SPATIAL ECONOMETRICS

Zhaoqiu Luo Sayli Javadekar

Overview

I. Introduction
II. Data description
III. Exploratory Analysis
IV. Modeling
V. Prediction
VI. Conclusion

Objective

- Find the determinants of home to work commuting in Toulouse
- Spatial data analysis
- Fitting models
  - Simple OLS model without adjustment
  - OLS model with adjustment
  - Lag model with adjustment
  - Durbin model with adjustment
- Prediction

Data Description

Original Data Set

- **Xfile :** Explanatory Variables
- **District :** (.shp) Geographic variables
- **Flux :** (.txt) Target Variable
    - 0's added to missing flows

Data we created:

- **Xo** : origin characteristics data (3600*3600)
- **Xd** : destination characteristics data (3600*3600)
- **Xo_intra , Xd_intra** : internal flows for origin and destination
- **Xo_inter , Xd_inter**: interregional flows for origin and destination

Data Analysis

- Xo : explanatory characteristics in origin

  $\underbrace{(1,\ldots,1)}_{\text{60 times}},(2,\ldots,2)\ldots\ldots(60,\ldots60)$

	name_district	ID_district	labour_force	activity_rate	employment	unemployment_rate	housing_units	origin	destination	flow
1	CAPITOLE	1	2364	43.10722101	1928	18.44331641	4588	1	1	336
2	CAPITOLE	1	2364	43.10722101	1928	18.44331641	4588	1	2	72
3	CAPITOLE	1	2364	43.10722101	1928	18.44331641	4588	1	3	108
4	CAPITOLE	1	2364	43.10722101	1928	18.44331641	4588	1	4	56
5	CAPITOLE	1	2364	43.10722101	1928	18.44331641	4588	1	5	40
6	CAPITOLE	1	2364	43.10722101	1928	18.44331641	4588	1	6	16
7	CAPITOLE	1	2364	43.10722101	1928	18.44331641	4588	1	7	12
8	CAPITOLE	1	2364	43.10722101	1928	18.44331641	4588	1	8	60
9	CAPITOLE	1	2364	43.10722101	1928	18.44331641	4588	1	9	20
10	CAPITOLE	1	2364	43.10722101	1928	18.44331641	4588	1	10	32
11	CAPITOLE	1	2364	43.10722101	1928	18.44331641	4588	1	11	52
12	CAPITOLE	1	2364	43.10722101	1928	18.44331641	4588	1	12	8

Data Analysis

- Xd : explanatory characteristics in destination

$$\underbrace{(1,2,\ldots,60),(1,2,\ldots,60)\ldots\ldots(1,2,\ldots60)}_{\text{60 times}}$$

	name_district	ID_district	labour_force	activity_rate	employment	unemployment_rate	housing_units	origin	destination	flow
1	CAPITOLE	1	2364	43.10722101	1928	18.44331641	4588	1	1	336
2	ARNAUD-BERNARD	2	2936	37.97206415	2328	20.70844687	5664	1	2	72
3	SAINT-GEORGES	3	1464	41.30925508	1256	13.93442623	2964	1	3	108
4	SAINT-ETIENNE	4	2054	49.24478542	1713	16.21226874	2978	1	4	56
5	CARMES	5	2504	50.08	2126	14.89616613	3976	1	5	40
6	SAINT-CYPRIEN	6	1924	50.10416667	1536	19.75051975	2940	1	6	16
7	AMIDONNIERS	7	2288	41.66059723	1984	13.11188811	3612	1	7	12
8	COMPANS	8	2940	48.0706344	2452	16.05442177	4884	1	8	60
9	LES-CHALETS	9	3560	51.47484095	2896	18.08988764	5444	1	9	20
10	MATABIAU	10	3892	52.88043478	3152	18.80781089	6376	1	10	32
11	SAINT-AUBIN-DUPUY	11	4052	52.40558717	3236	19.84205331	6436	1	11	52
12	LE-BUSCA	12	4076	52.68872802	3472	14.62217861	5632	1	12	8

