An Analysis of Chicago Taxi Trips Data: Modeling area-wide count outcomes with a spatial autoregressive model *

Dongping Zhang†

June 2017

(version 07.06.a)

Abstract

Taxicab is an important component of urban transit system since it caters to a large amount of demand and covers a wide geographic area. In this paper, the spatial variation of urban taxi ridership is studied by using a large scale Chicago taxi trips data. Taxi ridership is analyzed by relating it to various spatially explicit socio-demographic, built-environment, and urban-transportation variables. A spatial autoregressive (SAR) model is implemented to account for the spatial dependence of taxi demands among seventy-seven Chicago communities together with three different spatial weights for sensitivity check. The results suggest that SAR model outperforms the ordinary least square model in both goodness of fit and explanatory accuracy, and rook-contiguity weights seems to be the optimal weights modelling taxi ridership in Chicago.

keywords: spatial dependence, spatial autocorrelation, spatial lag, spatial variation, taxicab

*This paper is the final project for *Spatial Regression Analysis* course in Spring Quarter 2017 at the University of Chicago. I want to express my sincerest gratitude to the course instructor Dr. Luc Anselin for your teaching, guidance and support throughout the quarter.

†The University of Chicago, MA in Computational Social Science, dpzhang@uchicago.edu.

1 Introduction

Taxicab is an indispensable mode of transportation to a city's transit system because it could complement other public transportation alternatives through its flexibility and availability. Qian and Ukkusuri (2015) suggests the significance of taxi industry is reflected by its fleet size and the number of passengers served in a daily basis. According to Deng and Ji (2011), more than 50,000 taxis were serving in Shanghai, operated by over 100 taxi companies in 2011. As for New York City, Bloomberg and Yassky (n.d.) claims 485,000 trips are made per day on average.

The existing literatures have widely examined the casual factors that influence transit demand. McNally (2007) claims external factors such as population and employment rate are viewed as reasonable proxies for the amount of passengers. On the other hand, Kanafani (1983) believes internal factors such as travel time, costs and the level of service are also closely associated with transit ridership. As for taxicabs, Yang (2000) studied passenger demand, taxi utilization, and level of services. Deng and Ji (2011) used exploratory spatial data analysis (ESDA) and explored the spatiotemporal structure of taxi services in Shanghai, China from a macro perspective. Lee and Park (2008) analyzed passenger pick-up pattern in Jeju Island, Korea. Liu and Ratti (2010) systematically studied the driving strategy of the top-performing cab drivers in Shengzhen, China. Li (2011) developed a novel method to represent the passenger-finding strategies using time-location-strategy triplets using taxi trips data in Hangzhou, China. Veloso and Bento (2011) performed an exploratory analysis to visualize the spatiotemporal variation of taxi services and explored the relationships between pickup and dropoff locations in Lisbon, Portugal. Meanwhile, few efforts have been made to explore the determinants for taxi ridership in major cities of the United States except New York City (NYC). Ferreira and Silva (2013) proposed a new model that allows users to visually query taxi trips using NYC taxi trips data and Qian and Ukkusuri (2015) used a geographically weighted regression to model the spatial heterogeneity of the NYC taxi ridership and to visualize the spatial distributions of parameter estimations.

According to Ferreira and Silva (2013), taxis are valuable sensors and information associated with taxi trips can provide unprecedented insight into many different aspects of city life, from economic activity and human behavior to mobility patterns. A lack of studies in taxi trips in the United States could be attributed to three reasons. (1) data is very limited, and the current available taxi trips data set are from NYC and Chicago, which was released in 2015 and 2013 correspondingly. (2) analyzing taxi data presents many challenges because trip data are complex, containing geographical and temporal components in addition to multiple variables associated with each trip.

McNally (2007) stated that the traditional approaches of ridership analysis are dominated by the canonical four-step model and ordinary least square (OLS) multiple regression models. Gutiérrez and García-Palomares (2011) thinks OLS model is preferred because it is relatively quick, less expensive and more suitable for the analysis at finer scale. Nevertheless, OLS has an important assumption of independence of error terms, which does not apply to data varies with time or space. For example, Kabacoff (2015) gave an example that time series data will often display autocorrelation – observations collected closer in time will be more correlated with each other than with observations distant in time. Same concept applies to spatial data, Anselin et al. (2013) defines spatial autocorrelation as an association between value similarity and spatial similarity, or the correlation of a variable with itself through space. Kutner and Neter (2004) confirmed that failure to meet OLS assumption of independence of errors would make OLS coefficients no longer efficient and no longer have the minimum variance property, which biases the standard errors of regression coefficients. So, a spatial autoregressive model (SAR), suggested by Anselin and Rey (1991), is one of the alternatives to overcome this shortcoming.

In this paper, taxi ridership is analyzed by relating it to various spatially explicit socio-demographic, built-environment, and urban-transportation variables. A spatial autoregressive (SAR) model is implemented to account for the spatial dependence of taxi demands among seventy-seven Chicago communities together with three different spatial weights for sensitivity check. The results suggest that SAR model outperforms the ordinary least square model in both goodness of fit and explanatory accuracy, and

rook-contiguity weights is the optimal weights modelling taxi ridership in Chicago.

2 Data

All data sets used in this study are publicly available government administrative data, and the two main data sources to obtain all needed data sets are *The Chicago Data Portal* and *The Chicago Metropolitan Agency for Planning (CMAP)*. Multiple datasets and files need to be downloaded from each of the sources and Appendix A provides detailed elaborations on how to obtain, clean, and process raw data sets in order to reproduce the analysis.

Five data sets need to be downloaded from *The Chicago Data Portal*, and they are (1) *Chicago Taxi Trips Dataset*, which contains every single detailed taxi rides collected by Global Positioning System (GPS) digital traces installed in each taxi from 2013 to the present and the data set is still updated monthly (Note: this dataset is approximately 40 GBs in size). (2) *Chicago Transit Authority Bus Stops Dataset*, which contains every spatial coordinates of bus stops in the Chicago area. (3) Chicago Transit Authority L'Train Stations Dataset, which contains every spatial coordinates of L'Train stations in the Chicago area. (4) *Hardship Index*, which contains the hardship index of each community. This statistic incorporates six selected socioeconomic indicators and ranges from 1 to 100 with a higher index number representing a greater level of hardship. (5) *Current Chicago Community Boundary Shapefile*, which is a digital vector storage format for storing geometric locations. It contain the current *Chicago 77* by geospatial vectors, which spatially describe each Chicago community by polygons.

CMAP Community Data is consisting of 77 PDF files corresponding to 77 Chicago communities. Each file contains variables that summarize demographics, housing, employment, transportation habits, retail sales, property values, and land use of a specific Chicago Community Area. Data are compiled from the U.S. Census Bureau's 2010-14 American Community Survey, Longitudinal Employment-Household Dynamics data for 2014, etc.

Since *Chicago Taxi Trips Dataset* is a fairly new data set released by the City of Chicago in late 2013, barely any literature has used this particular data set yet.

2.1 Processing

All raw variables are comprehensively presented in Table 11 of Appendix A.2, together with their descriptions and sources. Because different communities have varying characteristics, and some socio-demographic variables, such as populations, is highly correlated with the number of taxi pickups in a community, it is crucial to ensure variables that are large in size would not affect the aggregated counts of taxi demands or pickups. Thus, variable processing has to be implemented so as to avoid potential "size problem".

Taxi Pickup Counts Taxi pickups is divided by the total acres so as to obtain number of pickups per acre in a community.

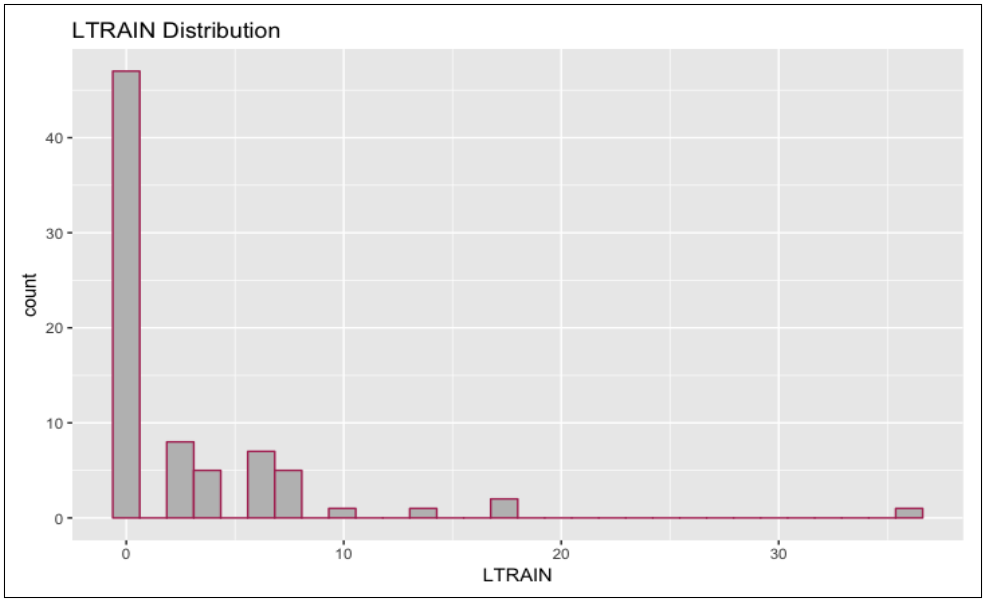
BLACK, BS, COMMUTERS, BUS Variables such as the number of African American (**BLACK**), the number of people holding a Bachelor's Degree or Higher (**BS**), the number of bus stops (**BUS**), and the number of commuters (**COMMUTER**) would be divided by the total acres so as to transform those variables to becomes measurements of density.

LTRAIN 47 out of 77 neighborhoods do not have L'Train stations, which is approximately 61% of all observations. Summary statistics of **LTRAIN** and its distribution could be viewed in Table 1 and Figure 1. In order to stabilize model, **LTRAIN** would be transformed into a binary dummy control variable in the regression model, where "1" indicates there is at least one L'train station in the neighborhood and "0" otherwise.

Table 1: Summary Statistics of **LTRAIN**

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
LTRAIN	0.000	0.000	0.000	2.779	4.000	36.000

Figure 1: LTRAIN Distribution



RES, COM, PARK Total residential area is obtained by summing single-family Residential acres and multi-family Residential acres. In addition, to avoid "size problem" mentioned above, total residential acres, total commercial acres, and total park acres are divided by total community acres, converting total acres to percentage terms.

All processed variables and corresponding descriptions are presented in Table 2. Dependent variables are taxi trip pickup counts. Independent variables are classified into three categories: (1) Socio-demographic, (2) Built-environment, (3) Urban-transportation.

A density map of aggregated four-year pickups could be viewed in Figure 3 of Appendix C. Density maps and unique value map for every independent variables could be assessed by Figure 4 in the same appendix. Possible influential observations are assessed in Appendix E.

Table 2: List of Dependent Variables and Potential Independent Variables

List of Dependent Variables

Variable	Definition
PICKUP13PA	2013 taxi pickup density in a community
PICKUP14PA	2014 taxi pickup density in a community
PICKUP15PA	2015 taxi pickup density in a community
PICKUP16PA	2016 taxi pickup density in a community
TOTALPA	aggregated taxi pickups density in a community

Note: Suffix **PA** indicates **P**er **A**cre

List of Potential Independent Variables

Category 1: Socio-demographic	
Variable	Definition
BLACK	African American density in a community
MINC	the median income of a community
BS	density of people holding a B.S. degree or higher in a community
COMMUTER	density of commuters in a community
HARD	the hardship index of a community
Category 2: Built-environment	
Variable	Definition
PARKP	percentage of community land that is park
RESP	percentage of community land used for residential purpose
COMP	percentage of community land used for commercial purpose
Category 3: Urban-transportation	
Variable	Definition
BUSPA	CTA bus stops density of a community
LTRAIND	dummy control indicating existence of L'Train stations in a community

Note: Suffix **P**, **PA**, and **D** indicates **P**ercentage, **P**er **A**cre, and **D**ummy correspondingly

3 Models

3.1 Global Model Specification

According to Table 2, there are 10 potential predictors possible, and thus initial global model specification is presented in Equation 1 where $i = 1 \dots 77$, the number of observations, and $j = 1 \dots 10$, the number of independent variables, which are listed in Table 2.

$$\textbf{PICKUPS}_i = \beta_0 + \sum_{j=1}^{10} \beta_j X_{ij} + \epsilon_i \tag{1}$$

Detailed diagnostics of Global Model (OLS) assumptions could be viewed in Appendix D. After backing up all assumptions, the global model specification becomes Equation 2 where $i = 1 \dots 77$, the number of observations, and $j = 1 \dots 7$, the number of independent variables. Independent variables in Equation 2 are (1) communter density, (2) hardship index, (3) percentage of residential acres, (4) percentage of commercial acres, (5) CTA bus stops density, (6) a dummy control for CTA L'Train stations.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{2}$$

or

$$log(\boldsymbol{PICKUPS}_i) = \beta_0 + \sum_{j=1}^{6} \beta_j X_{ij} + \beta_7 log(\boldsymbol{COMP})_i + \epsilon_i$$

3.2 Spatial Autocorrelation

After solidifying assumptions for the global model presented in Equation 2, an assessment of spatial autocorrelation using Moran's I on the residuals of the global model would be implemented in order to detect the presence of spatial dependence following methods derived by Anselin (1993).

As Anselin et al. (2013) stated, for a set of n spatial observations for a variable y,

Moran's I is given as Equation 3, where z_i is the deviation from the mean, or $y_i - y_{bar}$ and $w_{i,j}$ is the spatial weights.

In this study, three different types of spatial weights are used for sensitivity check, among which two are contiguity-based and the one is distance-based with arc-distance threshold value set to be 4.61588 miles. The two different types of contiguity-based spatial weights are rook-contiguity and queen-contiguity. Appendix F elaborates on the concept of spatial weights and it also contains detailed summary statistics of neighbor distributions for each weight.

$$I = \frac{n}{S_0} \times \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} z_i w_{i,j} z_j}{\sum_{i=1}^{n} z_i^2} \tag{3}$$

Using Moran's I test statistic stated in Equation 3, an inference is conducted. The null hypothesis of randomly distributed residuals of global model in space is tested against the alternative hypothesis of non-randomly distributed residuals. Three different spatial weights mentioned above are used to test sensitivity. The compiled results are presented in Table 3.

