Spatial Variation in Infant Mortality in Chicago

Spatial Regression Analysis: Spring 2020

Varada Shrotri

Introduction:

The mortality rates in United States has decreased exceptionally from almost 20 deaths per 1000 population in 1930 to roughly 8 deaths per 1000 population in 2020. Despite significant decrease in overall mortality rates through the rears, disparities in mortality have persisted along various dimensions such as gender, race/ethnicity, and geographic area (Yang et. al. 2015).

Losing an infant is devastating for parents, families and communities and can result in extreme and persistent sadness that does not get better with time. Parents that have lost a child also have increased risks of many poor health outcomes such as post-traumatic stress disorder, depression, psychiatric hospitalizations, guilt and heart attacks. In 2017, over 22,000 infants died in the United States. According to the Center for Disease Control and Prevention, the leading causes were birth defects, low birthrate and preterm birth, maternal pregnancy complications, sudden infant death syndrome, injuries. The United States infant mortality rate is consistently higher than other developed countries, 1.5 hours than the average (3.8 deaths per 1000 live births) among Organization for Economic Co-operation and Development (OECD) countries.

Within the United States, significant disparities persist in infant mortality among different racial and ethnic groups, with the most striking disparity between babies born to black women and babies born to white women. Research indicates socioeconomic inequality in the United States is likely a primary contributor to its higher infant mortality, along with reporting methods that differ from state to state. Considerable progress has been made in the United States over the past 50 years to reduce infant mortality, however, more needs to be done.

Income inequality and social capital have driven the recent investigation of the effects of social characteristics on mortality. Although mortality is an ecological and spatial feature of a population in a specific area, a large body of previous mortality research has not incorporated a spatial perspective into investigation of relationship between contextual characteristics and mortality rates (Yang et. al. 2015). The spatial interdependence in mortality rates may be relevant in understanding role of neighbourhood characteristics in mortality disparity. Although prior research makes it known that Chicago's primarily black neighbourhoods have a higher infant mortality, we can also inspect if living in a neighbourhood close to black neighbourhoods makes a difference on infant mortality.

Findings of this and similar research has important implications for policymaking and future research. The identification of the spatial spillover effects of social factors on mortality suggests that by improving social conditions such as increasing the level of social capital in a particular area may not only promote the health of the local population, but also that of the residents nearby.

Literature Review

Studies have discussed the determinants of infant mortality, they have overlooked some issues. First, although infant mortality is an ecological and spatial feature of population of a specific area, a large body of previous mortality research has not incorporated a spatial perspective into the relationship between contextual characteristics and mortality rates (Sparks and Sparks 2010; Sparks et al. 2012; Yang et al. 2011). In this study, I demonstrate the importance of incorporating a spatial perspective in the ecological level mortality research with conventional OLS approach.

In terms of theory, it is necessary to investigate the determinants of infant mortality in context with the impact of explanatory variables in 'adjacent areas'. Most studies explain the infant mortality rate of a given area only with the characteristics *within* this area – reflecting an 'aspatial' perspective. Thus there is a need to incorporate a spatial perspective in to infant mortality research both methodologically and theoretically. The determinants of infant mortality rates of a certain area could be explained not only by the features of this area, but also by the characteristics of the surrounding area. As social processes are spatially embedded, geographic proximity to neighbouring areas should play an important factor in unveiling the spatial mortality dynamics. Figure 1 shows the spatial structure of infant mortality rates in the city of Chicago. The rates are for the 77 neighbourhoods of Chicago which are loosely based on zip codes.

First, I look at the spatial spillover process. Of the various meanings of spatial spillover, growth spillover has the most general meaning and implies that the change in outcome of interest in a unit is related to the behaviours of neighbouring units. Similarly, the relevance of an explicit spatial perspective is recognized as in other social sciences, for example this is illustrated by the use of spatial models and neighbourhood effects in demography and criminology (Anselin 2003; Audretsch 2003). Despite the various meanings of spatial spillover, the core concept is to conceptualize spatial structure and the dynamics among spatial units in which diffusion of ideas, practices, and resources occurs (Capello 2009; Rogers 1995). Following the previous studies that utilized the spatial spillover perspective to regional developmental and economic outcomes, we conceptualize a spatial unit as a geographically limited system in which all necessary resources could not be produced, but those resources exceeding local demands would spill over to nearby units for survival and growth (Capello 2009). Specifically, when a spatial unit exceeds specific resources, it may generate spillover influence to its neighbours. Thus, a spatial unit is neither self-efficient nor isolated, but it is inherently interdependent on its surroundings.