Data Analysis

- Characteristics of Internal Flows: Xo_intra, Xd_intra
- Characteristics of Interregional Flows: Xo_inter, Xd_inter

$$
\begin{array}{c}
1\rightarrow 1\\
1\rightarrow 2\\
\vdots\\
2\rightarrow 1\\
2\rightarrow 2\\
2\rightarrow 3\\
\vdots\\
\vdots\\
60\rightarrow 60
\end{array}
\quad
\overset{X_{o/d\_}\text{ intra}}{\begin{bmatrix} x_{1,1}\\ 0\\ \vdots\\ 0\\ x_{2,2}\\ 0\\ \vdots\\ \vdots\\ x_{60,60}\end{bmatrix}}
\quad
\overset{X_{o/d\_}\text{ inter}}{\begin{bmatrix} 0\\ x_{1,2}\\ \vdots\\ x_{2,1}\\ 0\\ x_{2,3}\\ \vdots\\ \vdots\\ 0\end{bmatrix}}
$$

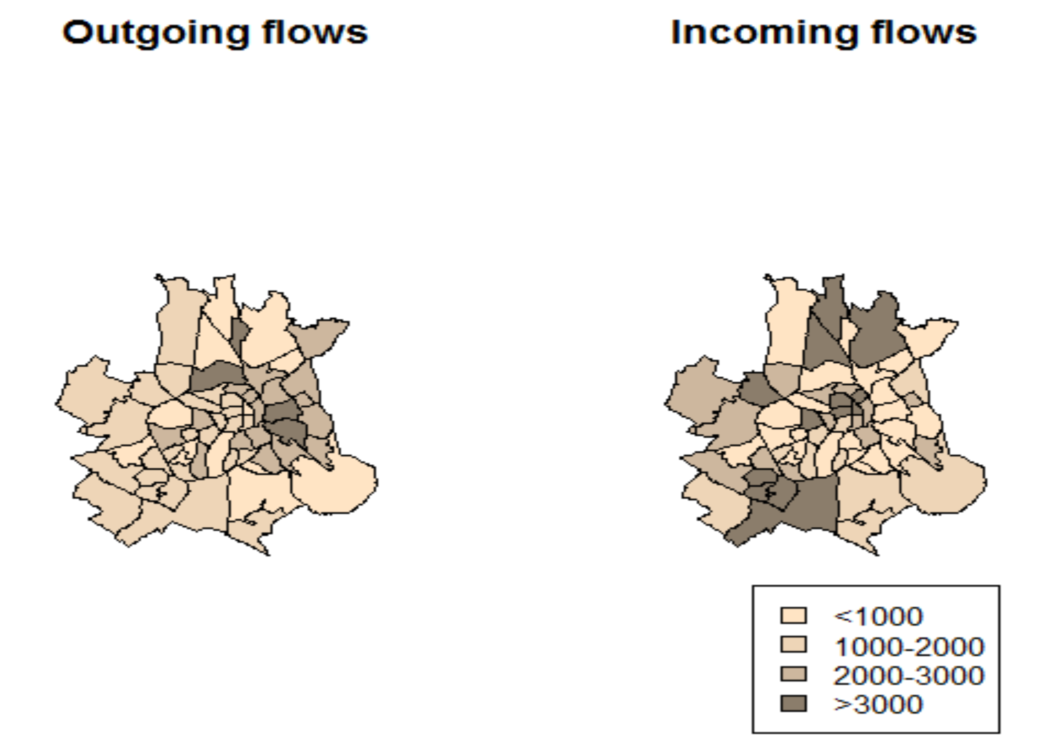
Note:
Xo_intra = Xd_intra

$$X_{i,j}\_\mathrm{int}ra = \begin{cases} x_{ij} & If \quad i=j \\ 0 & otherwise \end{cases}$$

$$X_{i,j}\_\mathrm{int}er = \begin{cases} 0 & If \quad i=j \\ x_{ij} & otherwise \end{cases}$$