Table 3: Spatial Autocorrelation Diagnostics

Weights / **Moran's I**	**Queen-contiguity**		**Rook-contiguity**		**Distance-based**	
Residuals	I	**P-value**	I	**P-value**	I	**P-value**
log(PICKUP13PA)	2.3394	0.01932*	2.4121	0.01586**	3.8482	0.00012***
log(PICKUP14PA)	3.1451	0.00166**	3.2968	0.00098***	4.1742	0.00003***
log(PICKUP15PA)	3.1466	0.00165**	3.2083	0.00134**	4.7330	0.00000***
log(PICKUP16PA)	2.7750	0.00552**	2.9857	0.00283**	3.6245	0.00029***
log(TOTALPA)	2.8637	0.00419**	2.9870	0.00282**	4.1178	0.00004***
Note: *p<0.05; **p<0.01; ***p<0.001						

As shown, Moran's I seems to be most sensitive to distance-based weights using centroids of the community polygons. All three sets of statistics seem to exhibit an increasing pattern between 2013 to 2015 and decreases in 2016. Most importantly, almost all p-values are statistically significant at 5% significance level, so it is safe to

claim that the residuals of the global models exhibit spatial autocorrelation.

After confirming spatial autocorrelation in the residuals of the global model, *Lagrange Multiplier* or *Rao Score* test statistics is used to discover SAR specification according to Anselin and Rey (2014). Appendix G.2 describes LM_ρ statistic used to detect spatial lag, Appendix G.3 describes LM_λ statistic used to detect spatial error, and Appendix G.4 describes the robustified statistics, LM_ρ^* and LM_λ^* in detail.

Table 4: Spatial Dependence Diagnostics

Weights / Lagrange		Queen-contiguity		Rook-contiguity		Distance-based	
Description	**Test**	χ^2	**P-value**	χ^2	**P-value**	χ^2	**P-value**
Spatial Lag	LM_ρ	7.7862	0.00526**	8.1787	0.00424**	7.4904	0.00620**
Robust Spatial Lag	LM_ρ^*	6.4823	0.01090*	5.8552	0.01553*	6.4515	0.01109*
Spatial Error	LM_λ	1.9637	0.16111	2.7183	0.09920	1.2543	0.26274
Robust Spatial Error	LM_λ^*	0.6598	0.41662	0.3948	0.52980	0.2154	0.64260
Note: *p<0.05; **p<0.01							

The results of *Lagrange Multiplier* test statistics are statistically consistent as presented in Table 4. LM_ρ are all statistically significant at 5% significance level for a χ^2 variate with one degree of freedom, strongly indicating the presence of spatial autocorrelation thus spatial dependence. The robustified statistics, LM_ρ^*, are also statistically significant at 5% significance level for all three tests, which decides the lag specification.

3.3 Spatial Autoregressive Model Specification

Anselin and Rey (2014) defines the spatial autoregressive model (SAR) lag specification to be Equation 4, where $\boldsymbol{W}\,log(\boldsymbol{PICKUPS}_i)$ is the spatially lagged dependent variable with associated autoregressive coefficient ρ and spatial weights $\boldsymbol{W}$. The rest of the notation is as the global model stated in Equation 2.

$$\boldsymbol{Y} = \rho \boldsymbol{W} \boldsymbol{Y} + \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4}$$

or

$$log(\boldsymbol{PICKUPS}_i) = \rho \boldsymbol{W}\, log(\boldsymbol{PICKUPS}_i) + \beta_0 + \sum_{j=1}^{5} \beta_j X_{ij} + \beta_6 log(\boldsymbol{COMP})_i + \epsilon_i$$

where

- ρ is the autoregressive coefficient
- $\boldsymbol{W}$ is the spatial weighting matrix
- $\boldsymbol{W}\, log(\boldsymbol{PICKUPS}_i)$ is the spatial lag

4 Estimation

Four regression models are implemented on annual taxi pickups per acre from 2013 to 2016 as well as on the aggregated pickups per acre across four years, thus 20 regression models total. The four regression models are the global model and the spatial autoregressive models using three different spatial weights. The regression results could be viewed in Appendix H from Table 24, Table 25, Table 26, Table 27, and Table 28.

The model diagnostics table for aggregated pickups per acre and yearly pickups per acre are presented in Table 5, Table 6, Table 7, Table 8, and Table 9 below.

All independent variables are statistically significant at 0.05 level except commercial acre percentage for all five global models. Based on the coefficient values, many of the variables exhibit an intuitive relationship with respect to the taxi ridership, such as the positive relationship with commuter density and commercial acre percentage as well as the negative relationship with hardship index and residential acre percentage. One interesting finding is that all public-transportation variables, CTA bus stops

Table 5: Model Diagnostics of Aggregated Pickups per acre

	Dependent variable:			
	Aggregated Total Pickups per acre			
	Global Model OLS	*Spatial Autoregressive Model SAR*		
Weights	(NA)	(Queen)	**(Rook)**	(Distance)
R^2	0.820603	0.866891	**0.870133**	0.864794
Log Likelihood	-117.293	−107.788	**−107.095**	−107.374
σ^2	1.35528	0.914171	**0.891908**	0.928578
Akaike Inf. Crit.	248.586	231.576	**230.19**	230.748
Schwarz criterion	264.993	250.326	**248.941**	249.499
Residual Std. Error	1.16417	0.956123	**0.944409**	0.963628

*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:			
	2013 Pickups per acre			
	Global Model OLS	*Spatial Autoregressive Model SAR*		
Weights	(NA)	(Queen)	(Rook)	**(Distance)**
R^2	0.825818	0.868419	0.870013	**0.871005**
Log Likelihood	-117.284	−108.292	−108.003	**−106.699**
σ^2	1.35497	0.930	0.919	**0.912238**
Akaike Inf. Crit.	248.569	232.622	232.006	**229.399**
Schwarz criterion	264.975	251.372	250.756	**248.149**
Residual Std. Error	1.16403	0.964638	0.958778	**0.955111**

*p<0.1; **p<0.05; ***p<0.01

Table 6: Model Diagnostics for 2013 Pickups per acre

	Dependent variable:			
	2014 Pickups per acre			
	Global Model OLS	*Spatial Autoregressive Model SAR*		
Weights	(NA)	(Queen)	**(Rook)**	(Distance)
R^2	0.813262	0.862084	**0.866726**	0.856449
Log Likelihood	-117.768	−108.165	**−107.18**	−108.586
σ^2	1.37211	0.921251	**0.890243**	0.958896
Akaike Inf. Crit.	249.537	232.329	**230.36**	233.173
Schwarz criterion	265.943	251.08	**249.11**	251.923
Residual Std. Error	1.17137	0.959818	**0.943527**	0.979232

*p<0.1; **p<0.05; ***p<0.01

Table 7: Model Diagnostics for 2014 Pickups per acre

	Dependent variable:			
	2015 Pickups per acre			
	Global Model OLS	*Spatial Autoregressive Model SAR*		
Weights	(NA)	(Queen)	**(Rook)**	(Distance)
R^2	0.821645	0.874534	**0.876862**	0.871246
Log Likelihood	-118.679	−107.42	**−106.921**	−107.197
σ^2	1.40495	0.898486	**0.881812**	0.922032
Akaike Inf. Crit.	251.358	230.839	**229.843**	230.395
Schwarz criterion	267.765	249.59	**248.593**	249.145
Residual Std. Error	1.18531	0.947885	**0.939049**	0.960225

*p<0.1; **p<0.05; ***p<0.01

Table 8: Model Diagnostics for 2015 Pickups per acre

	Dependent variable:			
	2016 Pickups per acre			
	Global Model OLS	*Spatial Autoregressive Model SAR*		
Weights	(NA)	(Queen)	**(Rook)**	(Distance)
R^2	0.808471	0.848962	**0.853404**	0.847443
Log Likelihood	-118.545	−111.067	**−110.204**	−110.664
σ^2	1.40008	1.00372	**0.974197**	1.01381
Akaike Inf. Crit.	251.09	238.133	**236.408**	237.328
Schwarz criterion	267.497	256.884	**255.159**	256.078
Residual Std. Error	1.12818	1.00186	**0.987014**	1.00688

*p<0.1; **p<0.05; ***p<0.01

Table 9: Model Diagnostics for 2016 Pickups per acre

density and CTA L'Train dummy, have a positive relationship with taxi ridership on average. It is intuitive to speculate these positive coefficients are because communities with greater density of bus stops and train stations are communities where people tend to cluster. The more people in such communities, the more likely the ridership. On the other hand, based on the regression result, it is also statistically sound to make a claim that public transportation system might have a positive spillover effect that the diverse modes of transportation might be able to boost ridership.

According to Table 5, Table 6, Table 7, Table 8, and Table 9, it is also evident that, by addressing spatial dependence of taxi pickup counts, SAR outperforms OLS in the explanatory power as well as the goodness of model fit in general. Among all three SAR models using three different spatial weights, rook-contiguity weights is seemingly to be the optimal spatial weights used to model Chicago taxi ridership.

5 Conclusion

In this paper, Chicago taxi ridership is analyzed by relating it to a total of six spatially explicit socio-demographic, built-environment, and urban-transportation variables. An OLS is used as a global model and is compared to an SAR model using three different spatial weights. Unsurprisingly, SAR model outperforms OLS in both goodness of fit and explanatory accuracy. The reason behind this superiority is because SAR is able to account for the spatial autocorrelation of taxi demands among seventy-seven Chicago communities and to address this spatial dependence by having a spatial lag variable in the model specification as a regressor. In addition, three different spatial weights are used for sensitivity check and it turns out that rook-contiguity weights is seemingly to be the optimal spatial weights used to model Chicago taxi ridership.

One of the limitations of this study is a lack of data in alternative transportation means. Ride-sharing companies, such as Uber and Lift would be two profound impacting variables that could potentially negatively influence taxi demands in each community of Chicago. Also, a highly segregated city Chicago is, spatial dependence

of taxi pickups may not be able to fully incorporate the nature of taxi demands in Chicago, meaning a geographically weighted regression model might be more suitable by accounting for spatial heterogeneity. As a future work, it will be meaningful to extend the study in other large cities and explore the similarities and differences on the impact of explanatory variables in different urban areas. This will help to explain how the form and function of urban spaces result in taxi demand.

References

Anselin, Luc, "Exploratory spatial data analysis in a geocomputational environment," *Geocomputation, a primer*, 1988, pp. 77–94.

_ , "Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity," *Geographical analysis*, 1988, *20* (1), 1–17.

_ , *Spatial econometrics: methods and models*, Kluwer Academic Publishers, Dordrecht, 1988.

_ , *The Moran scatterplot as an ESDA tool to assess local instability in spatial association*, Morgantown, WV: Regional Research Institute, West Virginia University, 1993.

_ , "Simple diagnostic tests for spatial dependence." Regional science and urban economics," *Regional science and urban economics*, 1996, *26.1* (77-104).

_ , "Rao's score test in spatial econometrics," *Journal of statistical planning and inference*, 2001, *97* (1), 113–139.

_ , "Thirty years of spatial econometrics," *Papers in regional science*, 2010, *89* (1), 3–25.

_ **, Allen Murray, and Sergio Rey**, "Spatial Analysis," *The Oxford Handbook of Quantitative Methods, Vol. 2: Statistical Analysis*, 2013, *2* (154-174).

_ **and Anil K. Bera**, "Spatial dependence in linear regression models with an introduction to spatial econometrics," *Statistics Textbooks and Monographs*, 1998, *155* (237-290).

_ **and Serge Rey**, "Properties of tests for spatial dependence in linear regression models," *Geographical analysis*, 1991, *23* (2), 112–131.

_ **and Sergio Joseph Rey**, *Modern Spatial Econometrics in Practice*, GeoDa Press, 2014.

Bloomberg, Michael and David Yassky, "Taxicab Factbook," New York City Taxi and Limousine Commission.

Burridge, P., "On the Cliff-Ord test for spatial autocorrelation among regression residuals," *Geographical Analysis*, 1980, *4*, 267–284.

Deng, Zhongwei and Minhe Ji, "Spatiotemporal structure of taxi services in Shanghai: Using exploratory spatial data analysis," *Geoinformatics*, 2011.

Donald, Farrar E. and Robert R. Glauber., "Multicollinearity in regression analysis: the problem revisited.," *The Review of Economic and Statistics*, 1967, pp. 92–107.

Fox, John, *Applied regression analysis and generalized linear models*, Sage Publications, 2015.

H., Chris Nachtsheim Kutner Michael and John Neter, *Applied linear regression models*, McGraw-Hill/Irwin, 2004.

Javier, Osvaldo Daniel Cardozo Gutiérrez and Juan Carlos García-Palomares, "Transit ridership forecasting at station level: an approach based on distance-decay weighted regression," *Journal of Transport Geography*, 2011, *19.6*, 1081–1092.

Junghoon, Inhye Shin Lee and Gyung-Leen Park, "Analysis of the passenger pick-up pattern for taxi location recommendation," *Networked Computing and Advanced Information Management*, 2008.

Kabacoff, Robert, *R in Action: Data Analysis and Graphics with R*, 2, illustrated ed., Manning, 2015.

Kanafani, Adib, "Transportation demand analysis," 1983.

Li, Bin, "Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset," *Pervasive Computing and Communications Workshops*, 2011.

Liang, Clio Andris Liu and Carlo Ratti, "Uncovering cabdrivers' behavior patterns from their digital traces," *Computers, Environment and Urban Systems*, 2010, *34.6* (541-548).

M., Richard P. Nathan Montiel Lisa and David J. Wright, *An update on urban hardship*, 411 State Street Albany, New York 12203-1003: The Nelson A. Rockefeller Institute of Government, 2004.

Marco, Santi Phithakkitnukoon Veloso and Carlos Bento, "Urban mobility study using taxi traces," *Proceedings of the 2011 international workshop on Trajectory data mining and analysis*, 2011.

