Spatial Autocorrelation Diagnostics and Spatial Lag/Error Models for the analysis of crime in the city of Chicago

Location: City of Chicago

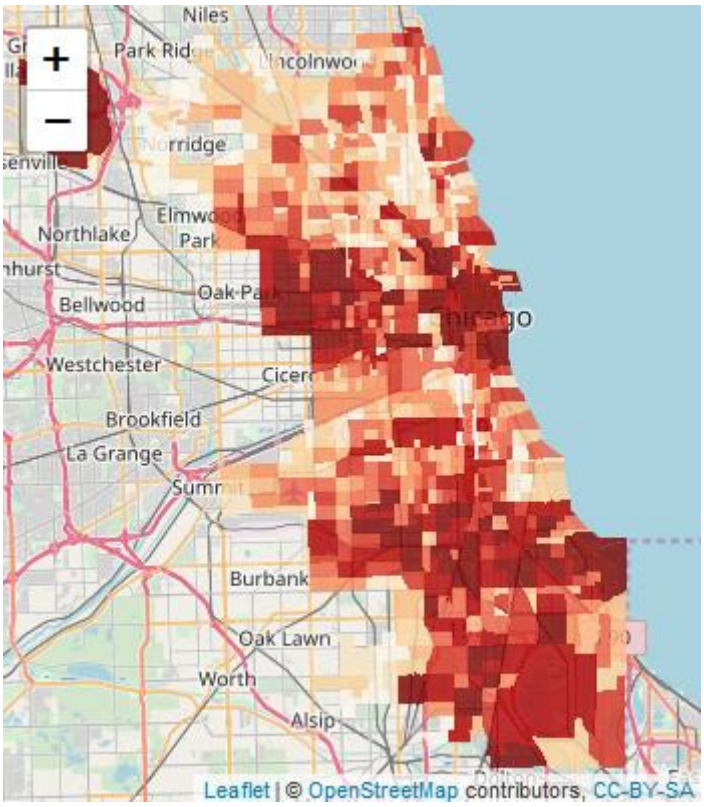
Level of detail: Census tracts

Coordinate Reference System (CRS): NAD 1983 State Plane Illinois East (EPS102671)

Introduction

The purpose of this study is to analyze crime in Chicago by running three different regression models: an ordinary least squares (OLS) regression, a spatial autocorrelation model to account for spatial dependence between tracts, and a geographically weighted regression to investigate changes in predictors' coefficients over space. The models seek to explore the same dependent variable "Count_crim" which is a count of crime incidents per tract in Chicago from January to September 2018. The independent variables also remain the same across models and include variable "Mean.Annua" (the mean annual income per tract), "Unemp.in.W" (a count of unemployed people per tract) and "Bachelor" (a count of college graduates per tract). We chose unemployment, income and education as predictors because we believe that they are statistically significant. We hypothesize that unemployment will be positively correlated with crime, whereas higher income and a more educated population should be negatively correlated with crime. We also add a control variable "Land_area", the surface area in m^2 of each tract, because we expect that larger tracts have more crime incidents than smaller ones. Before doing any analysis, we recall

from our previous paper what criminality looks like in Chicago using a choropleth map of crime counts below.



1. **Ordinary Least Squares (OLS) model**

First, we run an OLS regression in R and obtain the output below.

```
Call:
lm(formula = Count_crim ~ as.numeric(Bachelor) + as.numeric(Mean.Annua) +
    as.numeric(Unemp.in.W) + as.numeric(Land_area_), data = chi.poly@data)

Residuals:
   Min     1Q Median     3Q    Max
-612.8 -148.2  -52.7   76.9 3345.5

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.404e+02  2.870e+01  15.346  < 2e-16 ***
as.numeric(Bachelor)  -2.805e-01  6.036e-02  -4.647 3.95e-06 ***
as.numeric(Mean.Annua) -2.456e-01  4.159e-02  -5.904 5.26e-09 ***
as.numeric(Unemp.in.W)  1.588e-01  7.853e-02   2.022   0.0435 *
as.numeric(Land_area_)  5.147e-05  1.030e-05   4.995 7.22e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 266.1 on 793 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.1053,    Adjusted R-squared:  0.1008
F-statistic: 23.33 on 4 and 793 DF,  p-value: < 2.2e-16
```