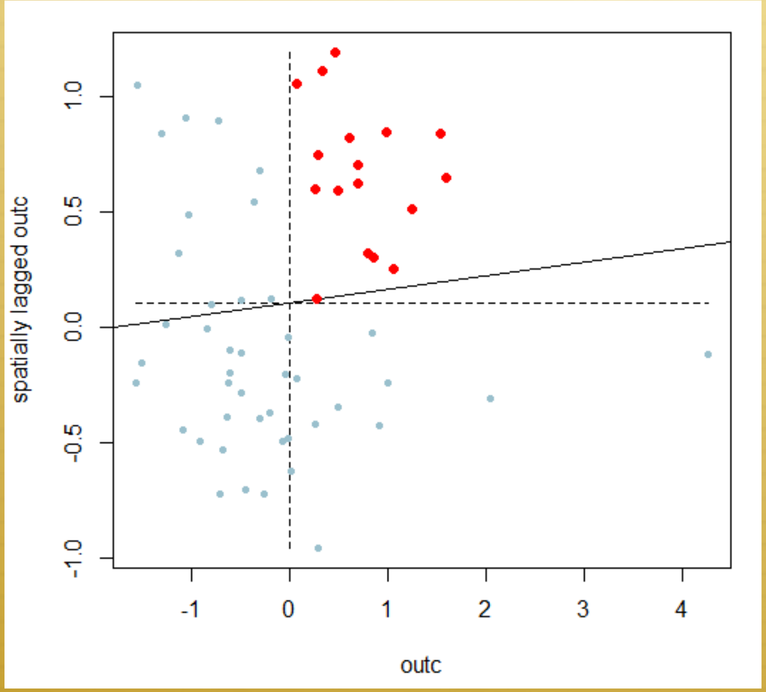
Outgoing Flow & Incoming Flow

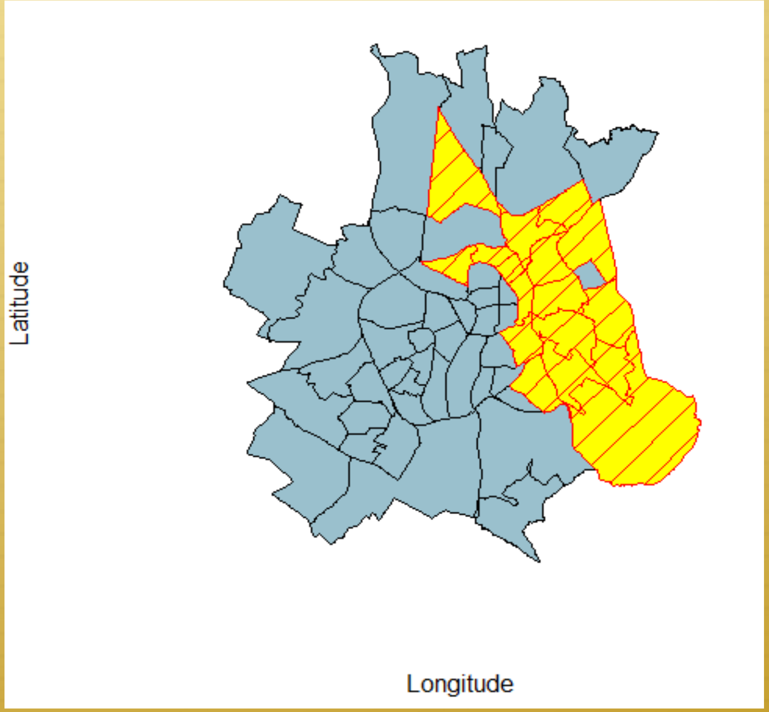
- Outgoing Flow: sum of the flows go to different destination for each origin
- Incoming Flow: sum of the flows from different origin for each destination