McNally, Michael G., "The four-step model," *Handbook of Transport Modelling*, 2007, pp. 35–53.

Nivan, Jorge Poco Huy T. Vo Juliana Freire Ferreira and Cláudio T. Silva, "Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips," *IEEE Transactions on Visualization and Computer Graphics 19*, 2013, (12), 2149–2158.

Qian, Xinwu and Satish V. Ukkusuri, "Spatial variation of the urban taxi ridership using GPS data," *Applied Geography*, 2015, *59*, 31–42.

Soltoff, Benjamin, "Diagnostic tests for OLS," 2017.

Yang, Hai, "A macroscopic taxi model for passenger demand, taxi utilization and level of services," *Transportation*, 2000, *27.3*, 317–340.

Appendices

A Data

In the Data section of this appendix, detailed elaborations would be presented to inform readers how to obtain, clean, and process raw data sets in order to reproduce the analysis. All data sets used in this study are publicly available government administrative data, and the two main data sources to obtain all needed datasets are *The Chicago Data Portal* and *The Chicago Metropolitan Agency for Planning (CMAP)*. Multiple datasets and files need to be downloaded from each of the sources.

Five data sets need to be downloaded from *The Chicago Data Portal*, and they are (1) *Chicago Taxi Trips Dataset*, which contains every single detailed taxi rides from 2013 to the present and the dataset is still updated monthly (Note: this dataset is approximately 40 GBs in size). (2) *Chicago Transit Authority Bus Stops Dataset*, which contains every spatial coordinates of bus stops in the Chicago area. (3) Chicago Transit Authority L'Train Stations Dataset, which contains every spatial coordinates of L'Train stations in the Chicago area. (4) *Hardship Index*, which contains the hardship index of each community. This statistic incorporates six selected socioeconomic indicators and ranges from 1 to 100 with a higher index number representing a greater level of hardship. (5) *Current Chicago Community Boundary Shapefile*, which is a digital vector storage format for storing geometric locations. It contain the current *Chicago 77* by geospatial vectors, which spatially describe each Chicago community by polygons.

CMAP Community Data is consisting of 77 PDF files corresponding to 77 Chicago communities. Each file contains variables that summarize demographics, housing, employment, transportation habits, retail sales, property values, and land use of a specific Chicago Community Area. Data are compiled from the U.S. Census Bureau's 2010-14 American Community Survey, Longitudinal Employment-Household Dynamics data for 2014, etc. Variables of each community could be obtained by downloading PDF files of that community and that makes 77 PDF files total. PDF scrapping would be required to compile actual variables for analysis, which would be detailed in later section of this appendix.

A.1 Creation

Chicago Taxi Trips Dataset is a 40GB dataset that contains 23 variables, but only three variables would be used in this study and they are *"Trip Start Timestamp"*, *"Pickup Centroid Latitude"* and *"Pickup Centroid Longitude"*.

"Trip Start Timestamp" is used to categorize pickup counts by year. The other two coordinate variables would be used together with the *Current Chicago Community Boundary Shapefile*. By checking whether each of the pickup coordinate is falling in any of the 77 community spatial polygon, counts by community could be obtained by aggregation by year (2013, 2014, 2015, and 2016) using the timestamp variable. However, not every taxi trip has pickup and dropoff coordinates, and those observa-

tions lacking spatial coordinates are removed from the analysis. Table 10 gives a brief summary on the percentage of observations lacks spatial coordinates.

Table 10: Brief Summary of Taxi Trips Missing Spatial Coordinates

Variables	Trips	# of NA	% of NA
2013	26,870,287	4,646,758	17.29
2014	31,020,726	4,527,054	14.59
2015	27,400,744	4,226,875	15.43
2016	19,878,240	2,757,663	13.87
Total	105,169,997	16,158,350	15.36

Chicago Transit Authority Bus Stops Dataset and Chicago Transit Authority L'Train Stations Dataset are processed using the same procedure as the *Chicago Taxi Trips Dataset*. There are 11074 bus stops and 234 L'Train stations in the Chicago area. By using *Current Chicago Community Boundary Shapefile*, the aggregated counts of bus stops and L' Train stations in each of the 77 areal units could be generated.

As noted, *CMAP Community Data* is consisting of 77 PDF files corresponding to 77 Chicago communities. Each of the file contains some key socio-demographic variables of that specific neighborhood. All PDF files have the same format, so any PDF scrapping method could be used to obtain variables of interest.

In short, a total of 18 raw variables need to be extracted and merged. Table 11 comprehensively presents all raw variables, together with their descriptions and sources.

A.2 Processing

Different communities have varying characteristics, and some socio-demographic variables, such as populations, is highly correlated with the number of taxi pickups in a community. It is crucial to ensure variables that are large in size would not affect the aggregated counts of trips. Thus, variable processing would be implemented so as to avoid "size problem".

PICKUP13, PICKUP14, PICKUP15, PICKUP16, TOTAL Taxi Pickups is divided by the total acres so as to obtain number of pickups per acre in a neighborhood.

BLACK, BS, COMMUTERS, BUS Variables such as the number of African American in a community (**BLACK**), the number of people holding a Bachelor's Degree or Higher in a community (**BS**), the number of bus stops in a community (**BUS**), and the number of commuters in a community (**COMMUTER**) would be

Table 11: List of Raw Variables

Variable	Description	Source	Where
BLACK	black population of a community	2014 American Community Survey five-year estimates	CMAP Community Data
MINC	the median income of a community		
BS	number of people holding a Bachelor's Degree or Higher in a community		
COMMUTER	total number of commuters in a community		
PARK	park acreage per 1,000 Residents	CMAP calculations of 2010 Land Use Inventory	
SRES	single-Family Residential land acres in a community	Chicago Metropolitan Agency for Planning Parcel-Based Land Use Inventory	
MRES	multi-Family Residential acres in a community		
COM	commercial acres in a community		
ACRE	total acres if a community		
POP	total population of a community	2000 and 2010 Census	
HARD	hardship index of a community	U.S. Census Bureau 2006-2010 American Community Survey 5-year estimates	Chicago Data Portal
PICKUP13	number of taxi pickups in a community in 2013	City of Chicago	
PICKUP14	number of taxi pickups in a community in 2014		
PICKUP15	number of taxi pickups in a community in 2015		
PICKUP16	number of taxi pickups in a community in 2016		
TOTAL	total number of taxi pickups from 2013 to 2016		
BUS	number of CTA bus stops in a community	Chicago Transit Authority	
LTRAIN	number of CTA L'Train stations in a community		

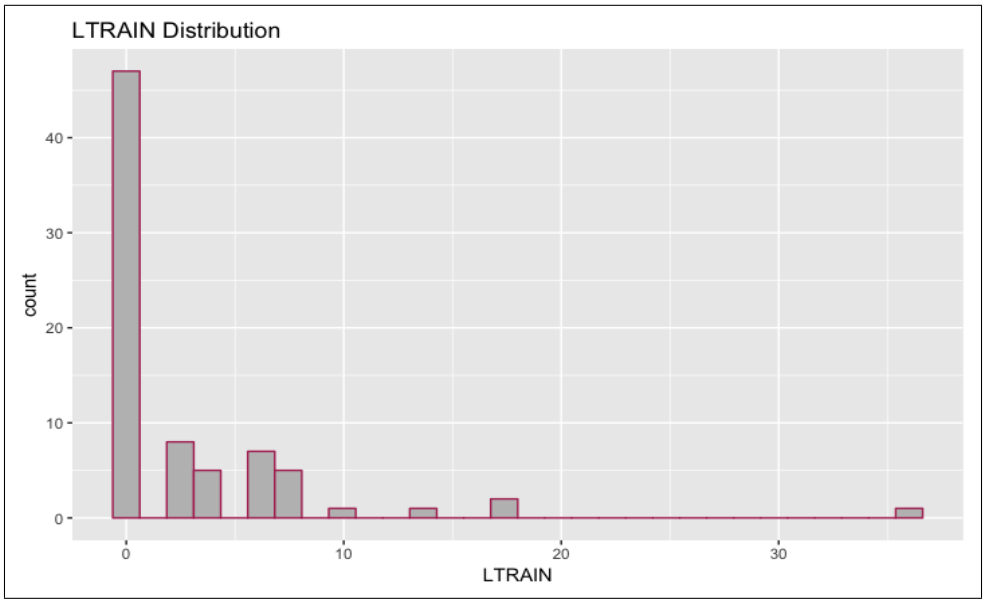
divided by the total acres so as to transform those variables to becomes measurements of density.

LTRAIN 47 out of 77 neighborhoods do not have L'Train stations, which is approximately 61% of all observations. Summary statistics of **LTRAIN** and its distribution could be viewed in Table 12 and Figure 2. In order to stabilize model, **LTRAIN** would be transformed into a binary dummy control variable in the regression model, where "1" indicate there is at least one L'train station in the neighborhood and "0" otherwise.

Table 12: Summary Statistics of **LTRAIN**

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
LTRAIN	0.000	0.000	0.000	2.779	4.000	36.000

Figure 2: **LTRAIN** Distribution



RES, COM, PARK Obtain total residential area by summing single-family Residential acres and multi-family Residential acres. In addition, to avoid "size problem" mentioned above, total residential acres, total commercial acres, and total park acres are divided by total community acres, converting total acres to percentage terms.

All processed clean variables and corresponding descriptions are presented in Table 13. Dependent variables are taxi trip pickup counts. Independent variables are classified into three categories: (1) Socio-demographic, (2) Built-environment, (3) Urban-transportation.

Table 13: List of Dependent Variables and Potential Independent Variables

List of Dependent Variables

Variable	Definition
PICKUP13PA	2013 taxi pickup density in a community
PICKUP14PA	2014 taxi pickup density in a community
PICKUP15PA	2015 taxi pickup density in a community
PICKUP16PA	2016 taxi pickup density in a community
TOTALPA	aggregated taxi pickups density in a community

Note: Suffix **PA** indicates **P**er **A**cre

List of Potential Independent Variables

Category 1: Socio-demographic	
Variable	Definition
BLACK	African American density in a community
MINC	the median income of a community
BS	density of people holding a B.S. degree or higher in a community
COMMUTER	density of commuters in a community
HARD	the hardship index of a community
Category 2: Built-environment	
Variable	Definition
PARKP	percentage of community land that is park
RESP	percentage of community land used for residential purpose
COMP	percentage of community land used for commercial purpose
Category 3: Urban-transportation	
Variable	Definition
BUSPA	CTA bus stops density of a community
LTRAIND	dummy control indicating existence of L'Train stations in a community

Note: Suffix **P**, **PA**, and **D** indicates **P**ercentage, **P**er **A**cre, and **D**ummy correspondingly

B Exploratory Data Analysis (EDA)

Table 14 provides brief summary statistics of all dependent variables and Table 15 provides brief summary statistics of all independent variables. For detailed variable description, please refer to Table 13 in Section A.2.

Table 14: Descriptive Statistics of All Dependent Variables

Statistic	Mean	St. Dev.	Min	Max
PICKUPS13PA	171.464	681.360	0.031	4,253.669
PICKUPS14PA	204.412	822.954	0.043	5,201.368
PICKUPS15PA	179.444	753.443	0.050	4,905.752
PICKUPS16PA	134.458	597.683	0.026	4,049.513
TOTALPA	689.777	2,852.277	0.168	18,410.300
Note: #Observations = 77 for all variables presented above.				

Table 15: Descriptive Statistics of All Independent Variables

Statistic	Mean	St. Dev.	Min	Max
BLACK	6.303	6.700	0.015	25.697
MINC	46,712.290	20,336.540	14,390	94,823
BS	4.867	6.309	0.057	31.322
COMMUTER	8.801	6.450	0.609	32.550
HARD	49.506	28.691	1	98
PARKP	6.725	15.963	0.200	124.700
RESP	0.334	0.129	0.027	0.576
COMP	0.053	0.043	0.006	0.287
BUSPA	0.076	0.030	0.004	0.154
Note: #Observations = 77 for all variables presented above. For variable LTRAIN, refers to Table 2 and Figure 12				

C Exploratory Spatial Data Analysis (ESDA)

Because the data set used for this study is highly spatial, EDA is not sufficient to get a distribution of statistics spatially. For spatial data, Anselin (2010) suggests to conduct Exploratory Spatial Data Analysis (ESDA). Following Anselin (1988a), exploratory spatial data analysis (ESDA) is a collection of techniques to describe and visualize spatial distributions; identify atypical locations or spatial outliers; discover patterns of spatial association, clusters or hot-spots; and suggest spatial regimes or other forms of spatial heterogeneity.

Figure 3 and Figure 4 below are some very rough and unpolished density maps and unique value maps conducted on some key variables of this study, and those variables are TOTALPA, COMMUTER, HARD, RESPA, COMPA, BUSPA, LTRAIND.