We notice that all predictors, i.e unemployment, income, education and land area, are statistically significant with p-value < 0.05. The adjusted R-squared of the regression is about 10% which is low but reasonable given that there are only four predictors. Additionally, we observe that our former hypothesis is confirmed by the signs of the estimated coefficients with positive coefficient estimates for land area and unemployment, and negatives coefficients for education and income. These coefficients respectively are 0.00005147, 0.1588, -0.2805 and -0.2456. We can interpret them as follows:

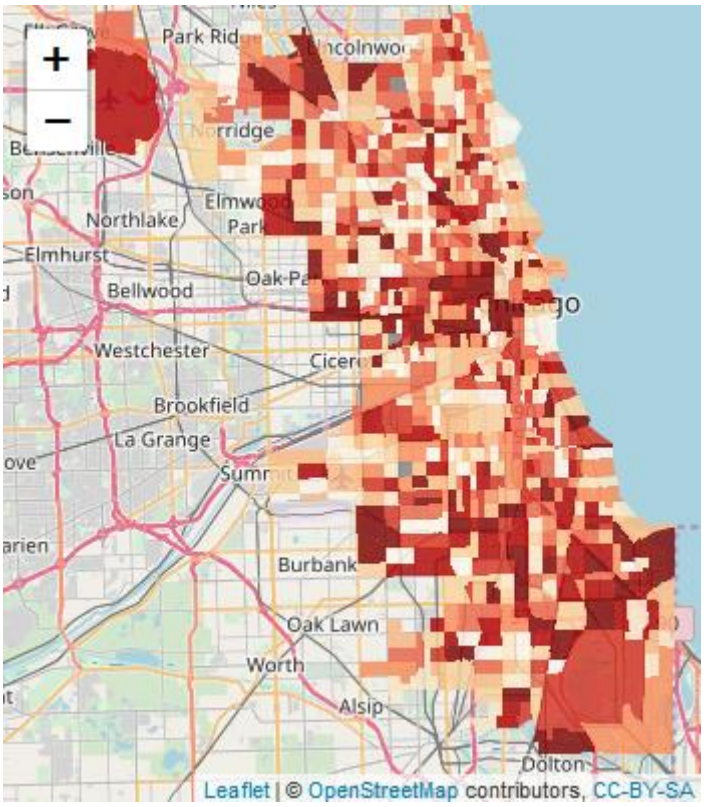
- Increasing by 1m² the land area per tract increases the number of crimes per year in this tract by 0.00005147
- Controlling for tracts' land area, the addition of 1 unemployed person in the tract's population increases the tract's annual number of crimes by 0.1588.
- Controlling for tracts' land area, an increment of $1 in the tract's mean annual income decreases the annual number of crimes in the tract by 0.2456.
- Controlling for tracts' land area, each college graduate in the tract's population decreases the annual number of crimes in the tract by 0.2805.

2. Spatial autocorrelation model

One of the caveats of the OLS model in the context of spatial regression is that the assumption of residuals independence is often violated. Let's verify if this assumption holds using a Moran's I test and a visual inspection of the distribution of residuals across the city. To build the Moran's I test we first create a spatial weight matrix of order 1 with Queen contiguity because we expect all connected tracts to have an influence on each other. The output of the test is shown below:

```
        Global Moran I for regression residuals

data:
model: lm(formula = Count_crim ~ as.numeric(Bachelor) + as.numeric(Mean.Annua) + as.numeric(Unemp.in.W),
data = chi.poly@data)
weights: W

Moran I statistic standard deviate = 20.745, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
Observed Moran I      Expectation         Variance
    0.4110262128    -0.0022084150     0.0003968052
```

The null hypothesis of the test is spatial independence and the p-value of the test is essentially zero therefore we conclude that there is spatial dependence in our dataset. This is confirmed by a visual inspection of the residuals which are clustered near South Side and Downtown.