Applying this concept of spatial interdependence to mortality research, the local social and institutional resources that promote population health may exceed local needs and hence generate spatial spillover effects on infant mortality in nearby areas. Drawing from the spillover perspective, I hypothesize that high levels of social and institutional resources that promote health (social capital and accessibility to health care) in an area would spill over to its neighbours, generating a positive effect on reducing the infant mortality rates in adjacent areas.

The goal of this study is to test whether the spatial spillover processes could be used to guide the modelling of geographic variation in Chicago neighbourhoods infant mortality rates by moving beyond the typical theoretical conceptualization of context where a neighbourhoods infant mortality is only associated with its own features. Specifically I investigate whether the mortality rates of a neighbourhood is associated with the features of surrounding neighbourhoods after accounting for its specific characteristics.

Hypothesis: A high level of social capital in a specific area is negatively associated with the mortality rate of that immediate area and its adjacent areas. This means that the social capital within an area will have a 'spatial spillover effect' on its adjacent areas.

Data

All Chicago neighborhood level data is take from Chicago Health Atlas web portal which is maintained by the Chicago Department of Health.

1. Infant Mortality Rate: This is the rate of infant deaths per live 1000 births. This data is extracted by the Health Atlas from the 2012-2016 American Community Survey (ACS) 5 year estimate (US Census Bureau). These are not age-sex standardized rates. It is suggested that age-sex standardized rates are appropriate for ecological mortality research (Kawachi and Blakely 2001). Since racial/ethnic structure is a factor that determines social stratification in the US, racial composition of a neighborhood is used as a dummy variable to account for it.
2. Racial Composition: A neighborhood was deemed to be a black dominated neighborhood if more than 50% of its population identify themselves as persons of colour. It is essential to control for racial composition as it is an important factor in Chicago.
3. Metropolitan Status: : Rural/urban residence has been found to be associated with mortality in the US (McLaughlin et al. 2007; Yang et al. 2011). Specifically, while rural counties are often characterized with poor socioeconomic profiles and access to health care services, their age-sex standardized mortality rates are unexpectedly lower than those of their urban counterparts, creating the so-called "rural paradox" (McLaughlin et al. 2001; Yang et al. 2011). Following the US office of Management and Budget, metro counties are defined as those 50,000. A dummy variable was created to identify between metro (1) and non-metro neighborhoods.
4. Economic Hardship Index: The Chicago Community Area economic hardship index calculation is based on the "Intercity Hardship Index." by Richard P. Nathan

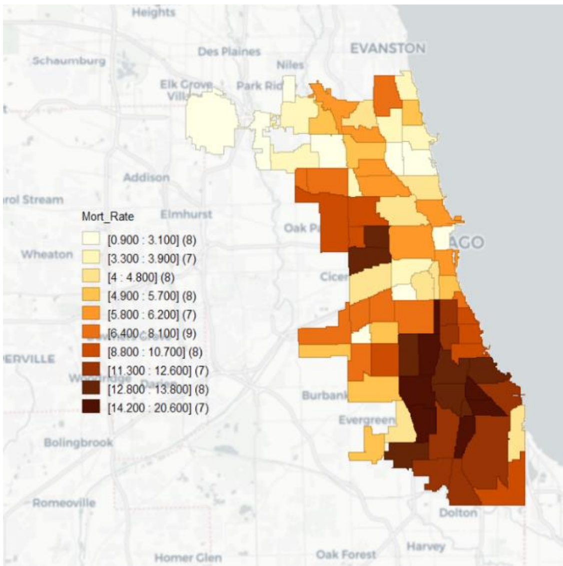
and Charles F. Adams, Jr. in “Understanding Urban Hardship,” Political Science Quarterly 91 (Spring 1976): 47-62. The economic hardship score is the median of six variables that have been standardized on a scale from 0 to 100. The variables are:

a. Unemployment (over the age of 16 years)
b. Education (over 25 years of age without a high school diploma)
c. Per Capita Income
d. Poverty (below the federal poverty level
e. Crowded housing (housing units with more than one person in one room)
f. Dependency (population under 18 or over 64 years of age)