Figure 3: Density of Total Pickups per acre

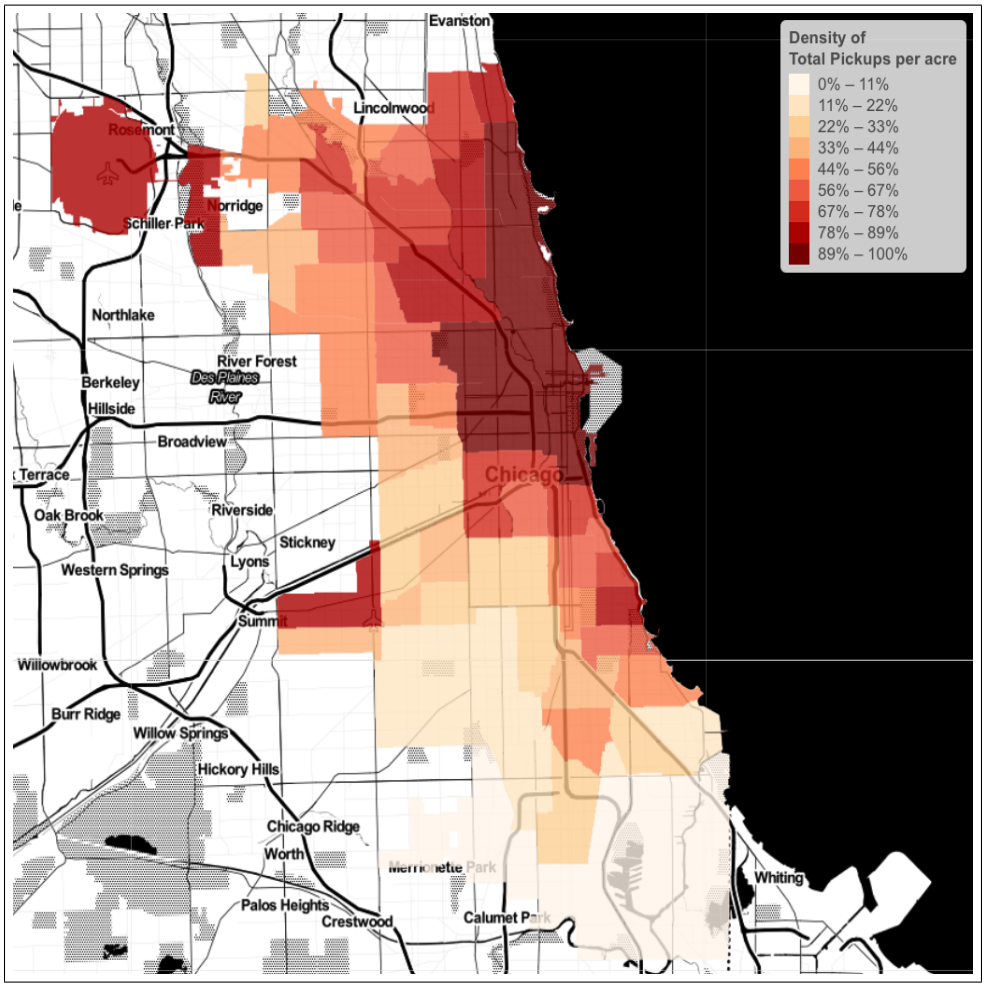
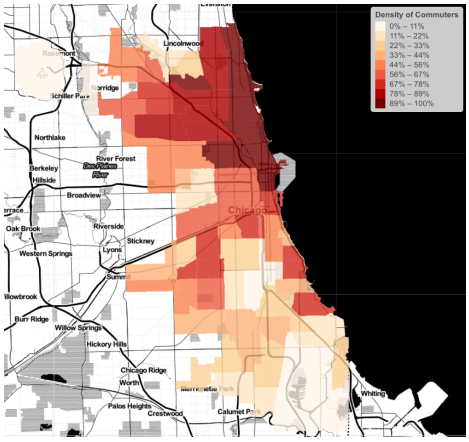
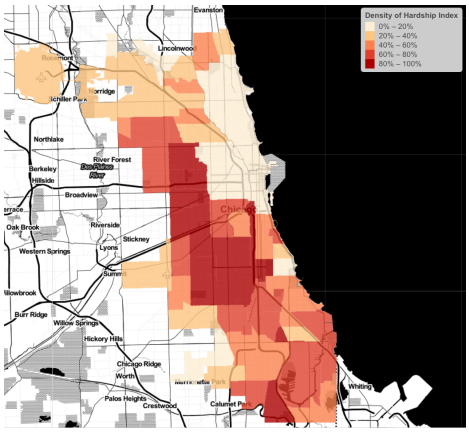


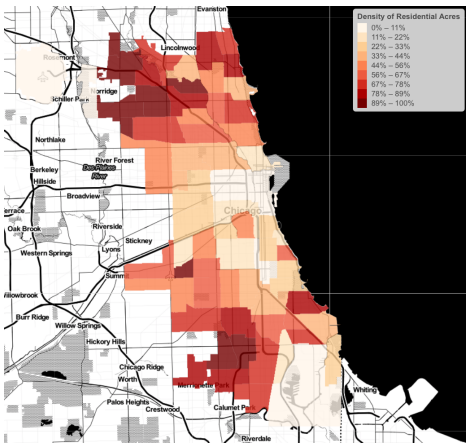
Figure 4: Density Map/Unique Value Map for Key Variables



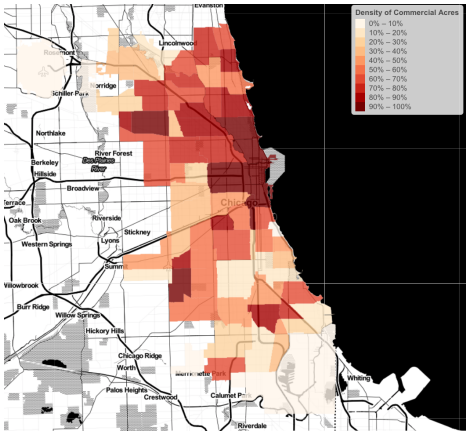
(a) Density of Commuters per acre



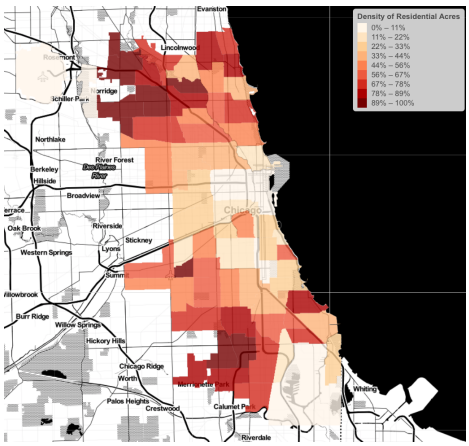
(b) Density of Hardship Index



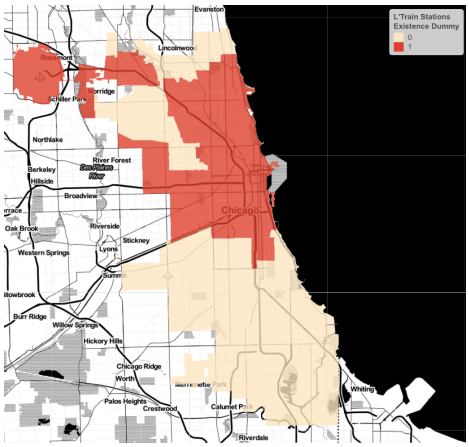
(c) Density of % of Residential Acres



(d) Density of % of Commercial Acres



(e) Density of Bus Stops per acre



(f) Unique Value Map for L'Train Stations

D Diagnostics of Global Model

D.1 Global Model (OLS) Specification

According to Table 2, there are 10 potential predictors possible, and thus the global model specification is presented in Equation A.4.1 where i represents the number of observations and $j = 1 \dots 10$, the number of independent variables.

$$\textbf{PICKUPS}_i = \beta_0 + \sum_{j=1}^{10} \beta_j X_{ij} + \epsilon_i \tag{A.4.1}$$

D.2 Detecting Multicollinearity

According to Donald and Glauber. (1967), although moderate multicollinearity would not be problematic to model estimation, it introduces biases by increasing variances to model coefficients and causes models very sensitive to minor changes. In this study, multicollinearity is assessed following two procedures, and they are Pearson product-moment correlation coefficients and variance inflation factor (VIF).

VIFs are first computed for 10 potential raw variables, and the results are presented in Table 16. According to Kabacoff (2015), the VIF value is an indicator for the severity of multicollinearity. As a rule of thumb suggested by Qian and Ukkusuri (2015), variables with VIF greater than 10 should be eliminated. It could be shown that there are at least some pairwise combinations of variables that have severe multicollinearity because variable **BS** and variable **COMMUTER** have significantly higher values, 15.011 and 13.881 correspondingly.

Table 16: VIF Table For Raw Global Model

BLACK	MINC	BS	COMMUTER	HARD	RESP	COMP	PARKP	BUSPA	LTRAIND
2.794	6.468	15.011	13.881	6.785	3.847	1.724	1.911	3.129	1.669

In order to decipher possible combinations of pairwise variables that are collinear, a pairwise correlation coefficients table is constructed.Table 17 presents pairwise correlation coefficients of every possible combination of independent variables. According to Kabacoff (2015), any pairwise coefficients greater than 0.7 imply the existence of collinearity. It is observed that most of the correlation coefficients are below 0.7; however median income, **MINC**, and hardship index, **HARD**, are highly negatively correlated with $\rho = -0.867$ and level of education, **BS**, and number of commuters, **COMMUTER**. are highly positively corelated with $\rho = 0.887$. The highly positive and highly negative coefficients make intuitive sense because wealthy communities with high median income certainly do not have high hardship indices while college graduates are more likely to have a job that requires daily commutes to their workplaces.

Table 17: Pearson product-moment correlation coefficient for explanatory variables

	BLACK	MINC	BS	COMMUTER	HARD	PARKP	RESP	COMP	BUSPA	LTRAIND
BLACK	1	-0.513	-0.126	-0.208	0.342	-0.123	0.003	-0.135	0.278	-0.123
MINC	-0.513	1	0.523	0.404	**-0.865**	0.044	0.301	0.355	-0.022	0.178
BS	-0.126	0.523	1	**0.887**	-0.654	-0.088	0.062	0.489	0.554	0.450
COMMUTER	-0.208	0.404	**0.887**	1	-0.489	-0.239	0.260	0.462	0.579	0.431
HARD	0.342	**-0.865**	-0.654	-0.489	1	-0.139	-0.262	-0.333	-0.076	-0.247
PARKP	-0.123	0.044	-0.088	-0.239	-0.139	1	-0.435	-0.022	-0.361	0.065
RESP	0.003	0.301	0.062	0.260	-0.262	-0.435	1	-0.146	-0.069	-0.255
COMP	-0.135	0.355	0.489	0.462	-0.333	-0.022	-0.146	1	0.424	0.375
BUSPA	0.278	-0.022	0.554	0.579	-0.076	-0.361	-0.069	0.424	1	0.459
LTRAIND	-0.123	0.178	0.450	0.431	-0.247	0.065	-0.255	0.375	0.459	1

In order to avoid multicollinearity in the global model (OLS), only one of the two variables from each pair should be used. In this study, hardship index and number of commuters would be kept and used as independent variables because: (1) hardship index is more representative because Montiel and Wright (2004) incorporates six selected socioeconomic indicators into hardship: (i) the percent of occupied housing units with more than one person per room (i.e., crowded housing); (ii) the percent of households living below the federal poverty level; (iii) the percent of persons aged16 years or older in the labor force that are unemployed; (iv) the percent of persons aged 25 years or older without a high school diploma; (v) the percent of the population under 18 or over 64 years of age (i.e., dependency); and (vi) per capita income. (2) commuters are more representative target population who are more likely to take taxi rides.

The variance inflation factor (VIF) is computed again in Table 18 after removing collinear variables, **BS** and **MINC**, to reassess the global model. As shown in table 18, all variables have VIF values less than 10, which suggests multicollinearity problem is resolved.

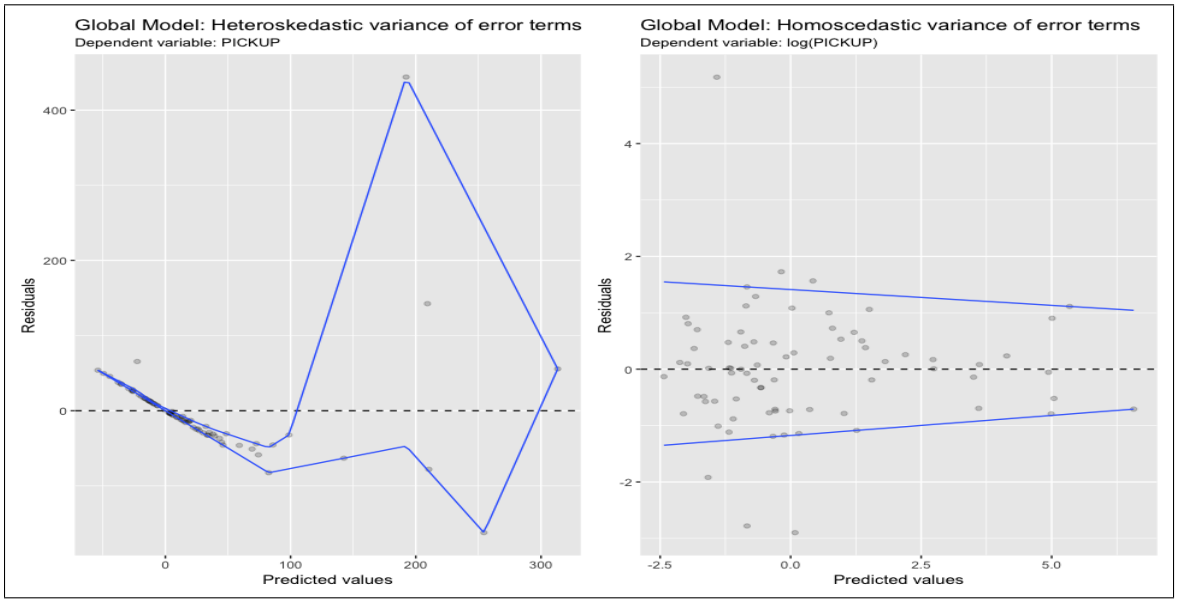
Table 18: VIF Table After Removing Milticollinear Variables

BLACK	COMMUTER	HARD	RESP	COMP	PARKP	BUSPA	LTRAIND
1.608	2.937	1.837	2.096	1.544	1.832	3.124	1.642

D.3 Detecting Heteroskedasticity

As a rule of thumb, residual vs. fitted value plots would be first generated to visually detect possible heteroskedasticity problem. The left side of Figure 5 presents the residual vs. fitted value plot of un-calibrated global model, which shows obvious pattern of heteroskedasticity. To statistically validate, a Breusch-Pagan test with H_0 hypothesis of constant variance is implemented. The p-value obtained is 5.38×10^{-57}, which is highly statistically significant and it rejects null hypothesis. The p-value confirms heteroskedasticity, suggesting a power transformation might be needed on the dependent variable.

Figure 5: Log Transformation of Dependent Variable PICKUP



According to Kabacoff (2015), a spread-level plot (SL plot) is created to get a sense of power transformation, which is also showed in Figure 6. The slope of SL plot is 0.9745824, which makes the optimal power $\boldsymbol{p}$ needed to transform the dependent variable to be $p = 1 - slope = 1 - 0.9745824 = 0.02541763 \approx 0$, and hence a log-transformation.