Thus, we need a new model that accounts for this dependency. There are two possible models that could be a good fit, spatial lag or spatial error, so we run additional test to

identify which one is the most appropriate. The output of this diagnostics is below.

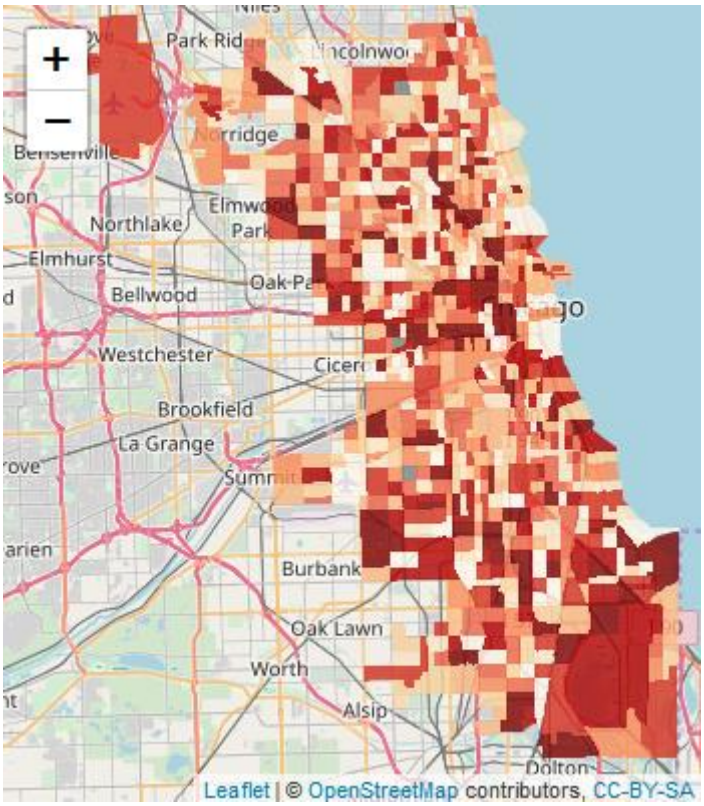
```
        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = Count_crim ~ as.numeric(Bachelor) + as.numeric(Mean.Annua) + as.numeric(Unemp.in.W) +
as.numeric(Land_area_), data = chi.poly@data)
weights: W

LMerr = 468.2, df = 1, p-value < 2.2e-16


        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = Count_crim ~ as.numeric(Bachelor) + as.numeric(Mean.Annua) + as.numeric(Unemp.in.W) +
as.numeric(Land_area_), data = chi.poly@data)
weights: W

LMlag = 459.92, df = 1, p-value < 2.2e-16


        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = Count_crim ~ as.numeric(Bachelor) + as.numeric(Mean.Annua) + as.numeric(Unemp.in.W) +
as.numeric(Land_area_), data = chi.poly@data)
weights: W

RLMerr = 18.047, df = 1, p-value = 2.156e-05


        Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = Count_crim ~ as.numeric(Bachelor) + as.numeric(Mean.Annua) + as.numeric(Unemp.in.W) +
as.numeric(Land_area_), data = chi.poly@data)
weights: W

RLMlag = 9.7696, df = 1, p-value = 0.001774
```