Exploratory Analysis - FLOWS

- Moran plot for outgoing flows

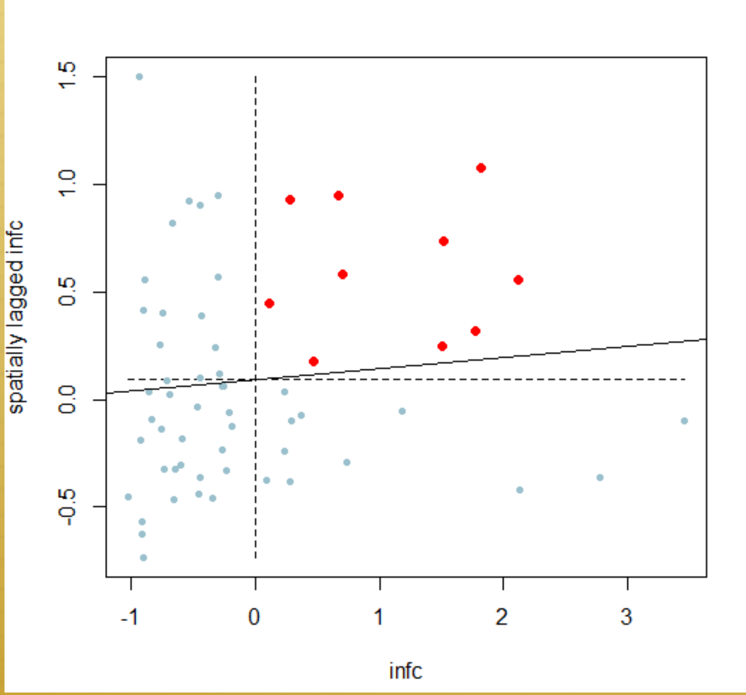


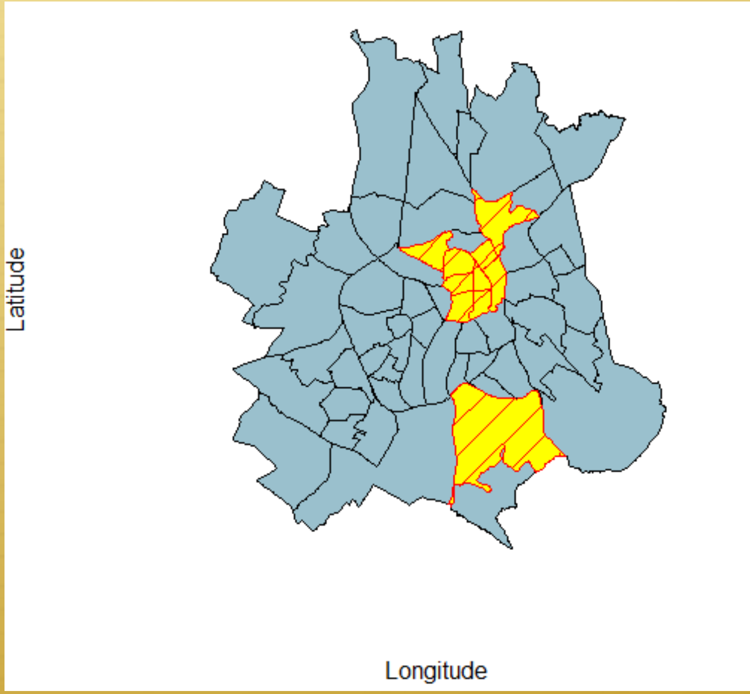


For regions with high outgoing flows, neighbours also have high outgoing flows!