Figure 7 examines the distributions of total pickup per acre before and after the log-transformation. As shown, the distribution looks more normally distributed after the transformation, which validates log-transformation. The right side of Figure 5 also visually presents the the effect of log-transformation. To statistically validate, another Breusch-Pagan test is implemented, and this time the p-value obtained is 0.2489394, leading to a rejection of null hypothesis and proving homoskedasticity.

Figure 6: Spread-level Plot for Global Model

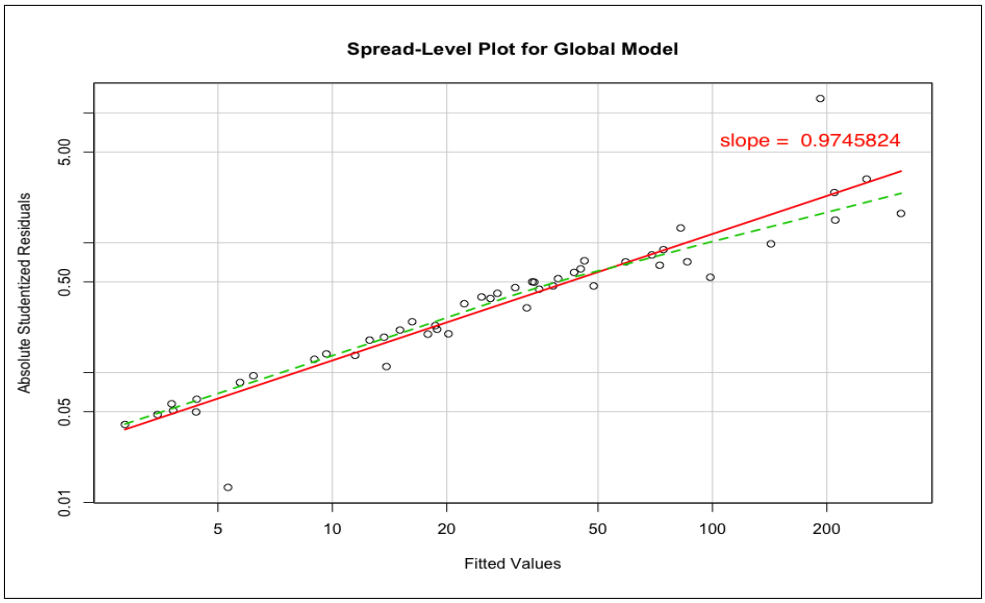
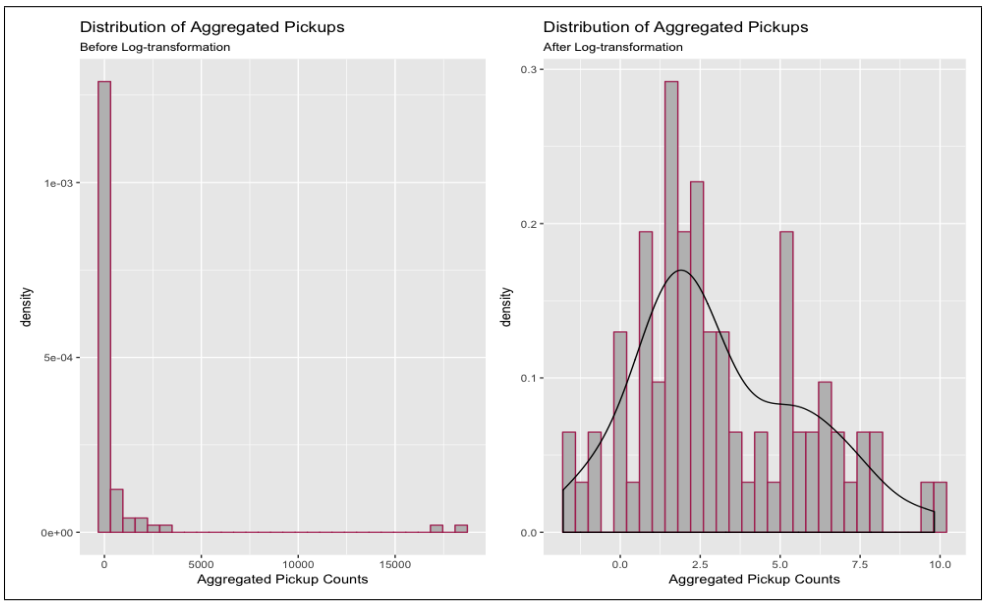


Figure 7: Histogram of Total Pickups

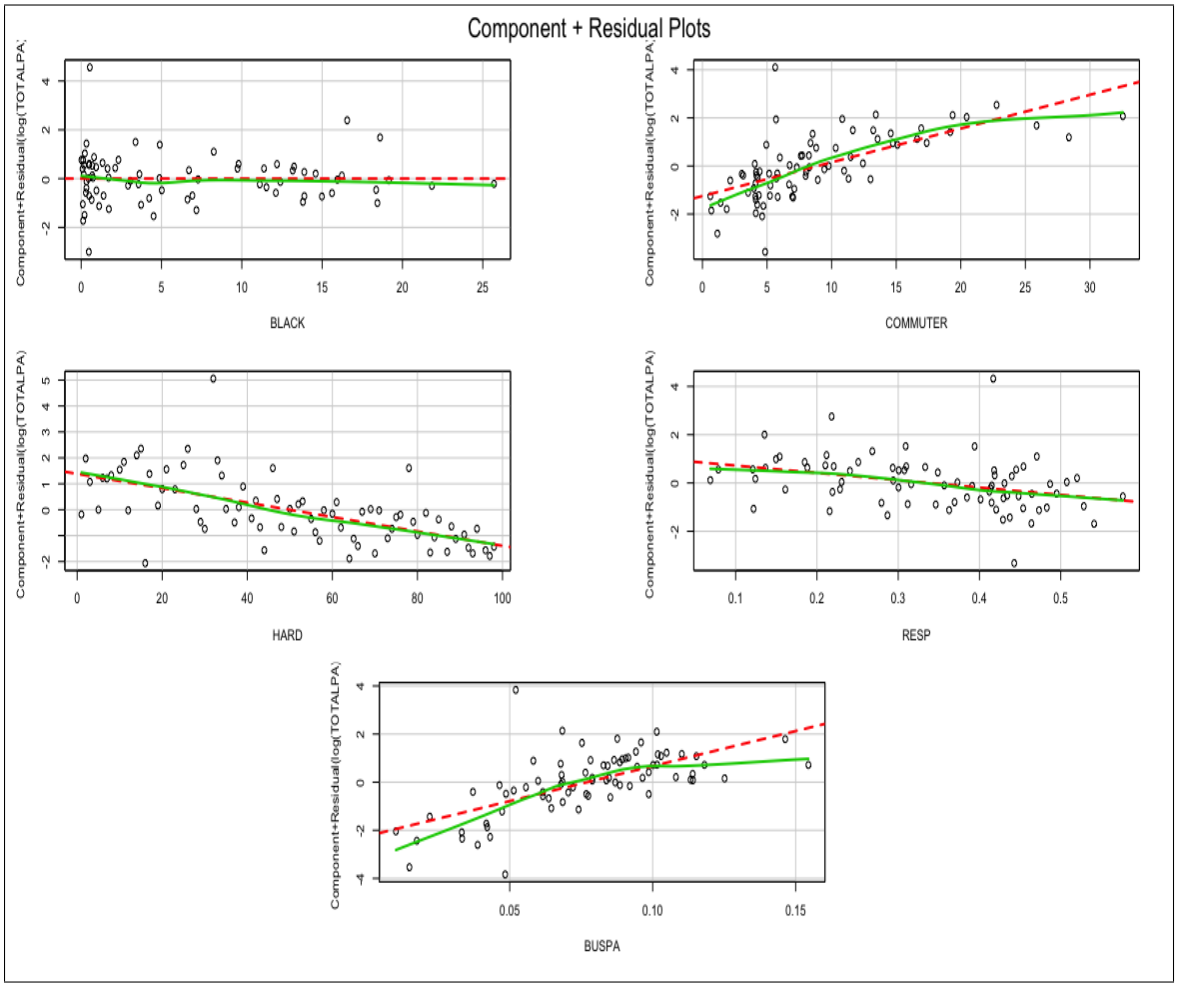


D.4 Detecting Non-linearity

According to Kabacoff (2015), evidence of nonlinearity in the relationship between the dependent variable and the independent variables could be detected by using partial residual plot by presenting $\epsilon_i + (\hat{\beta}_j \times X_{ji})$ on X_{ji} where the residuals ϵ are based on the full model, and i =1 ... n. The resulting partial residual plots for all independent variables except the percentage of park acres and the percentage of commercial acres are provided in Figure 8. As shown, each of the plot presented in Figure 8 shows smooth linear relationship.

The partial residual plots of variable, commercial acres percentage, **PARKP** and park acres percentage, **COMP**, are presented on the left side of Figure 9 and Figure 10 correspondingly. Both of them show nonlinearity. Kabacoff (2015) claims nonlinearity in partial residual plots suggests the variables per se may not have adequately modeled the functional form of that predictor in the regression.

Figure 8: Partial Residual Plot of Dependent Variables without **PARK**



To fix this nonlinearity, a log-transformation is implemented on both commercial acres percentage, **COMP**, and park acres percentage, **PARKP**, and the new partial

residual plots for log(**PARKP**) and log(**COMP**) are presented on the right side of Figure 9 and Figure 10. The linearities presented in the new partial residual plots proves the remedies of non-linearities issue.

Figure 9: Partial Residual Plot:
Log Transformation of PARKP

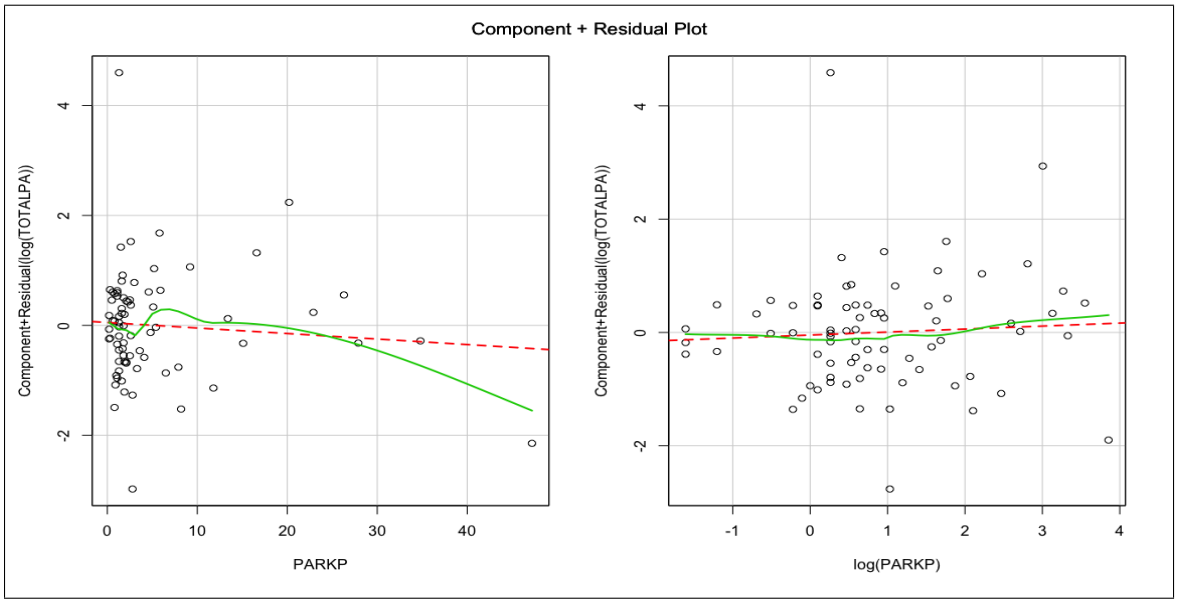
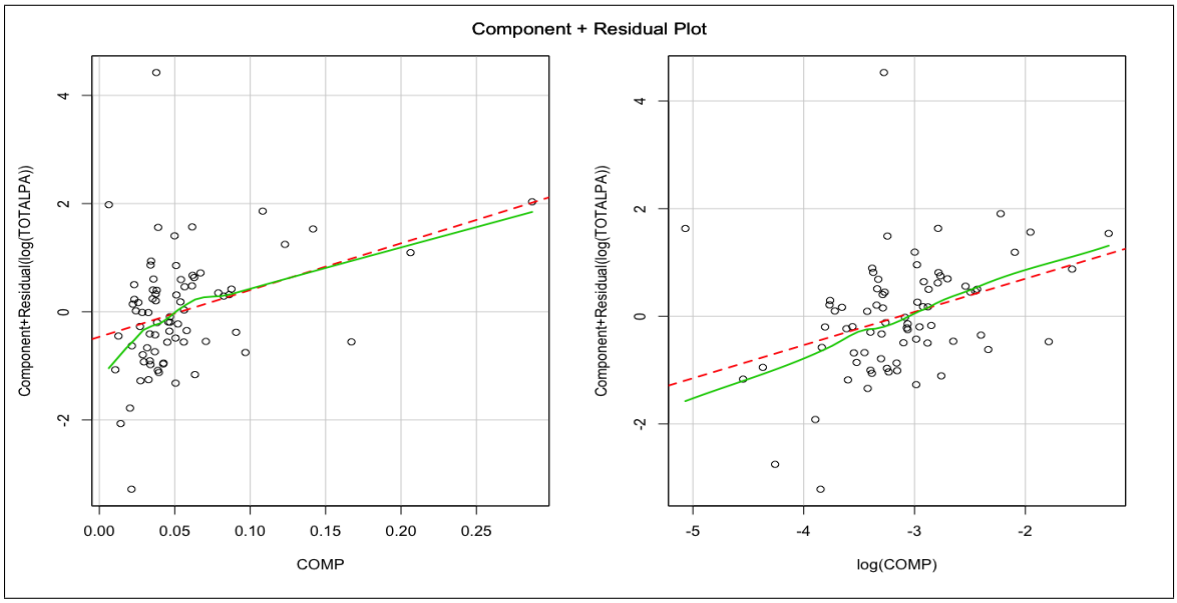