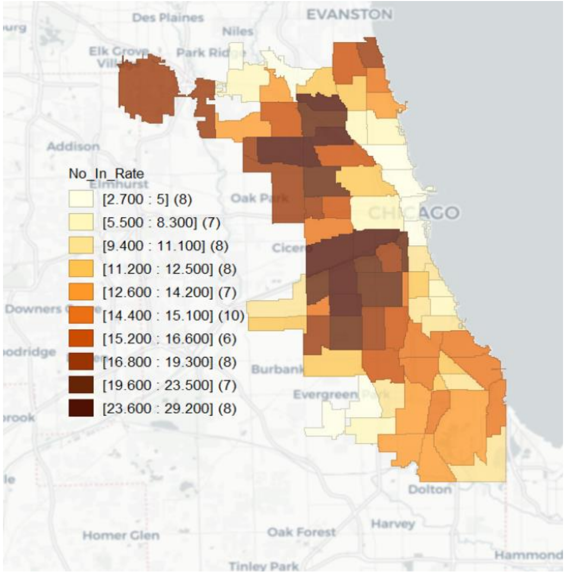
A higher hardship index score indicates worse economic conditions.

5. Food Security: Data for this pertains to percentage of people with low income and living more than half mile from the nearest supermarket, supercenter, or large grocery store among the total population.
6. Health Insurance: Percentage of people with no health insurance coverage among the total civilian noninstitutionalized population.

Spatial Dependence in Variables

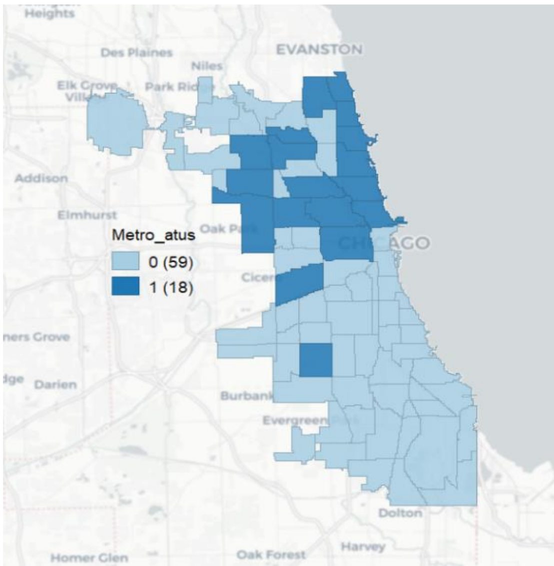


1 - Infant Mortality Rates

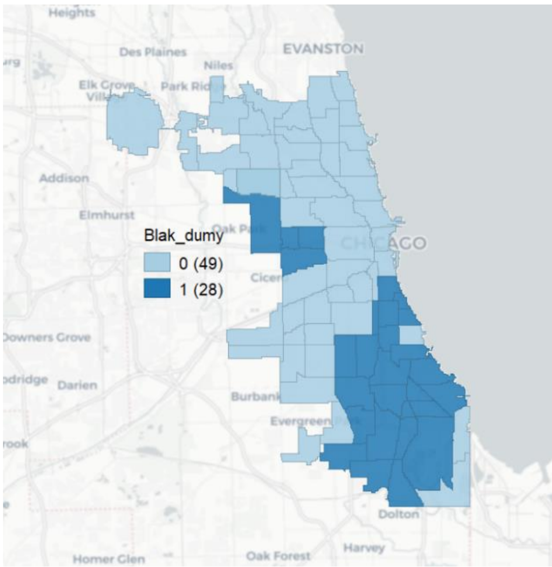


2 – No Health Insurance (Rate)

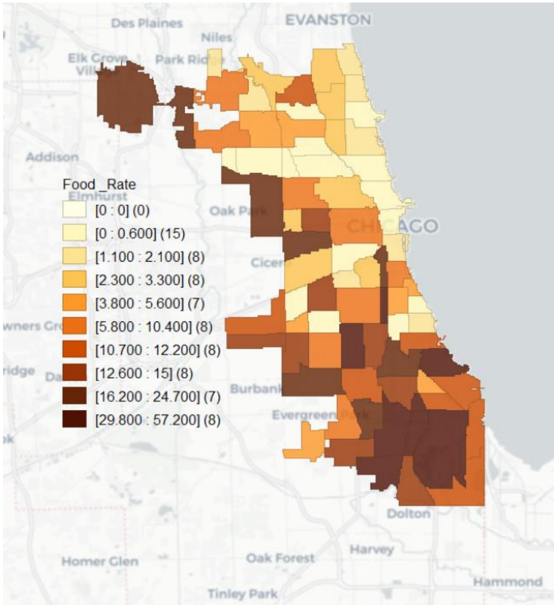
Based on the map, it can be seen that higher infant mortality rates are clustered together on the south-side of Chicago. We can infer that there is some spatial dependence and it is worth exploring how this relationship is exhibited by the independent variables.