Exploratory Analysis - FLOWS

- Moran plot for incoming flows

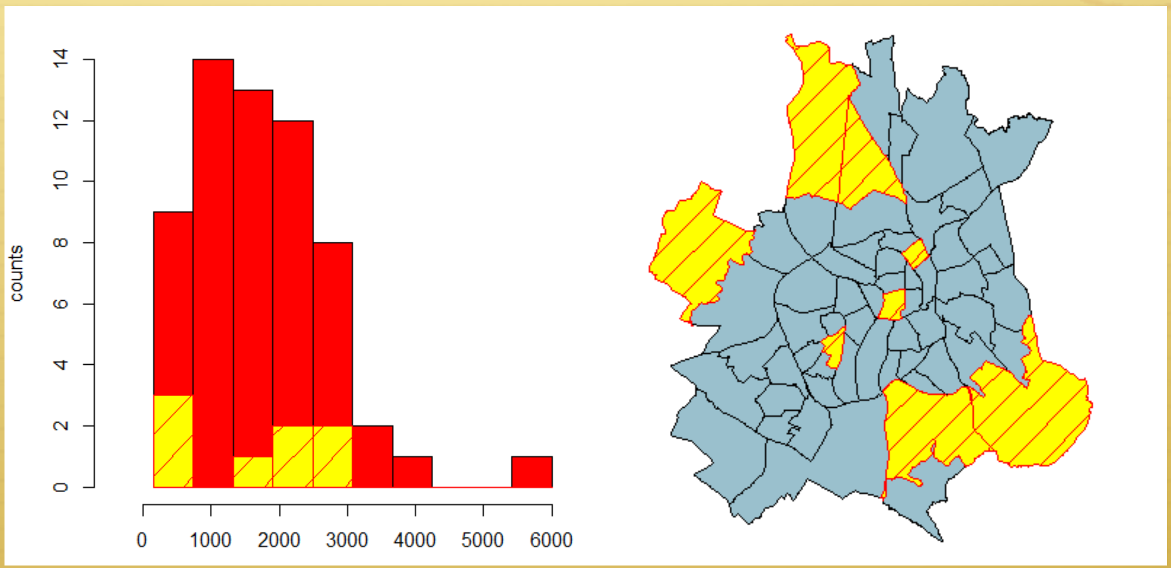




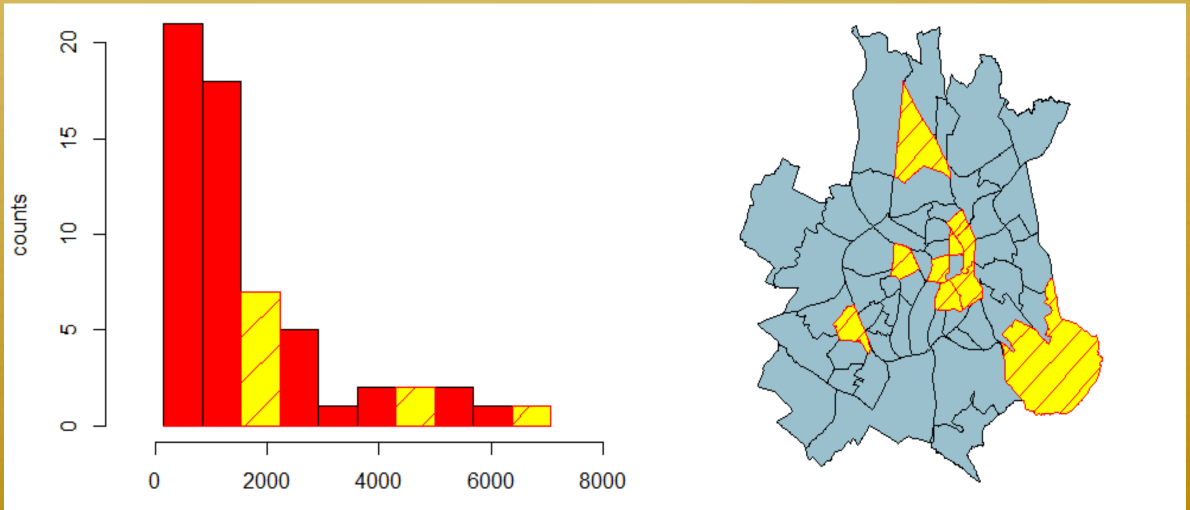
For regions with high incoming flows, neighbours also have high incoming flows!