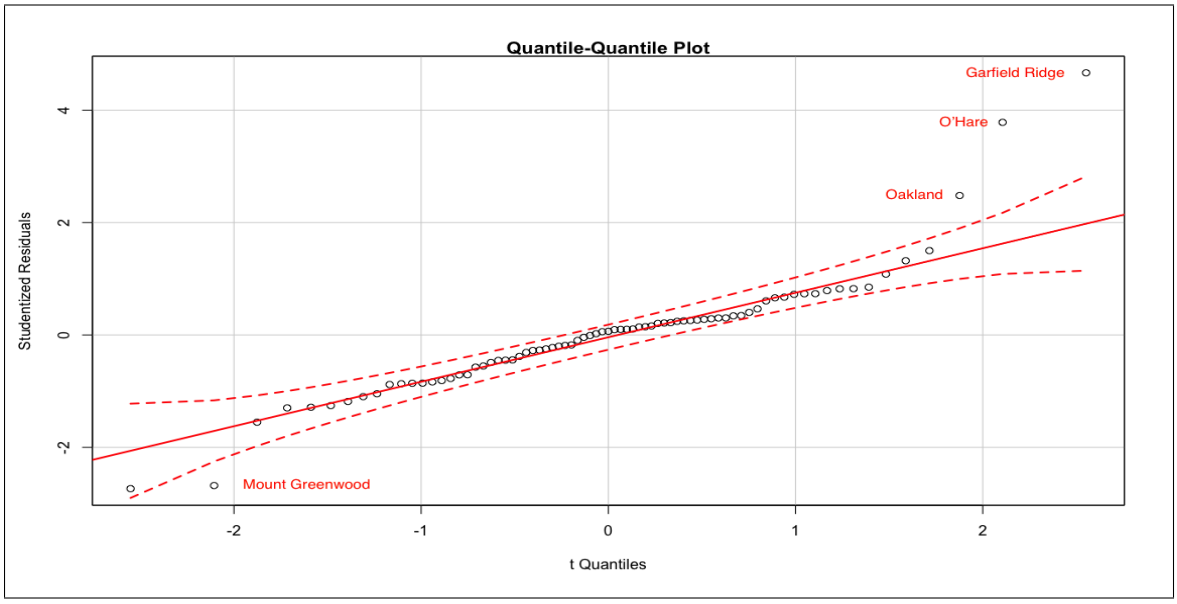
Figure 10: Partial Residual Plot:
Log Transformation of COMP



D.5 Detecting Non-normally Distributed Errors

After log-transforming the dependent variable, taxi pickups, and after log-transforming independent variables percentage of park acres and percentage of commercial acres, Figure 11 assesses whether errors are normally distributed. As advised by Kabacoff (2015), this figure has studentized residuals on the y-axis and a t-distribution with 66 degrees of freedom ($dof = n - p - 1 = 77 - 8 - 1 = 66$ where n is the sample size and p is the number of regression parameters including the intercept) on the y-axis.

Figure 11: Quantile-quantile Plot



As shown in Figure 11, there are four exception points, among which two of them are O'Hare and Garfield Ridge. Almost all other communities fall nicely inside of the 95% confidence envelope, suggesting the data is meeting the normality assumption fairly well. In addition, those two outliers make intuitive sense because O'Hare and Garfield Ridge are two communities having two airports – O'Hare International Airport (ORD) and Midway International Airport (MDW).

D.6 Detecting Independence of Errors

Kabacoff (2015) suggests the best way to assess whether the dependent variable values (and thus the residuals) are independent is from the knowledge of how the data were collected. For example, time series data will often display autocorrelation – observations collected closer in time will be more correlated with each other than with observations distant in time.

In order to detect autocorrelation, or to test the independence of errors assumption, Fox (2015) proposes a generalized Durbin-Watson test on the residuals of the global model with a null hypothesis of no autocorrelation while an alternative hypothesis of autocorrelation. The p-value obtained is 0.626, which is greater than 0.05

significance level. According to the p-value, null hypothesis fails to be rejected, and thus there is no autocorrelation and the independence of errors assumption is not violated.

D.7 Global Model Re-specification

Initially, the global model is indicated in Equation A.4.2 where i is the number of observations and $j = 1 \dots 10$, the number of independent variables.

$$\textbf{PICKUPS}_i = \beta_0 + \sum_{j=1}^{10} \beta_j X_{ij} + \epsilon_i \tag{A.4.2}$$

After detecting multicollinearity, two independent variables, **MINC** and **BS**, are excluded from the global model as in Equation A.4.3. This exclusions of independent variables lead to $j = 1 \dots 8$ while i still represents the number of observations.

$$\boldsymbol{PICKUPS}_i = \beta_0 + \sum_{j=1}^{8} \beta_j X_{ij} + \epsilon_i \tag{A.4.3}$$

After detecting heteroskedasticity, a log-transformation on the dependent variable is implemented, which is presented in Equation A.4.4. In Equation A.4.4, i represents the number of observations while $j = 1 \dots 8$, the number of independent variables.

$$log(\boldsymbol{PICKUPS}_i) = \beta_0 + \sum_{j=1}^{8} \beta_j X_{ij} + \epsilon_i \tag{A.4.4}$$

After detecting non-linearity of data, a log-transformation is required to independent variables **PARKP** and **COMP**, shown in Equation A.4.5. Similarly, i represents the number of observations while $j = 1 \dots 8$, the number of independent variables.

$$log(\boldsymbol{PICKUPS}_i) = \beta_0 + \sum_{j=1}^{6} \beta_j X_{ij} + \beta_7 log(\boldsymbol{PARKP})_i + \beta_8 log(\boldsymbol{COMP})_i + \epsilon_i \tag{A.4.5}$$

Those four models, Equation A.4.1, Equation A.4.3, Equation A.4.4, and Equation A.4.5, are presented in the first four columns of Table 19 correspondingly.

After scrutinizing the p-values of all coefficients in Equation A.4.5 in the fourth column of Table 19 below, it could be seen that **BLACK** and log(**COMP**) is not statistically significant. Thus, variable **BLACK** and log(**COMP**) are discarded, which makes the ultimate global model is defined to be Equation A.4.6, where i is the number of observations and $j = 1 \dots 7$, the number of independent variables. This model is presented in the last column of Table 19.

$$log(\boldsymbol{PICKUPS}_i) = \beta_0 + \sum_{j=1}^{6} \beta_j X_{ij} + \beta_7 log(\boldsymbol{COMP})_i + \epsilon_i \tag{A.4.6}$$

Table 19: Global Model Calibration

	Dependent variable:				
	TOTALPA			log(TOTALPA)	
	(1)	(2)	(3)	(4)	(5)
Description:	**Raw Global Model**	**Resolve Multicollinearity**	**Resolve Homoskedasticity**	**Resolve Non-linearity**	**Final Global Model**
BLACK	18.901	4.634	−0.003	0.003	
	(51.965)	(44.461)	(0.024)	(0.024)	
MINC	0.081***				
	(0.026)				
BS	306.716**				
	(127.917)				
COMMUTER	−108.021	101.753	0.145***	0.146***	0.142***
	(120.329)	(62.426)	(0.034)	(0.033)	(0.030)
HARD	44.039**	−29.092**	−0.025***	−0.027***	−0.028***
	(18.912)	(11.099)	(0.006)	(0.006)	(0.005)
RESP	−5,554.603*	−8,615.535***	−2.022	−3.416**	−3.620***
	(3,164.544)	(2,634.686)	(1.437)	(1.369)	(1.148)
COMP	16,347.980**	20,402.530***	6.812*		
	(6,351.733)	(6,780.837)	(3.699)		
PARKP	12.670	−3.176	0.032***		
	(18.041)	(19.921)	(0.011)		
log(COMP)				0.616**	0.592**
				(0.263)	(0.252)
log(PARKP)				0.053	
				(0.138)	
BUSPA	22,730.760*	21,605.270	30.252***	27.881***	28.269***
	(12,324.670)	(13,891.010)	(7.577)	(7.308)	(6.052)
LTRAIND	−991.296*	−1,163.819*	0.970***	0.739**	0.714**
	(548.256)	(613.321)	(0.335)	(0.333)	(0.324)
Constant	−6,368.315**	1,829.456	0.511	3.724***	3.851***
	(2,416.163)	(1,675.712)	(0.914)	(1.316)	(1.170)
R^2	0.648	0.539	0.840	0.851	0.851
Adjusted R^2	0.595	0.484	0.821	0.833	0.838
Residual Std. Error	1,815.922 (df = 66)	2,048.195 (df = 68)	1.117 (df = 68)	1.073 (df = 67)	1.058 (df = 69)
F Statistic	12.150*** (df = 10; 66)	9.923*** (df = 8; 68)	44.460*** (df = 8; 68)	47.871*** (df = 8; 67)	65.530*** (df = 6; 69)

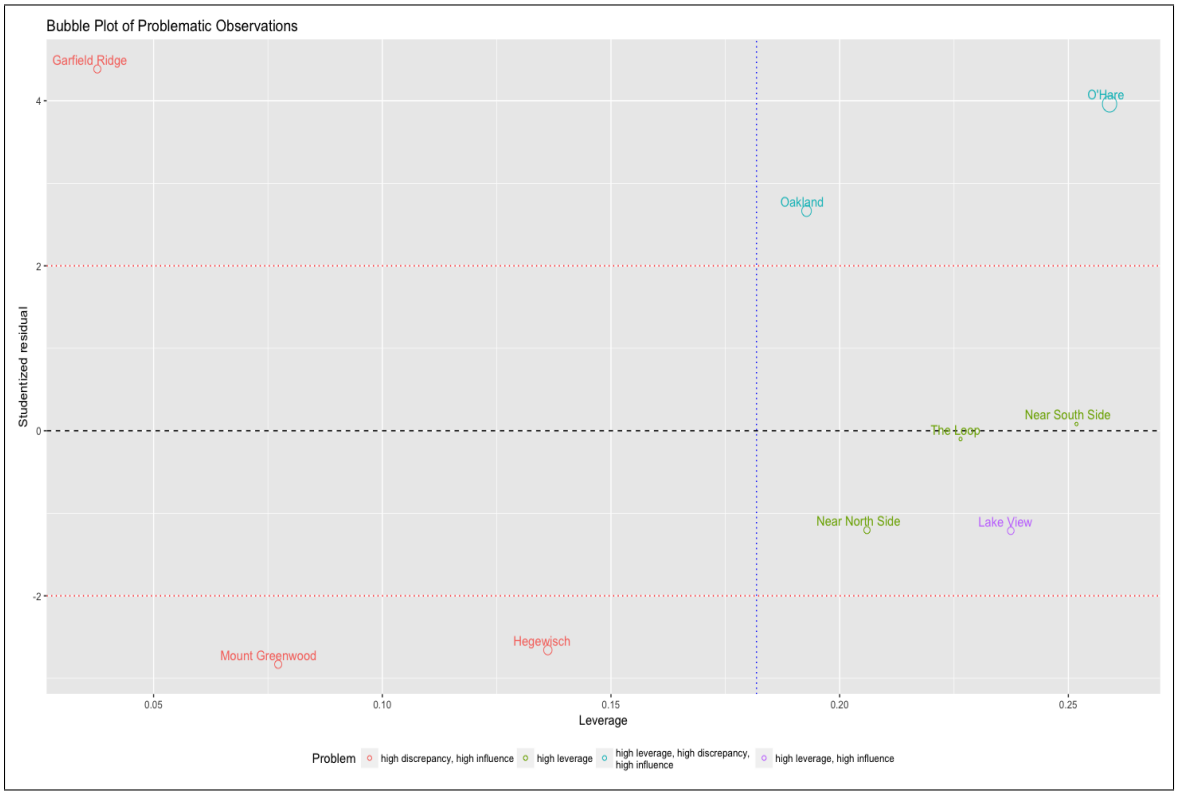
Note: #Observations = 77 for all 5 models presented above. *p<0.1; **p<0.05; ***p<0.01

E Influential Observations

Kabacoff (2015) defines influential observations to be observations that have a disproportionate impact on the values of the model parameters after examining the data. Soltoff (2017) uses three statistics to assess influential/problematic observations, and they are (1) Leverage (hat): degree of potential influence on the coefficient estimates that a given observation can (but not necessarily does) have; (2) Discrepancy: extent to which an observation is "unusual" or "different" from the rest of the data; (3) Influence: how much effect a particular observation's value(s) on Y and X have on the coefficient estimates (Influence = Leverage × Discrepancy).

Following three numerical rule of thumbs proposed by Soltoff (2017), problematic observations are (1) anything exceeding twice the average of leverage statistic; (2) anything outside of the range [-2,2] of discrepant statistic; (3) Cook's D greater than $\frac{4}{n-k-1}$ is influential where n is the number of observations and k is the number of coefficients in the regression model. Figure 12 shows all problematic points satisfy above three criteria.

Figure 12: Observations classified by 3 Statistics: Leverage, Discrepancy, and Influence



As shown in Figure 12, there are five outliers out of 77 observations (Garfield Ridge, O'Hare, Oakland, Hegewisch, and Mount Greenwood), which is approximately 9% of the total observations.

