The University of Chicago
MACS 30200 Research Paper Initial Result:
An Analysis of the Chicago Taxi Industry:
Predict Taxi Demand Per Capita from a Spatial Approach

Zhang, Dongping *

May 17, 2017

1 Research Question

What is the relationship between the aggregated taxi pick-ups of a community and spatially explicit socio-demographic, built-environment, and urban-transportation variables in Chicago?

2 Data

2.1 Source of 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 two main data sources are *The Chicago Data Portal* and *The Chicago Metropolitan Agency for Planning (CMAP)*. Multiple datasets and files would be downloaded from each of the sources.

2.1.1 The Chicago Open Data Portal

Chicago Taxi Trips Dataset 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).

Chicago Business Licenses Dataset contains every issuance of business license from 2002 to the present and the dataset is still updated monthly.

Chicago Transit Authority Bus Stops Dataset contains every spatial coordinates of bus stops in the Chicago area.

Chicago Transit Authority L'Train Stations Dataset contains every spatial coordinates of L'Train stations in the Chicago area.

Current Chicago Community Boundary Shapefile 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.

*The University of Chicago, MA in Computational Social Science, CNetID: 12144965, dpzhang@uchicago.edu.

2.1.2 The Chicago Metropolitan Agency for Planning

CMAP Community Data contains a set of variables that summarize demographics, housing, employment, transportation habits, retail sales, property values, and land use in all 77 Chicago Community Areas. Variables are compiled from the U.S. Census Bureau's 2010-14 American Community Survey, Longitudinal Employment-Household Dynamics data for 2014, and 2014/2015 data from the Illinois Department of Employment Security and the Illinois Department of Revenue (Sources: the U.S. Census Bureau, the Illinois Environmental Protection Agency, the Illinois Department of Employment Security, the Illinois Department of Revenue, and CMAP).

As noted, 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.

2.2 Data Processing

2.2.1 Chicago Taxi Trips Dataset

This 40GB dataset 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 and could be further aggregated by year (2013, 2014, 2015, and 2016) using the timestamp variable.

2.2.2 Chicago Business Licenses Dataset

Chicago Business Licenses Dataset is processed using the same procedure as the *Chicago Taxi Trips Dataset*. There are 32 variables in the raw dataset but only *"Latitude"* variable and *"Longitude"* variable of each business would be used. Using the *Current Chicago Community Boundary Shapefile*, number of business in community could be obtained.

2.2.3 Chicago Transit Authority Bus Stops Dataset

Chicago Transit Authority Bus Stops Dataset is processed using the same procedure as the *Chicago Taxi Trips Dataset*. There are 11074 bus stops in the Chicago area and using *Current Chicago Community Boundary Shapefile*, the aggregate counts of bus stops in each of the 77 spatial unit could be obtained.

2.2.4 Chicago Transit Authority L'Train Stations Dataset

Chicago Transit Authority L'Train Stations Dataset is processed using the same procedure as the *Chicago Taxi Trips Dataset*. There are 234 L'Train stations in the Chicago area and using *Current Chicago Community Boundary Shapefile*, the aggregate counts of L'Train stations in each of the 77 spatial unit could be obtained.

2.2.5 CMAP Community Data

There are 77 PDF files containing some key socio-demographic variables of each neighborhood. Because each PDF file has the same format, any PDF scrapper method could be used to obtain the variables of interest. There are a total of 8 variables needed from each of the 77 PDF file, and they are:

- Source: 2014 American Community Survey, five-year estimates
  - the number of black population from each community
  - the median income of each community
  - number of people holding a Bachelor's Degree or Higher
  - total number of commuters
- Source: 2000 and 2010 Census
  - Total population
  - Average Household Size
- Source: CMAP analysis of US Census Bureau, HERE, and Illinois Environmental Protection Agency data
  - annual vehicle miles traveled (VMT) per household in 2013
- Source: CMAP calculations of 2010 Land Use Inventory
  - park Acreage per 1,000 Residents
- Source: Chicago Metropolitan Agency for Planning Parcel-Based Land Use Inventory
  - single-Family Residential land acres
  - multi-Family Residential acres
  - commercial acres
  - total acre

2.3 Variables

2.3.1 Raw Variables

As described, there would be a total of 20 raw variables extracted and merged from above 5 data sources. Table 1 presents the names and brief descriptions of raw variables.