Exploratory Analysis - FLOWS

- Outgoing flows

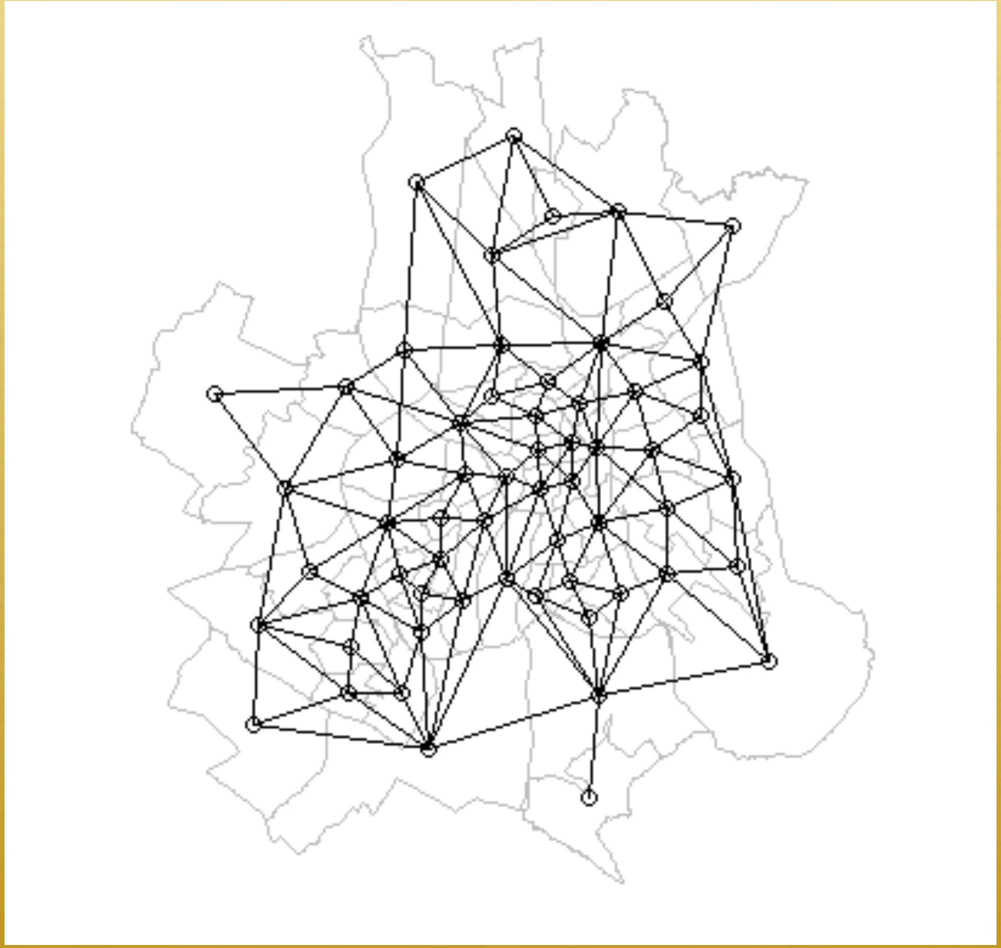


- Incoming flows



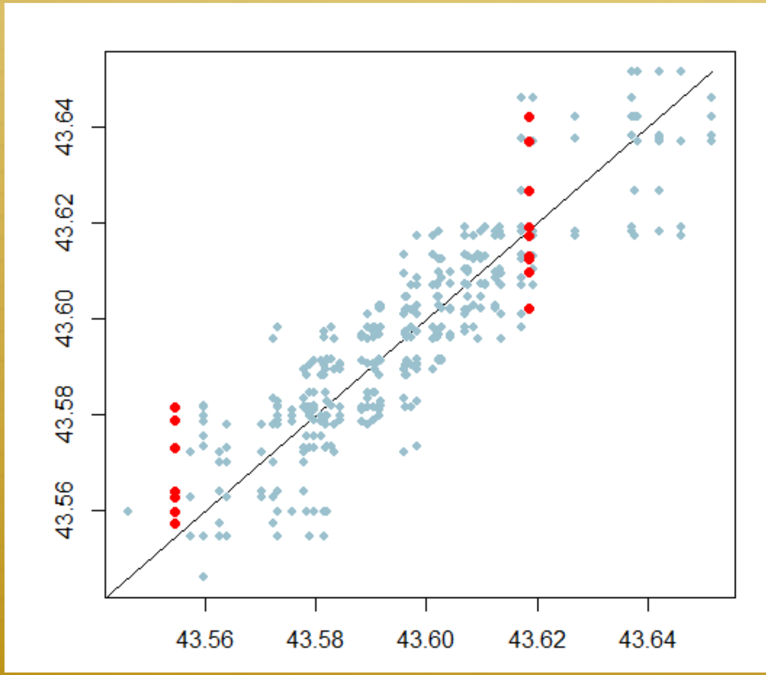
Exploratory Analysis – Weight Matrix

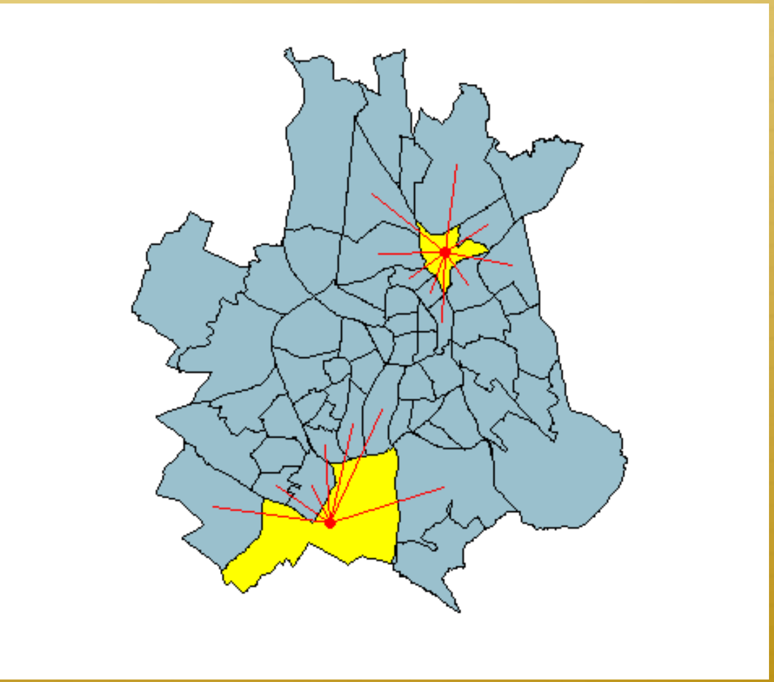
- Weight matrix constructed using common border method



