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

Dongping Zhang | <dpzhang@uchicago.edu>
M.A. in Computational Social Science | The University of Chicago

Background

- **Spatial Autocorrelation**: the correlation of a variable with itself through space, which violates OLS assumption of independence of observations.

Data & Methods

- Taxi ridership is aggregated by 77 Chicago communities and is analyzed by relating it to three categories of community-level variables.
- A spatial autoregressive model is implemented to account for the spatial dependence of taxi ridership.
- Three different spatial weights are used to test sensitivity, and they are queen-contiguity, rook-contiguity, and distance-based weights.

Global Model

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

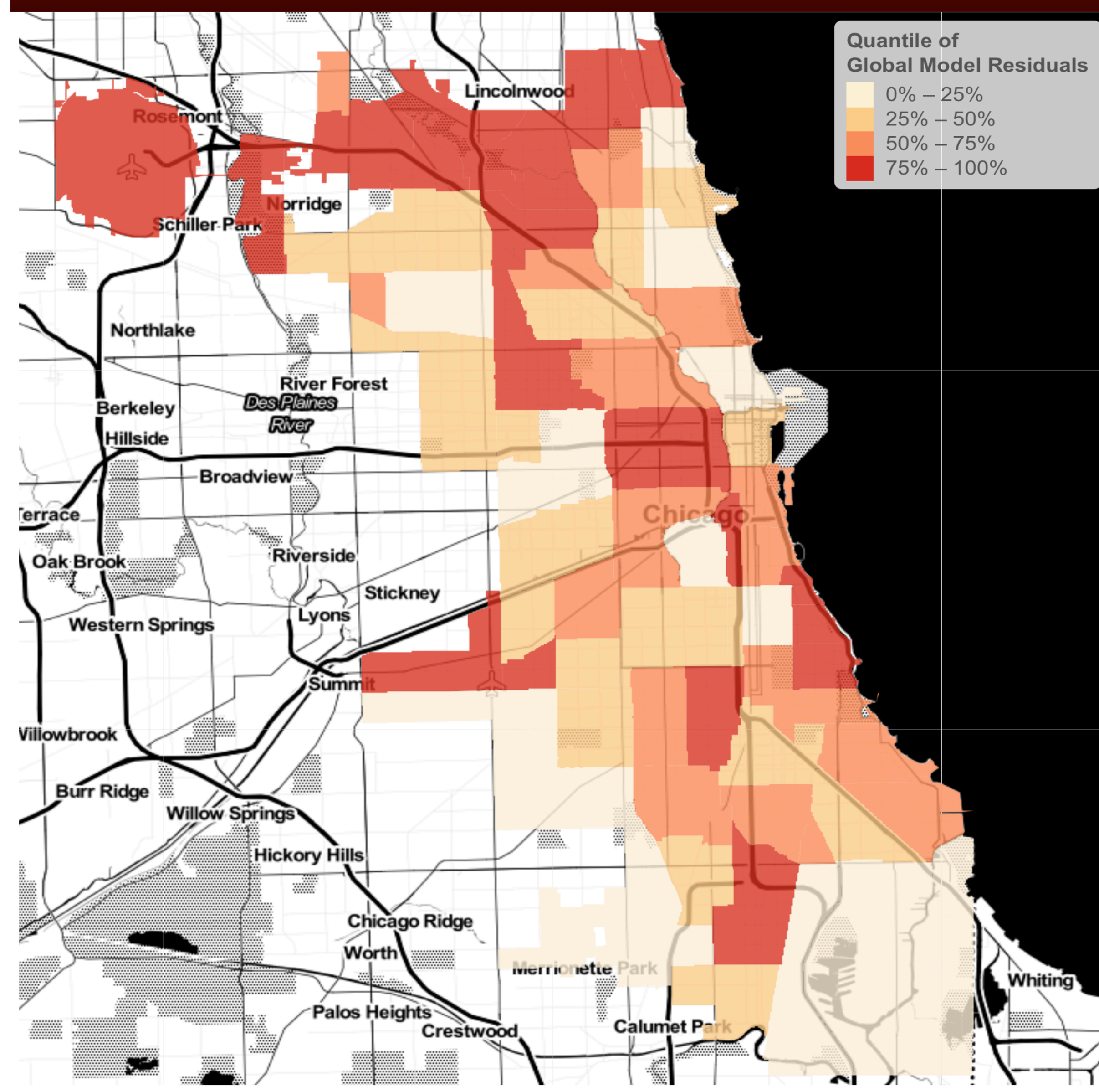
List of Dependent Variables

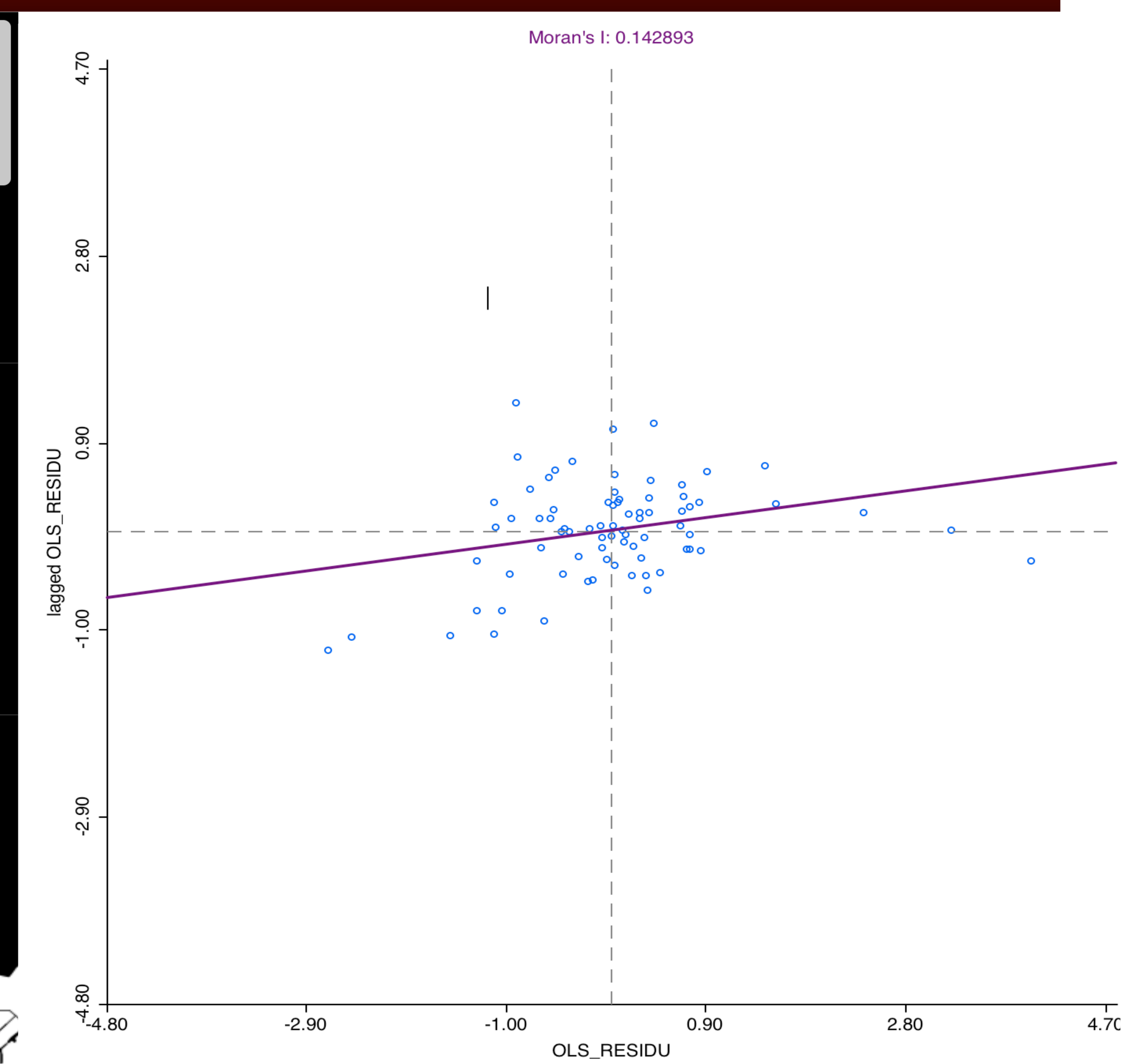
Variable	Definition
PICKUPS	aggregated pickup density

List of Independent Variables

Variable	Definition
Category 1: Socio-demographic	
COMMUTER	density of commuters
HARD	the hardship index
Category 2: Built-environment	
RESP	% residential land
COMP	% commercial land
Category 3: Urban-transportation	
BUSPA	bus stops density
LTRAIND	dummy L'Train stations

Spatial Autocorrelation of Pickups per acre by Moran's I





Lagrange Multiplier Test for Spatial Dependence

Lagrange Description (Weights)	Test	Queen-contiguity χ^2	Queen-contiguity P-value	Rook-contiguity χ^2	Rook-contiguity P-value	Distance-based χ^2	Distance-based 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

