveals that our error terms are correlated and therefore, the formula used to calculate our standard errors as shown in Table 2 are incorrect and so, the inference on our model is incorrect.

As census block groups are drawn based on population and have varying shapes and areas, we also conducted spatial diagnostic tests using a distance-based weights matrix in which a block group i was deemed to be a neighbor of block group j if the distance between i and j was less than 2.5 miles (using arcdistance as our distance measure). This band was chosen as it was the smallest bandwidth needed to avoid an isolates and ensure that each block group had at least one neighbor. Again, we found that our data was not spatially random through the Moran's I test. After conducting the Lagrange

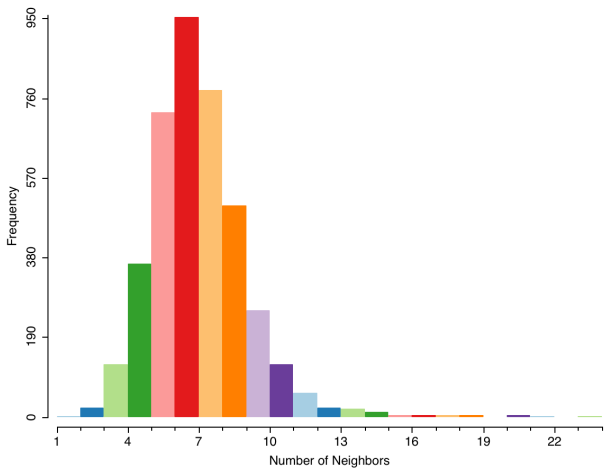

Figure 2: Histogram of Neighbors using First-Order, Queen Continguity Matrix

Multiplier spatial lag and spatial error tests, we found that we again must reject the null hypotheses for both tests. We then completed the robust versions of Lagrange Multiplier spatial lag and spatial error tests and found our p-values for both tests were 0.0, and therefore, we must reject both hypotheses: with our updated weights matrix, we again know that our OLS regression is biased and inefficient.

4.1 Future Work & Considerations

Although similar previous literature faced similar issues [5], because both robust Lagrange Multiplier in our tests were significant, we may have model misspecification and may need to add additional variables into our regression. As a basic test for misspecification, we ran the combined spatial lag and spatial error model in GeodaSpace using our first order queen weights matrix and found that the p-value for our λ term is 0.0761 and therefore not significant, reflecting that a spatial error model may not be appropriate and demonstrating that we currently have model misspecification.

In future regression iterations, additional explanatory variables may be added, and it may be worth differentiating between L stops and bus stops as opposed to equating the modes of transit as is done in the num_stops variable above. This differentiation may be worthwhile as bus stops are subject to the same congestion issues while subway stops do not suffer from the same traffic. On the other hand, as many commuters utilize bus services to complete the first or last portions of their trips before/after transferring to an L line, there may not be a statistically significant difference in decomposing our num_stops variable. Additionally, while two types of weight matrices were examined in the analysis above, further understanding and testing of appropriate weight matrices will completed going forward as well.

References

[1] National Equity Access. Commute time, United States. `https://nationalequityatlas.org/indicators/Commute_time`, 2020. Accessed: 2020-05-07.

[2] HHS ASPE. HHS Poverty Guidelines for 2020. `https://aspe.hhs.gov/poverty-guidelines`, 2020. Accessed: 2020-05-07.

[3] Boer Cui, Geneviève Boisjoly, Ahmed El-Geneidy, and David Levinson. Accessibility and the journey to work through the lens of equity. *Journal of Transport Geography*, 74:269–277, 2019.

[4] Lingqian Hu. Racial/ethnic differences in job accessibility effects: Explaining employment and commutes in the los angeles region. *Transportation Research Part D: Transport and Environment*, 76:56–71, 2019.

[5] Mizuki Kawabata and Qing Shen. Commuting inequality between cars and public transit: The case of the san francisco bay area, 1990-2000. *Urban Studies*, 44(9):1759–1780, 2007.

[6] Jay Koziarz. Chicago's traffic was the second worst in the nation in 2019, says report. https://chicago.curbed.com/2019/2/14/18224967/chicago-traffic-report-worst-nation-transportation, 2020. Accessed: 2020-05-07.

[7] Robert Krol and Shirley Svorny. The effect of rent control on commute times. *Journal of Urban Economics*, 58(3):421–436, 2005.

[8] Md Moniruzzaman and Antonio Páez. Accessibility to transit, by transit, and mode share: application of a logistic model with spatial filters. *Journal of Transport Geography*, 24:198–205, 2012.

[9] Annette Schaefer. Commuting Takes Its Toll. https://www.scientificamerican.com/article/commuting-takes-its-toll/, 2019. Accessed: 2020-04-26.