Exploratory Analysis – Weight Matrix

- Point selected on map
- Number of neighbours marked on the graph by
- Variable : Latitude





Modeling

- Simple OLS model without adjustment (OLS)

$$Y = X_o\beta_o + X_d\beta_d + \varepsilon$$

- OLS model with adjustment (OLS_a)

$$Y = X_o\beta_o + X_d\beta_d + X_{intra}\beta_{intra} + \varepsilon$$

- Lag model with adjustment (LAG_a)

$$Y = \rho WY + X_o\beta_o + X_d\beta_d + X_{intra}\beta_{intra} + \varepsilon$$

- Durbin model with adjustment (Durbin_a)

$$Y = \rho WY + (X_o + X_d + X_{intra})\beta + \delta W(X_o + X_d + X_{intra}) + \varepsilon$$

NOTE: $W = Wo + Wd = w \otimes I_{60} + I_{60} \otimes w$

w = common boarder neighborhood matrix (60*60)

Results: OLS model V.S. OLS with adjustment

	Model: OLS		Model:OLS_a	
	estimate	p-value	estimate	p-value
intercept	1.2700	0.501555	1.2093	0.427134
dist	-608.6000	< 2e-16 ***	-336.4124	< 2e-16 ***
acro	0.9775	< 2e-16 ***	0.8142	< 2e-16 ***
lfd	0.0302	0.000129 ***	0.0279	1.37e-05 ***
acrd	0.8526	6.37e-13 ***	0.7141	1.15e-13 ***
unempd	-2.4800	8.48e-16 ***	-0.0520	3.41e-10 ***
hd	0.0118	< 2e-16 ***	-2.2621	< 2e-16 ***
empd	-0.0533	1.63e-07 ***	0.0120	< 2e-16 ***
l_o_intra	-	-	-0.1388	0.001295 **
acr_o_intra	-	-	-1.6040	0.002999 **
em_o_intra	-	-	0.2082	0.000243 ***
unem_o_intra	-	-	2.0153	0.212075
hou_o_intra	-	-	0.0330	2.97e-07 ***
AIC	37584.53		36029.68	

Effects of
- distance decrease
- employment rate of destination become positive
- house unit of destination become positive

AIC decrease:
- model fit better

Procedure:

✓ OLS model with all the origin and destination explanatory variables

✓ Eliminate the non significant variables → OLS model above
- Origin variables: activity rate
- Destination variables: labor force, activity rate, employment, unemployment rate, house unit

✓ Add internal flows' characteristics for the significant variables (variables with intra) → OLS_a model above

Results: Lag Model and Durbin Model

✧ Why use lag model and durbin model?

Spatial autocorrelation check:

→Moran Test on the residual of traditional gravity model (OLS_a)

```
        Moran's I test under normality

data:  res
weights: mat2listw(Wt)

Moran I statistic standard deviate = 46.1522, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     3.387390e-01     -2.778550e-04      5.395822e-05
```

P-value is very small:

→ Reject the H0: no spatial autocorrelation

→ The residuals has spatial autocorrelation

→ Use lag model or durbin model

Results: OLS model V.S. Lag Model

	Model:OLS_a		Model:LAG_a	
	estimate	p-value	estimate	p-value
intercept	1.2093	0.427134	-9.5987	<2e-15 ***
dist	-336.4124	< 2e-16 ***	45.6277	0.0700
acro	0.8142	< 2e-16 ***	0.2384	0.0001 ***
lfd	0.0279	1.37e-05 ***	0.0091	0.0750
acrd	0.7141	1.15e-13 ***	0.2251	0.0032 **
empd	0.0120	< 2e-16 ***	-0.0191	0.0036 **
unempd	-0.0520	3.41e-10 ***	-0.7655	0.0001 ***
hd	-2.2621	< 2e-16 ***	0.0052	<1e-11 ***
l_o_intra	-0.1388	0.001295 **	-0.1424	<2e-5 ***
acr_o_intra	-1.6040	0.002999 **	-1.4310	0.0008 ***
em_o_intra	0.2082	0.000243 ***	0.2112	<2e-6 ***
unem_o_intra	2.0153	0.212075	2.5190	0.0484 *
hou_o_intra	0.0330	2.97e-07 ***	0.0293	<8e-9 ***
rho			0.38876	<2.22e-16 ***
AIC	36029.68		34590.06	