Table 1: Raw Variables

List of Raw Variables

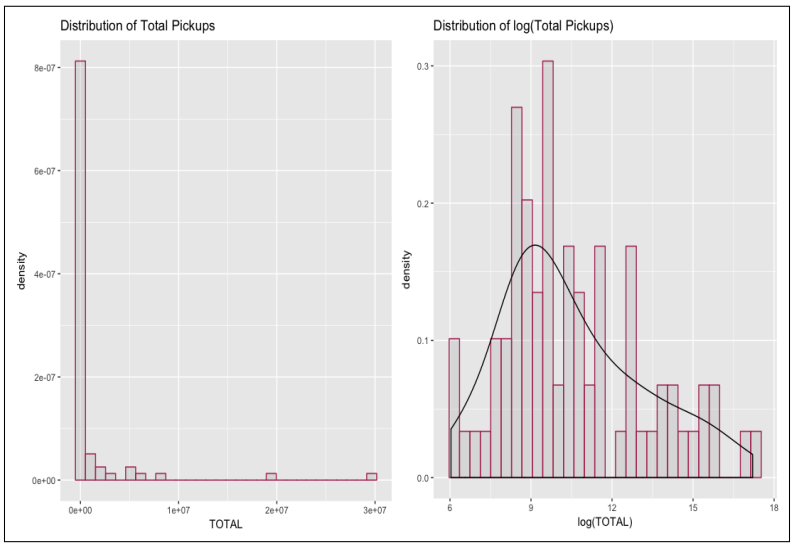
Variables	Descriptions
PICKUP13	number of taxi pickups in a community in 2013
PICKUP13	number of taxi pickups in a community in 2013
PICKUP13	number of taxi pickups in a community in 2013
PICKUP13	number of taxi pickups in a community in 2013
TOTAL	number of aggregated taxi pickups in a community
POP	total population of a community
AHHDS	average household size of a community
BLACK	the number of black population from each community
MINC	the median income of each community
BS	number of people holding a Bachelor's Degree or Higher
COMMUTER	total number of commuters
VMT	annual vehicle miles traveled (VMT) per household in 2013
PARK	park Acreage per 1,000 Residents
SRES	single-family Residential acres
MRES	multi-family Residential acres
COM	commercial acres
TOTAL	total acres
BUS	number of CTA bus stops in a community
LTRAIN	number of CTA L'Train stations in a community

2.3.2 Variables Processing

Different communities have varying characteristics, and some socio-demographic variables, such as total populations, is highly correlated with 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, several variable operations would be implemented so as to avoid size problem.

PICKUPS Taxi Pickups is divided by the total population so as to obtain number of pickups per capita in a neighborhood. In addition, a log transformation of picks per capita would also be necessary to the dependent variable so as to make the data more normally distributed, which is presented in Fig 1.

Figure 1: Histogram of Total Pickups



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

VMT VMT represents annual vehicle miles traveled (**VMT**) per household in 2013. Dividing VMT by average household size of a community to obtain average VMT per capita.

RES Obtain total residential area by summing single-family Residential acres and multi-family Residential acres.

2.3.3 Clean Variables

Table 2 and Table 3 contains processed and clean variables that would be used in my model.