Spatial Autoregressive Model (Lag Model)

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

or

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

where ρ is the autoregressive coefficient

$\boldsymbol{W}$ is the spatial weighting matrix

and $\boldsymbol{W}$ $log(\boldsymbol{PICKUPS}_i)$ is the spatially lagged dependent

Results

	Dependent variable: log(TOTALPA)	
	Global Model (OLS)	*Spatial Autoregressive (SAR)*
ρ (spatial lag)		0.470*** (0.092)
COMMUTER	0.144*** (0.033)	0.081*** (0.030)
HARD	−0.032*** (0.006)	−0.024*** (0.005)
RESP	−4.821*** (1.218)	−3.307*** (1.051)
log(COMP)	0.498* (0.276)	0.404* (0.227)
BUSPA	21.234*** (6.364)	11.151** (5.414)
LTRAIND	0.982*** (0.348)	0.567* (0.301)
Constant	4.593*** (1.271)	3.423*** (1.075)
R^2	0.82	0.87
σ^2	1.35528	0.914
Akaike Inf. Crit.	248.586	231.576
Residual Std. Error	1.164 (df = 70)	0.956 (df = 69)
F Statistic	53.366*** (df = 6; 70)	

Note: #obs. = 77 for both models — *p<0.1; **p<0.05; ***p<0.01

Conclusion

- Global Moran's I and the residuals of global model confirms spatial autocorrelation.
- A spatial autoregressive model greatly outperforms global OLS model in modelling community taxi demand.

Special Thanks

I would like to express my sincerest gratitude to Dr. Luc Anselin, Dr. Rick Evans, Dr. Ben Soltaff, and Ms. Ging Cee Ng for your teaching, guidance, and support throughout the quarter.