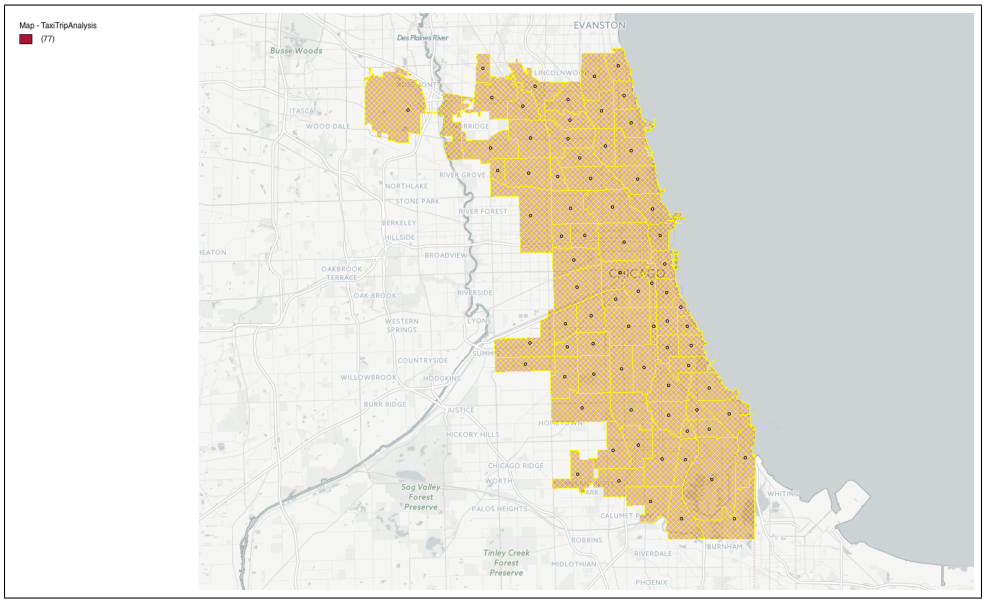
F Spatial Weights

According to Anselin and Rey (2014), spatial weights are a key component in cross-sectional analysis of spatial dependence. Spatial weights express the neighbor structure among the observations as a $n \times n$ matrix $\mathbf{W}$ in which the elements w_{ij} of the matrix are the spatial weights. In Matrix A.6.7, each spatial unit is represented in the matrix by a row i, and the potential neighbors by the columns j, with $j \neq i$. The spatial weights w_{ij} are non-zero when i and j are neighbors, and zero otherwise. Self-neighbor relation is excluded, so the diagonal elements of W are zero, $w_{ii} = 0$.

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{bmatrix} \tag{A.6.7}$$

There are commonly two types of spatial weights, and they are contiguity-based spatial weights and distance-based spatial weights. Contiguity means that two spatial units share a common border of non-zero length, so intrinsically, contiguity is most appropriate for geographic data expressed as polygons (so-called areal units), whereas distance-based spatial weights is suited for point data. However, according to Anselin and Rey (2014) the distinction is not that absolute in practice because polygon data can be represented by their centroid (central point) shown in Figure 13, while point data can be represented by a tessellation forming Thiessen polygons.

Figure 13: Polygon and its centroids



In this study, three different types of spatial weights are used for sensitivity check, among which two are contiguity-based and the others is distance-based weights with

arc-distance threshold value to be 4.61588 miles. Two different types of contiguity-based spatial weights are rook-contiguity and queen-contiguity. Anselin and Rey (2014) defines the rook criterion to be neighbors by the existence of a common edge between two spatial units while the queen criterion to be neighbors as spatial units sharing a common edge or a common vertex, so the queen criterion will always be at least as large as for the rook criterion. The comparison of neighbor distributions between queen and rook weights could be viewed from Figure 14, Figure 15, and Table 20, and some summary statistics of neighbors under all four types of weights could be seen from Table 21.

All three spatial weight matrices are row-standardized like Equation A.6.8, so each row sum of the row-standardized weights equals 1 and the sum of all weights $\sum_i \sum_j w_{ij} = n$.

$$w_{ij(s)} = w_{ij} / \sum_j w_{ij} \tag{A.6.8}$$

Table 20: Queen Weights vs. Rook Weights

Queen Contiguity-based Weights									
Number of Neighbors	1	2	3	4	5	6	7	8	9
#obs	1	3	9	18	13	16	12	4	1
% of total	1.299	3.896	11.688	23.377	16.883	20.779	15.584	5.195	1.299
sd from mean	-1.849	-1.251	-0.653	-0.054	0.000	0.544	1.142	1.740	2.339
Rook Contiguity-based Weights									
Number of Neighbors	1	2	3	4	5	6	7	8	9
#obs	1	4	11	20	19	16	4	1	1
% of total	1.299	5.194	14.286	25.974	24.675	20.779	5.195	1.299	1.299
sd from mean	-1.806	-1.124	-0.443	0.000	0.239	0.921	1.602	2.284	2.966

Table 21: Summary Statistics of All Four Spatial Weights

Summary Statistics of Neighbors under on 3 Types of Spatial Weights			
	Queen	**Rook**	**Distance**
Min	1	1	1
Max	9	9	26
Median	5	5	16
Mean	5.09	4.65	15.69
S.D.	1.67	1.47	5.17

Figure 14: Neighbor Distribution: Queen-contiguity Weights

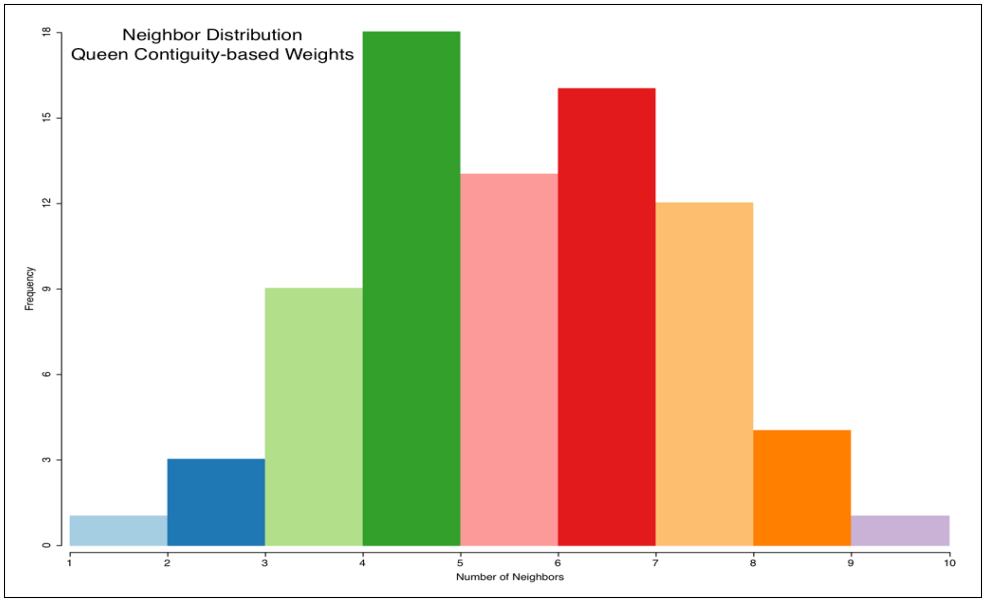
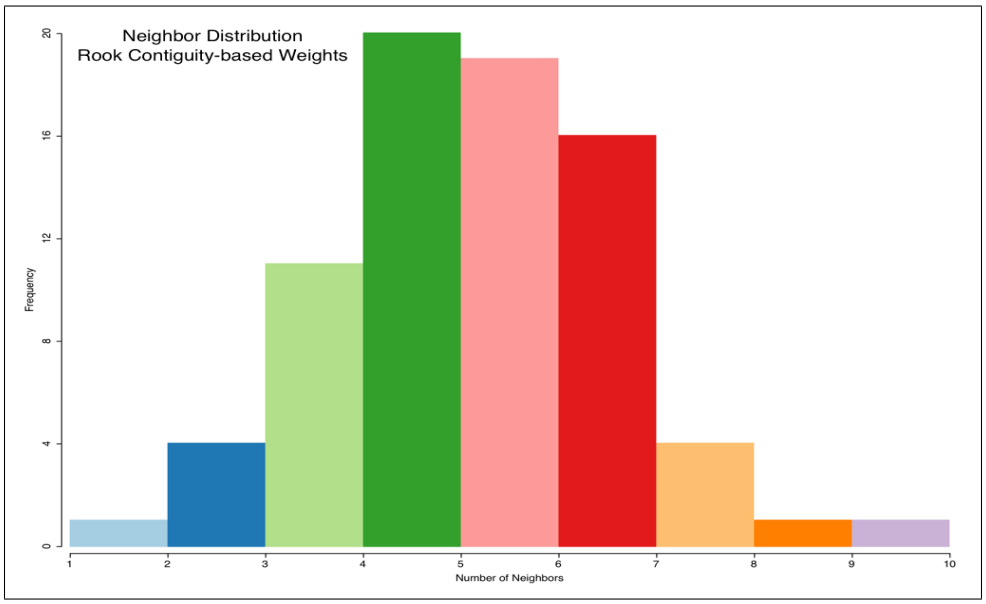


Figure 15: Neighbor Distribution: Rook-contiguity Weights



G Spatial Dependence

After solidifying assumptions for the global model presented in Equation A.7.9, an assessment of spatial autocorrelation using Moran's I on the residuals of the global model would be implemented to assess the presence of spatial dependence following methods derived by Anselin (1993). Then, the *Lagrange Multiplier* or *Rao Score* test statistics is used to discover the types of spatial dependence in the global model presented in Equation A.7.9, according to Anselin and Rey (2014).

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{A.7.9}$$

or

$$log(\boldsymbol{PICKUPS}_i) = \beta_0 + \sum_{j=1}^{6} \beta_j X_{ij} + \beta_7 log(\boldsymbol{COMP})_i + \epsilon_i$$

G.1 Detecting Spatial Autocorrelation

Anselin et al. (2013) defines spatial autocorrelation as an association between value similarity and spatial similarity, or the correlation of a variable with itself through space. As Anselin et al. (2013) stated, global spatial autocorrelation is a whole-map property and it says whether the spatial distribution of attribute values displays clustering or not. The most commonly used statistic to test global spatial autocorrelation is Moran's I. For a set of n spatial observations for a variable y, I is given as Equation A.7.10, where z_i is the deviation from the mean, or $y_i - y_{bar}$ and $w_{i,j}$ is the spatial weights.

$$I = \frac{n}{S_0} \times \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} z_i w_{i,j} z_j}{\sum_{i=1}^{n} z_i^2} \tag{A.7.10}$$

Using Moran's I test statistic stated in Equation A.7.10, an inference is conducted. The null hypothesis of randomly distributed residuals of global model in space is tested against the alternative hypothesis of not randomly distributed residuals. In addition, three different spatial weights were used to test sensitivity. The compiled results are presented in Table 22.

As shown, Moran's I statistics seems to be most sensitive to distance-based weights using centroids of the polygons (Refer to Appendix F). Interestingly, all three sets of statistics seem to exhibit an increasing pattern between 2013 to 2015 and decreases in 2016. Most importantly, almost all p-values are statistically significant at 5% significance level, thus it is safe to claim that the residuals of the global models exhibit spatial autocorrelation.

Table 22: Spatial Autocorrelation Diagnostics

Weights / Moran's I	Queen-contiguity		Rook-contiguity		Distance-based	
Residuals	I	**P-value**	I	**P-value**	I	**P-value**
log(PICKUP13PA)	2.3394	0.01932*	2.4121	0.01586**	3.8482	0.00012***
log(PICKUP14PA)	3.1451	0.00166**	3.2968	0.00098***	4.1742	0.00003***
log(PICKUP15PA)	3.1466	0.00165**	3.2083	0.00134**	4.7330	0.00000***
log(PICKUP16PA)	2.7750	0.00552**	2.9857	0.00283**	3.6245	0.00029***
log(TOTALPA)	2.8637	0.00419**	2.9870	0.00282**	4.1178	0.00004***

Note: *p<0.05; **p<0.01; ***p<0.001

G.2 Detecting Spatial Lag

Anselin and Rey (2014) defines a spatial lag model to be Equation A.7.11, where $\boldsymbol{W}\,log(\boldsymbol{PICKUPS}_i)$ is the spatially lagged dependent variable with associated autoregressive coefficient ρ and spatial weights $\boldsymbol{W}$. The rest of the notation is as the global model in A.7.9.

$$\boldsymbol{Y} = \rho \boldsymbol{W}\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{A.7.11}$$

or

$$log(\boldsymbol{PICKUPS}_i) = \rho \boldsymbol{W}\, log(\boldsymbol{PICKUPS}_i) + \beta_0 + \sum_{j=1}^{5} \beta_j X_{ij} + \beta_6 log(\boldsymbol{COMP})_i + \epsilon_i$$

where

- ρ is the autoregressive coefficient
- $\boldsymbol{W}$ is the spatial weighting matrix
- $\boldsymbol{W}\,log(\boldsymbol{PICKUPS}_i)$ is the spatial lag

A *Lagrange Multiplier* test is used to assess the null hypothesis of $\rho = 0$ and the alternative hypothesis of $\rho \neq 0$. Anselin (1988b) defines the *Lagrange Multiplier* test statistics for testing spatial lag to be Equation A.7.12. The statistic is distributed as a χ^2 with one degree of freedom.

$$LM_\rho = \frac{d_\rho^2}{D} \sim \chi^2(1) \tag{A.7.12}$$

where the numerator is Equation A.7.13 where $\boldsymbol{e}$ is a vector of OLS residual, $\boldsymbol{WY}$ is the spatial lag term, and $\hat{\sigma}^2_{\text{ML}} = \boldsymbol{e'e}/n$.

$$d_\rho = \frac{\boldsymbol{e'WY}}{\hat{\sigma}^2_{\text{ML}}} \tag{A.7.13}$$

and the denominator is Equation A.7.14. The denominator term D consists of two parts. The first part contains the sum of squared residuals in a regression of the spatially lagged predicted values $\boldsymbol{WX\hat{\beta}}$ on the $\boldsymbol{X}$. The second term in D is a trace expression.

$$D = (\boldsymbol{WX\hat{\beta}})'[\boldsymbol{I} - \boldsymbol{X(X'X)^{-1}X'}](\boldsymbol{WX\hat{\beta}})/\hat{\sigma}^2_{\text{ML}} + T \tag{A.7.14}$$

$$\text{where } T = tr(\boldsymbol{WW} + \boldsymbol{W'W}) \tag{A.7.15}$$

G.3 Detecting Spatial Error

According to Anselin and Rey (2014), the null hypothesis for the spatial error test is also a standard linear regression specification, as in Equation A.7.9. The specific alternative hypothesis is a regression model that includes a spatial autoregressive or a spatial moving average error term.