Table 2: Dependent Variables

List of Dependent Variables

Variables	Definition
PICKUP13	number of taxi pickups in a community in 2013
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	number of aggregated taxi pickups in a community

Table 3: Independent Variables

List of Independent Variables

Category 1: Socio-demographic	
Variables	Definition
BLACK	number of African American per acre in a community
MINC	median income of a community
BS	number of people holding a B.S. degree or higher per acre in a community
COMMUTERS	number of people who need to commute to work per acre in a community
VMT	annual Vehicle Miles Travelled per capita
BIZ	number of businesses per acre in a community
Category 2: Built-environment	
Variables	Definitions
PARK	park acre per 1,000 Residents
RES	total residential acre of a community
COM	total commercial acre of a community
Category 3: Urban-transportation	
Variables	Definitions
BUS	number of CTA bus stations in a community
LTRAIN	number of CTA L-Train stations in a community

2.3.4 Summary Statistics

Table 4: **Summary Statistics**

Statistic	N	Mean	St. Dev.	Min	Max
PICKUPS13	77	10.364	34.785	0.003	185.382
PICKUPS14	77	12.315	41.161	0.002	222.678
PICKUPS15	77	10.661	36.343	0.003	209.200
PICKUPS16	77	7.750	27.838	0.002	172.686
TOTAL	77	41.089	139.470	0.010	785.084
POP	77	20.310	10.741	1.679	49.798
AHHDS	77	2.755	0.582	1.600	4.300
BLACK	77	6.303	6.700	0.015	25.697
MINC	77	46,712.290	20,336.540	14,390	94,823
BS	77	4.867	6.309	0.057	31.322
COMMUTER	77	8.801	6.450	0.609	32.550
VMT	77	4,209.320	969.358	2,315.000	6,667.826
PARK	77	6.725	15.963	0.200	124.700
RES	77	604.621	363.850	56.800	1,759.600
COM	77	92.703	74.099	2.300	394.000
ACRES	77	1,918.349	1,262.035	365.500	8,536.200
BIZ	77	0.441	0.643	0.019	4.877
BUS	77	131.909	76.722	16	405
LTRAIN	77	2.779	5.536	0	36

2.3.5 Exploratory Spatial Data Analysis (ESDA)

Fig 2 is a quantile map showing the density of log(PICKUP) counts. It is very straightforward that taxi pickups are very segregated in Chicago as there are more pickups in northern neighborhood than southern neighborhoods.

Fig 3 presents Moran's I with LOWESS smoother to assess spatial autocorrelation of log(PICKUP) counts using a set of queen contiguity weights. The standarized log(PICKUP) counts is on the x-axis and its spatial lag values is on the y-axis. It can be shown that with the increase of log(PICKUP) counts, the weighted average of its neighborhood also would increase. The Moran's I statistics is 0.619775, indicating a positive spatial autocorrelation.

Fig 4 and Fig 5 presents the local spatial-autocorrelation or the clustering pattern of log(PICKUP) using the same queen contiguity weights. It could be sees clearly that high-high points are clustered in the northern communities while low-low points are clustered in the southern communities.

Figure 2: Pickup Density Map

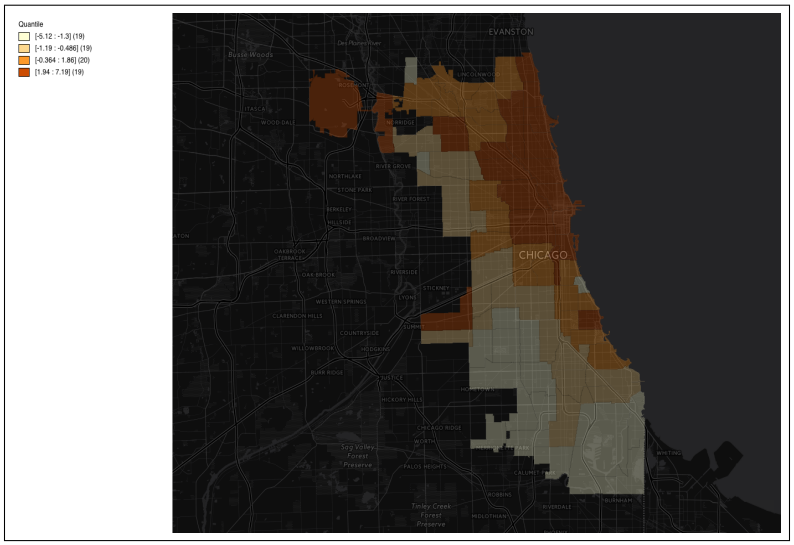


Figure 3: Moran's I: log(PICKUP)