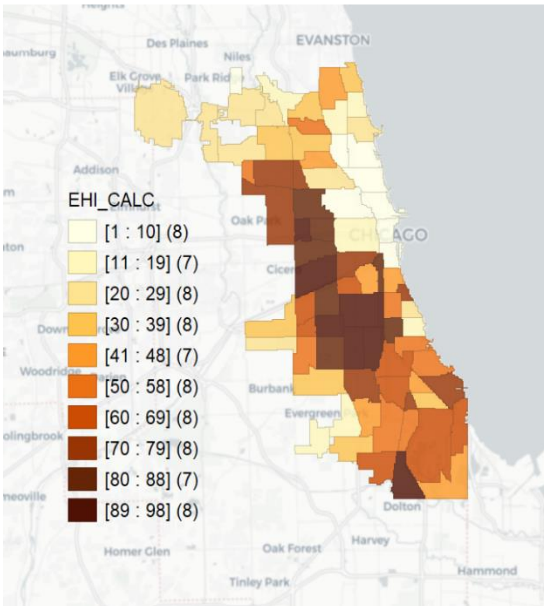
3 – Metropolitan Neighborhoods (shaded dark blue)



4 -Predominantly Black Counties (Shaded Dark Blue)



5 – Food Security

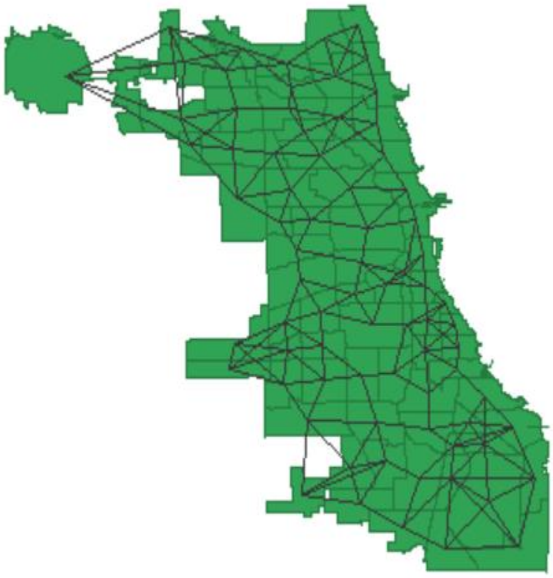


6 – Economic Hardship Index

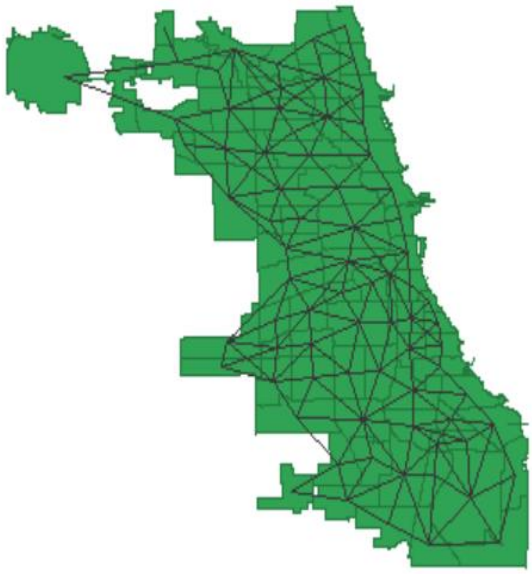
Based on the maps for dependent variables it is evident that infant mortality is lower in metropolitan neighbourhoods which are not predominantly black. Further these neighbourhoods also have better access to food sources, have a lower Economic Hardship Index Score.