The alternative hypothesis in spatial autoregressive form is presented in Equation A.7.16.

$$\boldsymbol{u} = \lambda \boldsymbol{W\mu} + \boldsymbol{v} \tag{A.7.16}$$

The alternative hypothesis in spatial moving average form is presented in Equation A.7.17.

$$\boldsymbol{u} = \lambda \boldsymbol{Wv} + \boldsymbol{v} \tag{A.7.17}$$

According to Anselin and Rey (2014), in both specifications, $\boldsymbol{v}$ is a vector of idiosyncratic error terms and $\boldsymbol{W\mu}$ and $\boldsymbol{Wv}$ are spatially lagged terms, respectively for the original regression error term (a spatial autoregression) or for the idiosyncratic term (a spatial moving average). The associated coefficient is λ.

A *Lagrange Multiplier* test is used to assess the null hypothesis of $\lambda = 0$ and the alternative hypothesis of $\lambda \neq 0$. Burridge (1980) defines the *Lagrange Multiplier* or Rao Score test statistics for spatial error autocorrelation to be Equation A.7.18 where $\boldsymbol{e}$ is a vector of OLS residuals, $\boldsymbol{We}$ is its spatial lag, and $\hat{\sigma}^2_{\text{ML}}$ is as before.The computed statistic is distributed as a χ^2 with one degree of freedom.

$$LM_\lambda = \frac{d^2_\lambda}{tr(\boldsymbol{WW} + \boldsymbol{W'W})} \sim \chi^2(1) \tag{A.7.18}$$

where

$$d_\lambda = \frac{\boldsymbol{e'We}}{\hat{\sigma}^2_{\text{ML}}} \tag{A.7.19}$$

As suggested by Anselin and Rey (2014), further technical details can be found in Anselin (1988c), Anselin and Rey (1991), Anselin and Bera (1998), and Anselin (2001).

G.4 Robust Testing of Spatial Dependence

As stated in Section G.2 and Section G.3 of this appendix, two spatial dependence test statistics using *Lagrange Multiplier* test LM_ρ from Equation A.7.12 and LM_λ

from Equation A.7.18 are used to assess the spatial dependence of the residuals of the global model. **?** introduced a robustified form of the test statistics for both the spatial lag LM_ρ and spatial error LM_λ, and they are LM_ρ^* and LM_λ^*, which are presented in Equation A.7.20 and A.7.21.

$$LM_\rho^* = \frac{(d_\rho - d_\lambda)^2}{D - T} \sim \chi^2(1) \tag{A.7.20}$$

$$LM_\lambda^* = \frac{(d_\lambda - TD^{-1}d_\rho)^2}{[T(1-T)]} \sim \chi^2(1) \tag{A.7.21}$$

where d_ρ has referenced in Equation A.7.13, d_λ has referenced in Equation A.7.19, D has referenced in Equation A.7.15, and T has referenced in Equation A.7.15.

According to Anselin and Rey (2014) and Anselin (1996), the purpose of using robustified form of the test statistics, LM_ρ^* and LM_λ^*, is because the LM_ρ statistic for testing spatial lag model is also sensitive to the presence of spatial error autocorrelation making the result of LM_ρ misleading. Similarly, the LM_λ statistic for spatial error has power against a spatial lag alternative and could also suggest the wrong alternative.

G.5 Diagnostics on Spatial Dependence

Lagrange Multiplier tests are implemented on the global model and three different spatial weights are used to test sensitivity.

Table 23: Spatial Dependence Diagnostics

Weights / Lagrange		Queen-contiguity		Rook-contiguity		Distance-based	
Description	**Test**	χ^2	**P-value**	χ^2	**P-value**	χ^2	**P-value**
Spatial Lag	LM_ρ	7.7862	0.00526**	8.1787	0.00424**	7.4904	0.00620**
Robust Spatial Lag	LM_ρ^*	6.4823	0.01090*	5.8552	0.01553*	6.4515	0.01109*
Spatial Error	LM_λ	1.9637	0.16111	2.7183	0.09920	1.2543	0.26274
Robust Spatial Error	LM_λ^*	0.6598	0.41662	0.3948	0.52980	0.2154	0.64260

Note: *p<0.05; **p<0.01

The results of *Lagrange Multiplier* test statistics are statistically consistent as presented in Table 23. LM_ρ are all statistically significant at 5% significance level for a χ^2 variate with one degree of freedom, strongly indicating the presence of spatial autocorrelation. The robustified statistics, LM_ρ^*, are also statistically significant at 5% significance level for all three tests, which confirms the lag specification.

H Compiling Models by Year

Table 24, Table 25, Table 26, Table 27, and Table 28 are regression tables containing Global Models (OLS) regressing annual pickup density from 2013 to 2016 on the same set of independent variables. In addition to OLS, three spatial autoregressive models are also listed using three different spatial weights to test sensitivity. On the bottom of each table, there are numerous criterion statistics to determine model performances.

Table 24: Model Compiling for 2013 Pickups

	Dependent variable:			
	2013 Pickups per acre			
	Global Model	*Spatial Autoregressive Model*		
	OLS	*SAR*		
Weights	(NA)	(Queen)	(Rook)	(Distance)
ρ (spatial lag)		0.452^{***}	0.518^{***}	0.537^{***}
		(0.092)	(0.0856)	(0.087)
COMMUTER	0.146^{***}	0.084^{***}	0.082^{***}	0.0925^{***}
	(0.033)	(0.031)	(0.030)	(0.094)
HARD	-0.031^{***}	-0.023^{***}	-0.022^{***}	-0.030^{***}
	(0.006)	(0.005)	(0.005)	(0.005)
RESP	-4.608^{***}	-3.171^{***}	-3.098^{***}	-4.286^{***}
	(1.218)	(1.058)	(1.055)	(1.000)
log(COMP)	0.527^{*}	0.447^{*}	0.460^{**}	0.483^{**}
	(0.276)	(0.229)	(0.228)	(0.226)
BUSPA	21.270^{***}	11.052^{**}	10.766^{**}	14.388^{***}
	(6.364)	(5.458)	(5.424)	(5.287)
LTRAIND	1.084^{***}	0.662^{**}	0.685^{**}	0.464
	(0.348)	(0.305)	(0.302)	(0.312)
Constant	3.121^{**}	2.732^{**}	2.735^{***}	3.337^{***}
	(1.271)	(1.061)	(1.057)	(1.04)
R^2	0.826	0.868	0.870	0.871
Log Likelihood	-117.284	−108.292	−108.003	−106.699
σ^2	1.35497	0.930	0.919	0.912238
Akaike Inf. Crit.	248.569	232.622	232.006	229.399
Schwarz criterion	264.975	251.372	250.756	248.149
Residual Std. Error	1.16403	0.964638	0.958778	0.955111

*p<0.1; **p<0.05; ***p<0.01

Table 25: Model Compiling for 2014 Pickups

	Dependent variable:			
	2014 Pickups per acre			
	Global Model	*Spatial Autoregressive Model*		
	OLS	*SAR*		
Weights	(NA)	(Queen)	(Rook)	(Distance)
ρ (spatial lag)		0.452^{***}	0.518^{***}	0.529^{***}
		(0.092)	(0.0856)	(0.099)
COMMUTER	0.153^{***}	0.079^{***}	0.075^{**}	0.079^{**}
	(0.034)	(0.030)	(0.030)	(0.031)
HARD	-0.032^{***}	-0.023^{***}	-0.022^{***}	-0.031^{***}
	(0.006)	(0.005)	(0.005)	(0.005)
RESP	-4.874^{***}	-3.251^{***}	-3.116^{***}	-4.412^{***}
	(1.241)	(1.054)	(1.039)	(1.026)
log(COMP)	0.519^{*}	0.359	0.368	0.409^{*}
	(0.281)	(0.228)	(0.224)	(0.224)
BUSPA	20.852^{***}	11.553^{**}	11.004^{**}	11.004^{**}
	(6.480)	(5.437)	(5.342)	(5.428)
LTRAIND	0.983^{***}	0.534^{*}	0.546^{*}	0.546^{*}
	(0.354)	(0.302)	(0.296)	(0.318)
Constant	3.392^{**}	2.548^{**}	2.517^{**}	3.198^{***}
	(1.294)	(1.056)	(1.040)	(1.069)
R^2	0.813262	0.862084	0.866726	0.856449
Log Likelihood	-117.768	−108.165	−107.18	−108.586
σ^2	1.37211	0.921251	0.890243	0.958896
Akaike Inf. Crit.	249.537	232.329	230.36	233.173
Schwarz criterion	265.943	251.08	249.11	251.923
Residual Std. Error	1.17137	0.959818	0.943527	0.979232

*p<0.1; **p<0.05; ***p<0.01

Table 26: Model Compiling for 2015 Pickups

	Dependent variable:			
	2015 Pickups per acre			
	Global Model	*Spatial Autoregressive Model*		
	OLS	*SAR*		
Weights	(NA)	(Queen)	(Rook)	(Distance)
ρ (spatial lag)		0.452^{***}	0.518^{***}	0.556^{***}
		(0.092)	(0.0856)	(0.093)
COMMUTER	0.141^{***}	0.083^{***}	0.081^{***}	0.085^{***}
	(0.033)	(0.030)	(0.030)	(0.0305)
HARD	-0.032^{***}	-0.023^{***}	-0.022^{***}	-0.031^{***}
	(0.006)	(0.005)	(0.005)	(0.005)
RESP	-4.781^{***}	-3.276^{***}	-3.191^{***}	-4.544^{***}
	(1.226)	(1.039)	(1.033)	(1.005)
log(COMP)	0.466^{*}	0.430^{*}	0.445^{**}	0.465^{**}
	(0.278)	(0.225)	(0.224)	(0.228)
BUSPA	21.351^{***}	9.650^{*}	9.307^{*}	12.686^{**}
	(6.404)	(5.346)	(5.296)	(5.310)
LTRAIND	0.943^{***}	0.524^{*}	0.548^{*}	0.336
	(0.350)	(0.299)	(0.295)	(0.313)
Constant	3.196^{**}	2.866^{***}	2.867^{***}	3.501^{***}
	(1.279)	(1.044)	(1.037)	(1.049)
R^2	0.821645	0.874534	0.876862	0.871246
Log Likelihood	-118.679	−107.42	−106.921	−107.197
σ^2	1.40495	0.898486	0.881812	0.922032
Akaike Inf. Crit.	251.358	230.839	229.843	230.395
Schwarz criterion	267.765	249.59	248.593	249.145
Residual Std. Error	1.18531	0.947885	0.939049	0.960225

*p<0.1; **p<0.05; ***p<0.01

Table 27: Model Compiling for 2016 Pickups

	Dependent variable:			
	2016 Pickups per acre			
	Global Model	*Spatial Autoregressive Model*		
	OLS	*SAR*		
Weights	(NA)	(Queen)	(Rook)	(Distance)
ρ (spatial lag)		0.452^{***}	0.518^{***}	0.513^{***}
		(0.092)	(0.0856)	(0.105)
COMMUTER	0.137^{***}	0.082^{***}	0.079^{**}	0.0790^{**}
	(0.034)	(0.032)	(0.031)	(0.032)
HARD	-0.032^{***}	-0.024^{***}	-0.023^{***}	-0.031^{***}
	(0.006)	(0.005)	(0.005)	(0.005)
RESP	-5.207^{***}	-3.758^{***}	-3.626^{***}	-4.781^{***}
	(1.238)	(1.111)	(1.097)	(1.056)
log(COMP)	0.508^{*}	0.415^{*}	0.423^{*}	0.454^{*}
	(0.281)	(0.238)	(0.235)	(0.239)
BUSPA	21.123^{***}	11.960^{**}	11.443^{**}	14.4097^{**}
	(6.469)	(5.692)	(5.604)	(5.595)
LTRAIND	0.903^{**}	0.547^{*}	0.557^{*}	0.357
	(0.354)	(0.314)	(0.309)	(0.327)
Constant	3.151^{**}	2.721^{**}	2.692^{**}	3.312^{***}
	(1.291)	(1.100)	(1.086)	(1.100)
R^2	0.808471	0.848962	0.853404	0.847443
Log Likelihood	-118.545	−111.067	−110.204	−110.664
σ^2	1.40008	1.00372	0.974197	1.01381
Akaike Inf. Crit.	251.09	238.133	236.408	237.328
Schwarz criterion	267.497	256.884	255.159	256.078
Residual Std. Error	1.12818	1.00186	0.987014	1.00688

*p<0.1; **p<0.05; ***p<0.01

Table 28: Model Compiling for Aggregated Pickups

	Dependent variable:			
	Aggregated Total Pickups per acre			
	Global Model	*Spatial Autoregressive Model*		
	OLS	*SAR*		
Weights	(NA)	(Queen)	(Rook)	(Distance)
ρ (spatial lag)		0.452^{***}	0.518^{***}	0.535^{***}
		(0.092)	(0.0856)	(0.097)
COMMUTER	0.144^{***}	0.081^{***}	0.078^{***}	0.081^{***}
	(0.033)	(0.030)	(0.030)	(0.031)
HARD	-0.032^{***}	-0.024^{***}	-0.022^{***}	-0.031^{***}
	(0.006)	(0.005)	(0.005)	(0.005)
RESP	-4.821^{***}	-3.307^{***}	-3.200^{***}	-4.461^{***}
	(1.218)	(1.051)	(1.041)	(1.00)
log(COMP)	0.498^{*}	0.404^{*}	0.416^{*}	0.446^{*}
	(0.276)	(0.227)	(0.225)	(0.228)
BUSPA	21.234^{***}	11.151^{**}	10.714^{**}	13.793^{***}
	(6.364)	(5.414)	(5.345)	(5.339)
LTRAIND	0.982^{***}	0.567^{*}	0.585^{**}	0.384
	(0.348)	(0.301)	(0.297)	(0.314)
Constant	4.593^{***}	3.423^{***}	3.397^{***}	3.963^{***}
	(1.271)	(1.075)	(1.066)	(1.056)
R^2	0.820603	0.866891	0.870133	0.864794
Log Likelihood	-117.293	−107.788	−107.095	−107.374
σ^2	1.35528	0.914171	0.891908	0.928578
Akaike Inf. Crit.	248.586	231.576	230.19	230.748
Schwarz criterion	264.993	250.326	248.941	249.499
Residual Std. Error	1.16417	0.956123	0.944409	0.963628

*p<0.1; **p<0.05; ***p<0.01