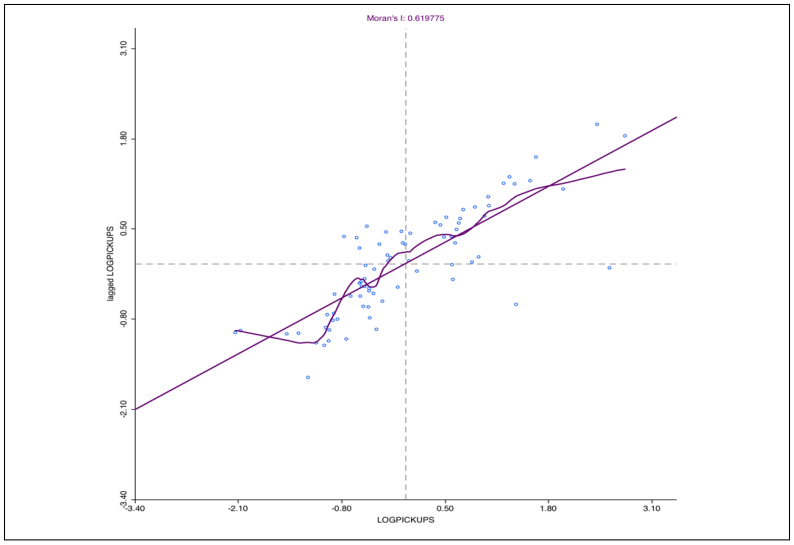


Figure 4: LISA Cluster Map: Log(PICKUP)

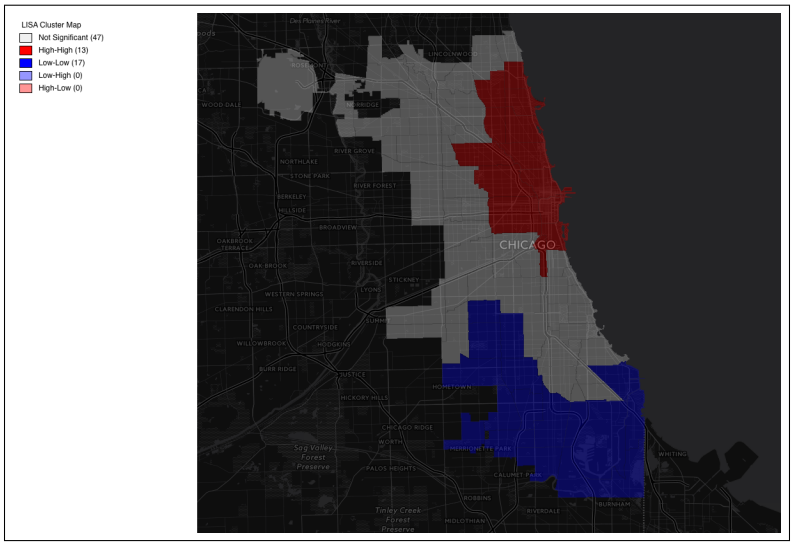
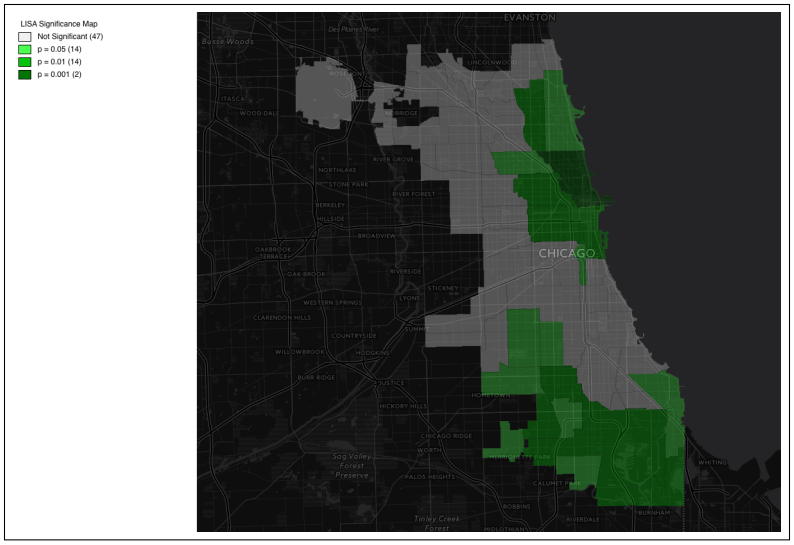