Methodology

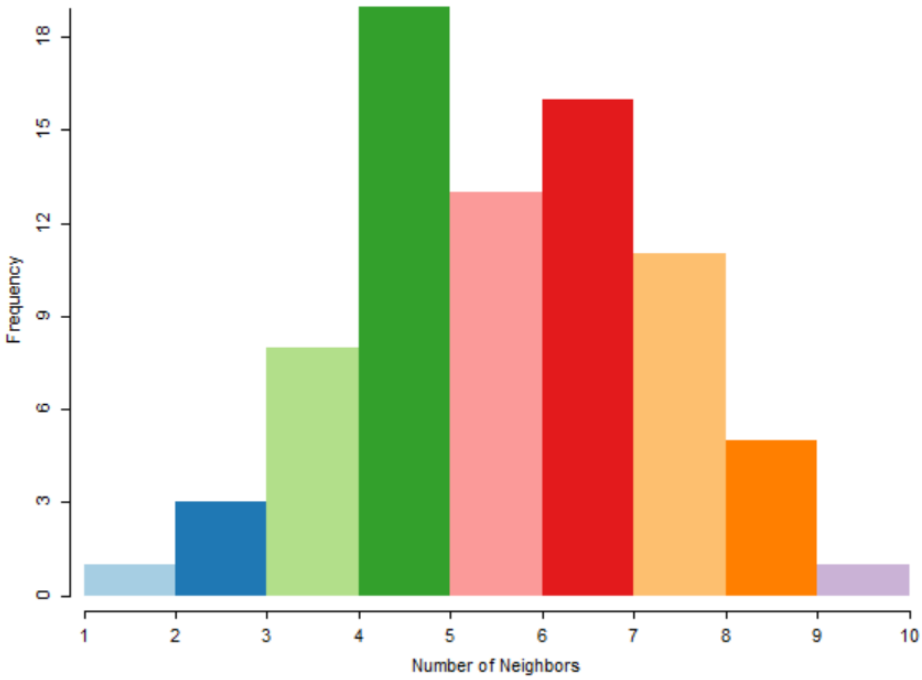
According to Marcos, Jesus and Manuel (2012), the dominant approach for defining a spatial weights matrix employed by practitioners involves an exogenous treatment of the problem. Thus the problem of selecting a weighting matrix is a problem of model selection. I started with a queen contiguity matrix to asses spatial dependence in my model. However, the results of my spatial lag model using the queen matrix were inconclusive. I therefore used a KNN-5 matrix to determine spatial dependence. This provided me conclusive results for my spatial lag model.



KNN 5 Matrix



Queen Matrix



Since the mean number of neighbours that each neighbourhood has is 5.12, I believe using a 5 nearest neighbour matrix would not bias my estimates.

First I ran a simple OLS model with a queen matrix of the following form:

$$\rightarrow Infant\_Mortality = \alpha + \beta_1 EHI\_Index + \beta_2 FoodRate +$$

$$\beta_3 No\ Health\ Ins + \beta_5 \mathrm{black\_dummy} + \beta_6 \mathrm{Metro\_status}$$

The results for the same are in Table 1. The OLS regressions suggests a highly significant relationship between infant mortality in the black dominated neighbourhoods. However none of my other variable are significant and therefore cannot be used to explain the variation in infant mortality rates.

Multicollinearity condition number is 12.241 which does raise a yellow flag that there might be multicollinearity between the independent variables. This could undermine the ability of the model to find significance. However it is well below the 30 cutoff.

Table 1 - Initial OLS Diagnostics

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :IDK WHat 1.dbf
Weights matrix      :IDK WHat 1.shp: distance: Threshold, 7.4986938332
Dependent Variable  :   Mort_Rate              Number of Observations:         77
Mean dependent var  :      8.0169              Number of Variables   :          6
S.D. dependent var  :      4.5855              Degrees of Freedom    :         71
R-squared           :      0.7008
Adjusted R-squared  :      0.6797
Sum squared residual:     478.166              F-statistic           :    33.2563
Sigma-square        :       6.735              Prob(F-statistic)     :    2.5e-17
S.E. of regression  :       2.595              Log likelihood        :   -179.565
Sigma-square ML     :       6.210              Akaike info criterion :    371.130
S.E of regression ML:      2.4920              Schwarz criterion     :    385.193