- Effects of distance and labor force of destination become not significant
- Rho is significant, which means there exits an autocorrelation of the flows
- AIC decreases a lot, model fits much better

Results: Durbin Model

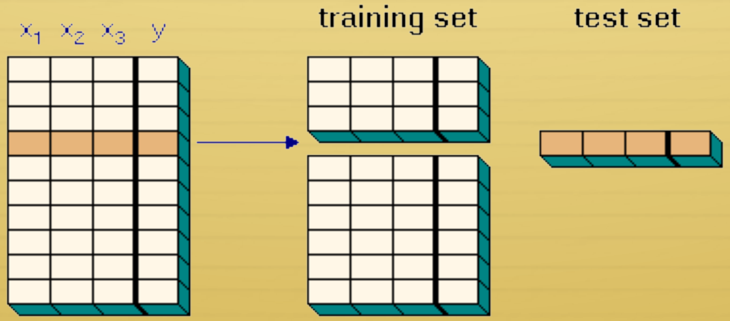
	Model:OLS_a			Model:OLS_a	
	estimate	p-value		estimate	p-value
intercept	-16.5511	<2e-8 ***			
dist	-16.2945	0.5418			
acro	0.1214	0.1936	acro_lag	0.0511	0.4539
lfd	0.0243	0.0066 **	lfd_lag	-0.0285	0.0002 ***
acrd	0.4343	0.0007 ***	acrd_lag	-0.2112	0.0497 .
empd	-0.0429	0.0002 ***	empd_lag	0.0399	<0.0000 ***
unempd	-1.3083	0.0001 ***	unempd_lag	1.0380	0.0004 ***
hd	0.0104	<2e-13 ***	hd_lag	-0.0050	<0.0000 ***
l_o_intra	-0.1351	<0.0000 ***	l_o_intra_lag	0.2425	<0.0000 ***
acr_o_intra	-1.4861	<0.0003 ***	acr_o_intra_lag	2.0040	0.0014 **
em_o_intra	0.2010	<0.0000 ***	em_o_intra_lag	-0.2994	<0.0000 ***
unem_o_intra	2.3885	0.0471 .	unem_o_intra_lag	-7.1636	0.0001 ***
hou_o_intra	0.0290	<0.0000 ***	hou_o_intra_lag	-0.0103	0.1104
rho	0.38876	< 2.22e-16 ***			
AIC	34325.3				

- Rho is significant, the most of lag explanatory variables are significant
- AIC decreases a little bit

Prediction

Procedure:

- Eliminate **randomly** 10% of data, denote the remaining data by training set and the 10% of eliminated data by test set.



- Construct model (eg, lag model) on training set

$$Y_k^{n.e} = \rho W^{n.e} Y_k^{n.e} + X_k^{n.e} \beta + \varepsilon$$

→ obtain the estimated coefficients ($\hat{\beta}$, $\hat{\rho}$)

- Predict on test set by using the coefficients we got

$$\hat{Y}_k^{e} = \hat{\rho} W^{e} Y_k^{e} + X_k^{e} \hat{\beta}$$

Evaluation of the Model

- Quadratic Mean Error (QME)

$$mean\left[(\hat{Y}_k - Y_k)^2\right] \qquad \textit{For k belongs to test set}$$

- Relative Quadratic Error (RQE)

$$sum\left[\left(\frac{\hat{Y}_k - Y_k}{Y_k}\right)^2\right] \qquad \textit{For k belongs to test set}$$

- Traditional Gravity Model V.S. Lag Model

	AIC	QME	RQE
Traditional Gravity Model	36029. 68	2266.525	3042.47
Lag Model	34590. 06	2395.717	309.29

Conclusion

- Spatial Autocorrelation Present
- AIC gets lower when autocorrelation is taken into consideration
- Relative error lower in the lag model than the traditional gravity model