Figure 5: LISA Significance Map: Log(PICKUP)



3 Model Specification

- Non-spatial Model specification:

$$log(PICKUPS) = \beta_0 + \sum_{i=1}^{11} \beta_i X_i + \epsilon_i$$

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  taxi_pickups
Dependent Variable  :  LOGPICKUPS  Number of Observations:   77
Mean dependent var  :    0.260989  Number of Variables   :   11
S.D. dependent var  :     2.49689  Degrees of Freedom    :   66

R-squared           :    0.345491  F-statistic           :      3.4839
Adjusted R-squared  :    0.246323  Prob(F-statistic)     :  0.00100053
Sum squared residual:     314.198  Log likelihood        :    -163.398
Sigma-square        :     4.76058  Akaike info criterion :     348.795
S.E. of regression  :     2.18188  Schwarz criterion     :     374.577
Sigma-square ML     :      4.0805
S.E of regression ML:     2.02002

------------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
          CONSTANT       1.80435        1.38991        1.29818     0.19874
             BLACK    -0.0627023      0.0516606       -1.21373     0.22917
              MINC   -1.2684e-05    2.31461e-05      -0.547995     0.58554
                BS      0.124139       0.121006        1.02589     0.30869
               BIZ      -3.65353       0.816788       -4.47305     0.00003
              PARK    -0.0121612      0.0191523      -0.634977     0.52764
               RES   -0.00184469     0.00116233       -1.58706     0.11728
               COM    0.00825415     0.00552996        1.49262     0.14030
               BUS   -0.00119182     0.00681122      -0.174979     0.86164
            LTRAIN      0.326678      0.0971541        3.36248     0.00129
          COMMUTER      0.015191       0.105049        0.14461     0.88546
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   23.351006
TEST ON NORMALITY OF ERRORS
TEST                   DF           VALUE             PROB
Jarque-Bera             2             7.7766          0.02048

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
```

```
TEST                    DF          VALUE            PROB
Breusch-Pagan test      10           13.0953         0.21839
Koenker-Bassett test    10            8.3496         0.59473
SPECIFICATION ROBUST TEST
TEST                    DF          VALUE            PROB
White                   65           73.7397         0.21403