Both the spatial lag and spatial error tests are significant and thus, inconclusive. We turn our attention to robust forms of the models. We find out that the robust forms are also both statistically significant, but the robust lag model has a higher p-value, so we pick the robust spatial error model. The output of this model is included below.

```
Call:errorsarlm(formula = Count_crim ~ as.numeric(Bachelor) + as.numeric(Mean.Annua) +
    as.numeric(Unemp.in.W) + as.numeric(Land_area_), data = chi.poly@data,     listw = W, zero.policy = TRUE)

Residuals:
     Min       1Q   Median       3Q      Max
-861.440 -102.700  -33.037   61.005 2863.269

Type: error
Regions with no neighbours included:
 169
Coefficients: (asymptotic standard errors)
                         Estimate  Std. Error z value  Pr(>|z|)
(Intercept)             3.1273e+02  3.3865e+01  9.2346 < 2.2e-16
as.numeric(Bachelor)   -1.2952e-01  4.8432e-02 -2.6743  0.007488
as.numeric(Mean.Annua) -1.2277e-01  4.6946e-02 -2.6152  0.008918
as.numeric(Unemp.in.W)  1.0709e-01  6.3615e-02  1.6833  0.092312
as.numeric(Land_area_)  7.2038e-05  9.3228e-06  7.7271  1.11e-14

Lambda: 0.654, LR test value: 283.56, p-value: < 2.22e-16
Approximate (numerical Hessian) standard error: 0.031372
    z-value: 20.847, p-value: < 2.22e-16
Wald statistic: 434.59, p-value: < 2.22e-16

Log likelihood: -5444.025 for error model
ML residual variance (sigma squared): 45232, (sigma: 212.68)
Number of observations: 798
Number of parameters estimated: 7
AIC: 10902, (AIC for lm: 11184)
```

First, we notice that lambda, the coefficient of spatially correlated error, is equal to 0.654 and it is statistically significant. This coefficient captures the dependency between tracts, therefore the error model will be improved compared to the OLS model as illustrated by the AIC of 10,902 in lower values than the OLS's AIC of 11,184. The estimates of the predictor's coefficients from the spatial error model are -0.12952 for the education variable, -0.12277 for the income, 0.107 for the unemployment and 0.000072 for land area. The signs of the predictor's coefficients have not changed so our hypothesis still holds with this model. However; all except the control variable land area are smaller in terms of magnitude.

We visualize below the distribution of residuals generated by the spatial error model and notice that the residuals seem less clustered, a sign that spatial dependency has weakened.



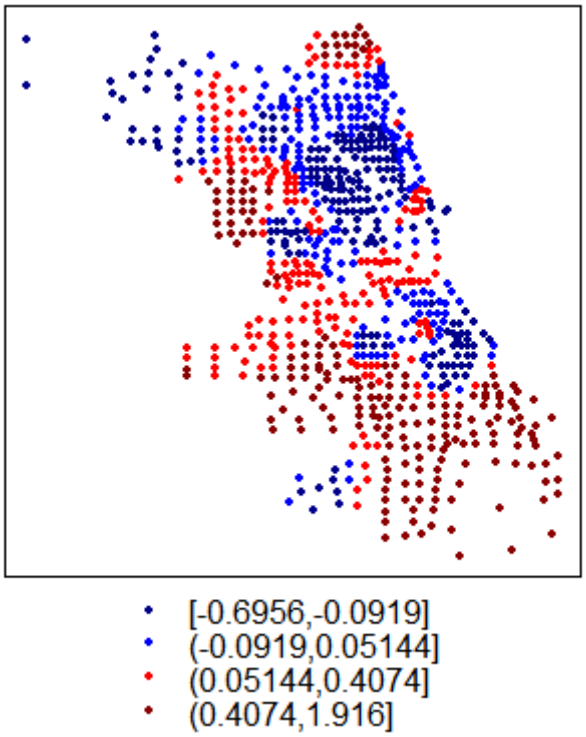
To conclude, a comparison of the residuals maps and of the AICs confirms that the spatial error model is an improvement over the OLS model.