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       4.0098131       0.8405880       4.7702479       0.0000095
           Blak_dumy       7.3668345       0.9376173       7.8569739       0.0000000
            EHI_CALC       0.0060843       0.0208797       0.2913980       0.7715966
          Food _Rate       0.0320247       0.0295243       1.0846881       0.2817293
         Metro_atus        0.2483032       0.7823409       0.3173850       0.7518830
          No_In_Rate       0.0447046       0.0863431       0.5177552       0.6062386
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER          12.241

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE            PROB
Jarque-Bera                       2          16.616            0.0002
```

The Jaque-Bera test on normality of errors suggests that my data is not normal. The adjusted R square suggests that around 68% of the variation in infant mortality rates can be explained by the independent variable. However since only the Black_Dummy variable is significant it is probable that the variation is mostly explained by only the racial characteristics of a neighbourhood.

Table 2 - Spatial Diagnostics

DIAGNOSTICS FOR HETEROSKEDASTICITY

RANDOM COEFFICIENTS

TEST	DF	VALUE	PROB
Breusch-Pagan test	5	20.932	0.0008
Koenker-Bassett test	5	10.824	0.0550

SPECIFICATION ROBUST TEST

TEST	DF	VALUE	PROB
White	18	21.956	0.2339

DIAGNOSTICS FOR SPATIAL DEPENDENCE

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.0495	2.160	0.0308
Lagrange Multiplier (lag)	1	9.976	0.0016
Robust LM (lag)	1	8.921	0.0028
Lagrange Multiplier (error)	1	1.257	0.2622
Robust LM (error)	1	0.202	0.6534
Lagrange Multiplier (SARMA)	2	10.178	0.0062

The White test and the Koenker-Bassett Test indicate that the data has homoscedastic errors. Therefore we do not need to correct for heteroskedasticity.

The Lagrange Multiplier tests indicate that the model might have a spatial lag, since the p-values for regular of the LM-Test are small enough to reject the null hypothesis that there is a spatial lag in the model.

Based on these results, I run a Spatial Lag model (GMM Method since the data is not normal to use Maximum Likelihood method).

Table 3 - Spatial Lag Model (GMM)

```
--------------------------------------------------
Data set            :IDK WHat 1.dbf
Weights matrix      :IDK WHat 1.shp: distance: Threshold, 7.4986938332
Dependent Variable  :   Mort_Rate               Number of Observations:          77
Mean dependent var  :      8.0169               Number of Variables   :           7
S.D. dependent var  :      4.5855               Degrees of Freedom    :          70
Pseudo R-squared    :      0.7332
Spatial Pseudo R-squared:  0.7382

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       0.8895156       1.1848269       0.7507557       0.4527997
           Blak_dumy       6.3669741       0.8985400       7.0859103       0.0000000
            EHI_CALC      -0.0031133       0.0191215      -0.1628187       0.8706612
          Food _Rate       0.0139095       0.0272856       0.5097745       0.6102094
          Metro_atus       0.9132049       0.7352850       1.2419740       0.2142461
          No_In_Rate       0.0759670       0.0788214       0.9637865       0.3351530
         W_Mort_Rate       0.4286294       0.1246044       3.4399229       0.0005819
------------------------------------------------------------------------------------
Instrumented: W_Mort_Rate
Instruments: W_Blak_dumy, W_EHI_CALC, W_Food _Rate, W_Metro_atus,
             W_No_In_Rate

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                          MI/DF       VALUE           PROB
Anselin-Kelejian Test            1          0.839          0.3597
================================ END OF REPORT =====================================
```

The results using the spatial lag model indicate that the lag (ρ) coefficient is significant (p-value < 0.05). The Anselin-Kelejian Test has a p-value of more than 0.05 thus indicating that the spatial dependence in the model is accounted for by the lag and there is no need to run anything further.

Results

Direct Effects:

An increase of 1 in infant mortality rate is associated with an increase of 0.428 rate in neighbouring 5 neighbourhoods. This is a significant impact indicating that ecological variables like infant mortality rates do impact the rates in other neighbourhoods.

Apart from rho and the variables for black dominated neighbourhoods, no other variables are significant to the analysis. This maybe suggests if I can revise the model and remove the insignificant variables.

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
--------------------------------------------------
Data set            :IDK WHat 1.dbf
Weights matrix      :IDK WHat 1.shp: distance: Threshold, 7.4986938332
Dependent Variable  :   Mort_Rate                Number of Observations:         77
Mean dependent var  :      8.0169                Number of Variables   :          3
S.D. dependent var  :      4.5855                Degrees of Freedom    :         74
Pseudo R-squared    :      0.7180
Spatial Pseudo R-squared:  0.7158

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       2.8887577       0.8834875       3.2697212       0.0010765
           Blak_dumy       6.4690689       0.7542112       8.5772643       0.0000000
         W_Mort_Rate       0.3370159       0.1203684       2.7998712       0.0051123
------------------------------------------------------------------------------------
Instrumented: W_Mort_Rate
Instruments: W_Blak_dumy

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Anselin-Kelejian Test            1          0.059          0.8076
================================ END OF REPORT =====================================
```