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : weights_rook
   (row-standardized weights)
TEST                           MI/DF       VALUE          PROB
Moran's I (error)              0.3343       5.1768        0.00000
Lagrange Multiplier (lag)        1         36.8685        0.00000
Robust LM (lag)                  1         22.3323        0.00000
Lagrange Multiplier (error)      1         18.5281        0.00002
Robust LM (error)                1          3.9920        0.04572
Lagrange Multiplier (SARMA)      2         40.8604        0.00000
============================== END OF REPORT ================================
```

3.1 Non-Spatial Regression Diagnostics

- Non-spatial Diaganostics
  - $R^2 = 0.345491$ suggest that OLS model explains about one third of the total variance
  - The multicollinearity condition number is 23.35, which does not suggest any potential problems
  - The value of the Jarque-Bera test statistic is 7.78 with a p-value of 0.02048, which suggest non-normal distribution of data
  - The test statistics for heteroskedasticity from Breusch-Pagan test and Koenker-Bassett test are 13.10 and 8.35 correspondingly to p-values of 0.22 and 0.59, which suggest homoskedasticity (The difference between the two statistics also also confirms potential error non-normality, since under normality, the value for both should be roughly the same)

3.2 Diagnostics for Spatial Effects

Using a rook contiguity weights,

- The Moran's I is 5.1768, which is highly significant, suggesting strong spatial autocorrelation
- The $LM_\rho = 36.8685$, which is highly significant for a χ^2 variate with one degree of freedom, strongly indicating the presence of spatial autocorrelation
- The $LM_\rho^* = 22.3323$, which suggests and confirms the lag specification

- The $LM_\lambda = 18.5281$, which is highly significant suggesting the error specification
- The $LM_\lambda^* = 0.04572$, which is weakly significant suggesting the error specification
- The $LM_{\rho\lambda} = 40.8604$ with two degrees of freedom, which is highly significant, suggesting a higher order model is the correct alternative specification

3.3 Interpretations of Spatial Effects Diagnostics

Because the p-values for both LM_ρ and LM_λ are less than 0.00000, I would prefer to look at the robust statistics, LM_ρ^* and LM_λ^*. The robust test statistics for the lag alternative is highly significant with p-value < 0.00000 while the robust test statistics for the error alternative is weakly significant with p-value < 0.04572. I would leaning towards a spatial lag model.

3.4 Spatial Lag Regression Diagnostics

After assessing the spatial effect of the OLS model, a spatial lag term is added to the regression model to account for the effect spatial heteroskedasticity. It is shown that R^2 increased significantly to about 70% (More complex spatial model such as semi-parametric geographically weighted Poisson regression model could be potentially used to obtain better result).

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : taxi_pickups
Spatial Weight      : weights_rook
Dependent Variable  :  LOGPICKUPS  Number of Observations:   77
Mean dependent var  :    0.260989  Number of Variables   :   12
S.D. dependent var  :     2.49689  Degrees of Freedom    :   65
Lag coeff.   (Rho)  :    0.762824

R-squared           :    0.692069  Log likelihood        :    -141.288
Sq. Correlation     : -            Akaike info criterion :     306.575
Sigma-square        :     1.91978  Schwarz criterion     :     334.701
S.E of regression   :     1.38556
```

Variable	Coefficient	Std.Error	z-value	Probability
W_LOGPICKUPS	0.762824	0.0727381	10.4873	0.00000
CONSTANT	2.27832	0.885243	2.57366	0.01006
BLACK	-0.0333139	0.0328931	-1.01279	0.31116
MINC	-2.21353e-05	1.47088e-05	-1.5049	0.13235
BS	0.158642	0.0769509	2.0616	0.03925
BIZ	-1.56611	0.528852	-2.96135	0.00306
COMMUTER	-0.105141	0.0667108	-1.57608	0.11501
PARK	-0.0138954	0.0121671	-1.14205	0.25343
RES	-0.000201158	0.000741878	-0.271147	0.78628

```
             COM    0.00611834     0.00351204         1.7421     0.08149
             BUS   -0.00695253     0.00433025       -1.60557     0.10837
          LTRAIN      0.158183      0.0624126        2.53448     0.01126
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                     DF      VALUE        PROB
Breusch-Pagan test                       10         8.9848     0.53355

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : weights_rook
TEST                                     DF      VALUE        PROB
Likelihood Ratio Test                    1        44.2203     0.00000
================================ END OF REPORT =================================
```