Research Proposal: Modeling the Spread of COVID-19 in Greater Chicagoland

SOCI 40217: Spatial Regressions

Cecile Murray

Introduction

Chicago has emerged as one of the nation's "hotspots" in the COVID-19 pandemic. As this disease is transmitted from person to person through physical proximity, any effort to model its spread should account for the spatial relationships between different areas. However, those relationships have different dimensions: in addition to physical adjacency, one might also consider how commuting patterns link spatially disparate areas. I propose to explore the determinants of the disease's prevalence across the Chicago metropolitan area, with a particular focus on the roles of local employment patterns and socio-demographic disparities.

Research design

I want to compare the results of three models predicting the number of COVID deaths in order to better understand how integrating spatial data affects model parameters, inference, and explanatory power. The first model would be a simple OLS regression with no spatial component, while the latter two would be spatial autoregressive (SAR) models, one of which would use physical adjacency to derive spatial weights, and one which would use commuting intensities. I plan to include additional explanatory variables such as population density, age, race/ethnicity, and health insurance coverage rates.

Due to the nature of the available COVID data, the unit of observation will be the ZIP code. While the number of tests, cases, and deaths are each important outcome variables, I plan to focus on the number of deaths because it is less likely to suffer from non-random measurement error. Finally, I do not have data on county- and city-level social distancing measures, but I will rely on the fact that the statewide stay-at-home order mandating the closure of most businesses applies uniformly across all Illinois ZIP codes.

Data

This analysis will draw on three datasets to capture COVID-19 prevalence, local employment characteristics, and socio-demographic disparities. Additionally, I will use ZIP code shapefiles available from the U.S. Census Bureau's TIGER database.[1]

[1] https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html

First, the Illinois Department of Public Health (ILDPH) maintains a ZIP-code level database of total coronavirus cases, tests, and deaths.[2] The data are updated daily and published in a table on the ILDPH website. Unfortunately, there are two barriers to acquiring time-series data. Specifically, data are presented cumulatively so that each day's update overwrites the last, and the data change as new cases are investigated, making it difficult to correctly align a case total with a specific date. Consequently, I plan to use the total counts for the initial approach. Should the project end up requiring it or if time allows, I will consider writing a basic web scraper to collect the data each day.

Second, I will use data from the U.S. Census Bureau's Longitudinal Employer Household Dynamics (LEHD) program to identify connections between ZIP codes based on commuting relationships.[3] The LEHD program's Origin-Destination Employment Statistics (LODES) offer Census-block-level counts of workers by home and work locations, as well as additional information about worker age, education, income, and major industry. The data are derived from state unemployment insurance (UI) records: while most workers are covered by UI, the data exclude self-employed individuals. The most recent data available are for 2017. I will need to aggregate the block-level counts of workers to the ZIP code level in order to leverage them for the spatial weights matrix, which I will do by performing a polygon-to-polygon spatial join in R.

Third, the Census Bureau's American Community Survey (ACS) provides detailed socio-economic and demographic data that I will use to explore the role of socio-demographic disparities. These data are collected from a survey of about 3 million households annually and contain information about income, age race and ethnicity, and health insurance status. Given that the disease has a higher mortality rate for older individuals and that it has afflicted Chicago's minority communities especially badly, I am particularly interested in including age variables, such as the share of a ZIP code's population that is over age 60, and race/ethnicity variables, such as the share of a ZIP code's population that is Black or Hispanic. Also, since access to health care is uneven and individuals who do not have insurance might delay seeking care for COVID symptoms, I plan to include a measure of the uninsured rate. Finally, given recent work suggesting that public transportation use may have contributed to the spread of COVID-19 in New York City, I will include the percentage of workers who commute by transit.[4]

---

[2] https://www.dph.illinois.gov/covid19/covid19-statistics

[3] https://lehd.ces.census.gov/data/

[4] Jeffrey E. Harris, The Subways Seeded the Massive Coronavirus Epidemic in New York City, NBER, https://www.nber.org/papers/w27021.