Doing so, decreases the adjusted R square, but only marginally. It also makes the constant insignificant. Therefore I stick to the model in Table 3.

Total Effects:

Since there is a spatial lag model in play, the total effects of each tested variable are greater than their direct effects. The effects of a neighbourhoods exogenous variables on its own infant mortality as well as the infant mortality in the neighbouring neighbourhoods can be determined by calculating the β x $^{1}/_{(1-\rho)}$.

Thus an increase in percentage of people who do not have sufficient access to healthy food leads to an increase in infant mortality in that neighbourhood and its neighbouring

5 neighbourhoods. An increase in the percentage of people with no health insurance coverage leads to an increase in infant mortality in that and surrounding neighbourhoods by 0.1328. However, the regression output indicates that Economic Hardship Index is inversely related to infant mortality ie. an increase in EHI index score will reduce the infant mortality.

Since I used pre-calculated EHI values, I am not sure how the data is represented and has what kind of measurement error (if any). This might be one of the many that reasons we end up getting a relationship which does not make economic sense.

Conclusion and Caveats

I examined whether the mortality rates of a neighbourhood is associated with the features of surrounding neighbourhoods after accounting for its specific characteristics. I find that there is certainly a spatial process of dependence going on and that it is worth exploring.

Finding more about this will provide significant insights about the spatial autocorrelation among poor socio-economic indicators, unemployment and inequities in income levels which further manifest into infant mortality rate.

The data pertains to 77 Chicago neighborhoods. Since the data is not normal, I cannot use the Maximum Likelihood estimation. I have used GMM method for estimation. One can argue 77 is not a large sample, thereby contesting the validity of the GMM model. I tried moving to a larger area and studying the geography of the entire Illinois state, however, I have major data constraints while using the ACS sample data. Although I was able to extract data for some variables, I was lost for most variables due to complex nature of over 25,000 codes available. The fact that my data is not large enough remains as a caveat and one needs to be cautious while interpreting the results of the model.

Similarly most of my explanatory variables ended up being insignificant. Using a larger dataset for a bigger geography or using a spatial panel to estimate this model might give a better conclusion.

References

1. Yang, T. C., Noah, A., & Shoff, C. (2015). Exploring geographic variation in US mortality rates using a spatial Durbin approach. *Population, space and place*, *21*(1), 18–37. https://doi.org/10.1002/psp.1809
2. Yang, T., & Jensen, L. (2015). Exploring the Inequality-Mortality Relationship in the US with Bayesian Spatial Modeling. *Population Research and Policy Review, 34*(3), 437-460. Retrieved April 28, 2020, from www.jstor.org/stable/43672138
3. Barufi, A.M., Haddad, E. & Paez, A. Infant mortality in Brazil, 1980-2000: A spatial panel data analysis. *BMC Public Health* **12,** 181 (2012). https://doi.org/10.1186/1471-2458-12-181
4. Herrera, Marcos, Jesus Mur, and Manuel Ruiz Marin. *Selecting the Most Adequate Spatial Weighting Matrix: A Study on Criteria*. University Library of Munich, Germany, 2012.
5. Anselin L. Spatial externalities. International Regional Science Review. 2003;26(2):147–152. [Google Scholar]
6. Audretsch DB. Innovation and spatial externalities. International Regional Science Review. 2003;26(2):167–174. [Google Scholar]

7. Capello R. Spatial spillovers and regional growth: a cognitive approach. European Planning Studies. 2009;17(5):639–658. [Google Scholar]
8. Rogers EM. Diffusion of Innovations. Simon and Schuster; 1995. [Google Scholar]
9. https://www.chicagohealthatlas.org/indicators