3. Geographically Weighted Regression (GWR) model

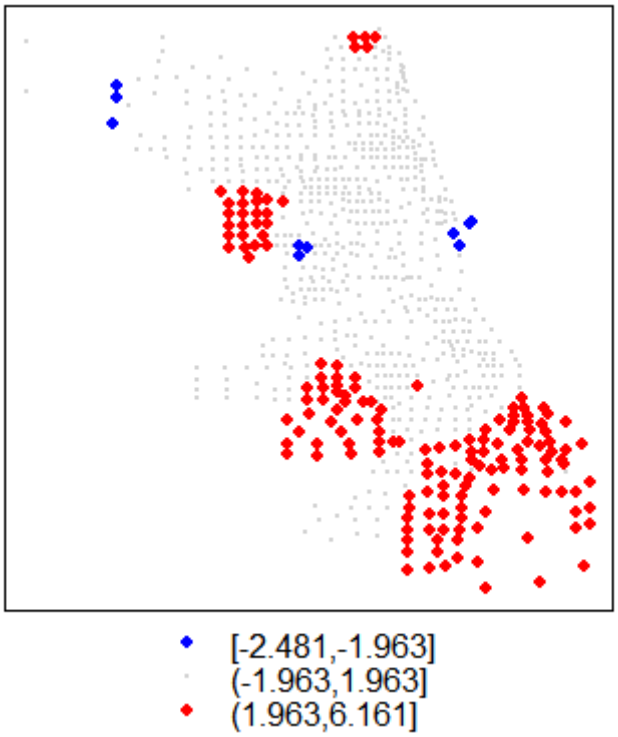
We run a GWR on the Unemployment variable and obtain the below output.

```
Call:
gwr(formula = Count_crim ~ as.numeric(Bachelor) + as.numeric(Unemp.in.W) +
    as.numeric(Land_area_), data = centr, coords = coordinates(centr),
    adapt = GWRbandwidth, hatmatrix = TRUE, se.fit = TRUE)
Kernel function: gwr.Gauss
Adaptive quantile: 0.009984621 (about 7 of 801 data points)
Summary of GWR coefficient estimates at data points:
                            Min.     1st Qu.      Median     3rd Qu.        Max.   Global
X.Intercept.           -8.1683e+01  1.0825e+02  1.6896e+02  2.5607e+02  6.2661e+02 351.5155
as.numeric.Bachelor.   -9.6088e-01 -1.9731e-01 -6.9821e-02 -6.9205e-03  2.4177e-01  -0.3127
as.numeric.Unemp.in.W. -6.9556e-01 -9.1898e-02  5.1442e-02  4.0742e-01  1.9156e+00   0.1615
as.numeric.Land_area_. -2.6808e-05  9.6164e-05  2.3370e-04  3.9614e-04  2.1699e-03   0.0001
Number of data points: 801
Effective number of parameters (residual: 2traceS - traceS'S): 234.4213
Effective degrees of freedom (residual: 2traceS - traceS'S): 566.5787
Sigma (residual: 2traceS - traceS'S): 177.5138
Effective number of parameters (model: traceS): 168.9325
Effective degrees of freedom (model: traceS): 632.0675
Sigma (model: traceS): 168.0662
Sigma (ML): 149.2952
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 10724.7
AIC (GWR p. 96, eq. 4.22): 10461.56
Residual sum of squares: 17853538
Quasi-global R2: 0.7197268
```

Variable Unemployment ranges from -0.696 to 1.916 with a mean of 0.206 and a median of 0.0514. The global estimate is 0.1615. The mean, median and global estimate are positive which supports our hypothesis and previous conclusions that Unemployment should be positively correlated with crime. However, this relationship appears to be negative in some location of Chicago so we map the various GWR coefficients of Unemployment to identify these areas.



The coefficient for Unemployment is positive in the South and West side of Chicago and negative Downtown and Uptown. This means that in Downtown and Uptown, more unemployment causes less crimes. However, these areas of the city have the lowest unemployment rate, so we suspect that these local coefficients may not be statistically significant. To verify this, we plot the statistical significance map below.



We find that only the West and South sides of the city are statistically significant. It confirms our intuition that Unemployment is only significant in areas where the coefficients are positive, and that negative coefficients are implausible. This result does not contradict any prior conclusions we made but it also does not provide any additional insights because the positive coefficients do not vary a lot among the significant tracts.