Correlating Opioid Prescriptions over Space and Time

S610 Final Project

David Terner*

Jiacheng Zhong†

Fall 2019

GitHub Repository : https://github.com/davidmterneriu/S610_Final_Project.git

Contents

*Department of Economics, Indiana University Bloomington

†Department of Operations and Decision Technologies, Indiana University Bloomington

1 Proposal Outline

1.1 Data Description

1. Medicare Part D Opioid Prescribing Data: Annual data, each year from 2013-2017, of opioid prescription rates at the state, county, and ZIP code levels. The data are de-identified. Opioid prescribing rates are derived using data from Medicare Part D claims prescribed by health care providers.

2. Medicaid Opioid Prescribing Data: Similar to the Medicare data above, the Medicaid data are about de-identified Medicaid opioid claims between 2013 and 2017. However, the data are only at the state level. Opioid prescribing rates are derived using Medicaid data on prescription drugs prescribed by health care providers and reported by states to the Centers for Medicare & Medicaid Services (CMS). (*More granular data at the county and ZIP code levels maybe available for this project.*)

3. The U.S. Income and Unemployment Data: For every county in a state, the median household incomes and unemployment rates are available from 2013 to 2017.

4. The U.S. Population Data: For every county in a state, the population are available from 2013 to 2017 at the county and Zip code levels.

5. "zipcode" R-Package: The package helps identify a distance between any two counties.[1]

1.2 Methods

1.2.1 Moran's I

Moran's I is defined as

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{1}$$

where x_i is the variable of interest at location i, N is the number of unique locations across indices i and j, w_{ij} is some measure of distance from location i to j (note $w_{ii} = 0$), and $W = \sum_i \sum_j w_{ij}$.[2] (1) is really nothing more than a distance-adjusted correlation statistic. The appropriate test statistic for I is:

$$z_I = \frac{I - \mathbb{E}[I]}{\sqrt{V[I]}} \tag{2}$$

where

$$\mathbb{E}[I] = -\frac{1}{N-1} \quad \text{and } V(I) = \underbrace{\frac{NS_4 - S_3 S_5}{(N-1)(N-2)(N-3)W^2}}_{=\mathbb{E}[I^2]} - (\mathbb{E}[I])^2$$

and furthermore

$$S_1 = \frac{1}{2} \sum_i \sum_j (w_{ij} + w_{ji})^2 \qquad S_4 = (N^2 - 3N + 3)S_1 + NS_2 + 3W^2$$

$$S_2 = \sum_i \left(\sum_j w_{ij} + \sum_j w_{ji} \right)^2 \qquad S_5 = (N^2 - N)S_1 = 2NS_2 + 6W^2$$

$$S_3 = \frac{\frac{1}{N} \sum_i (x_i - \bar{x})^4}{\left(\frac{1}{N} \sum_i (x_i - \bar{x})^2 \right)^2}$$

1. We use the `housingData` library instead, which contains the county centroid coordinates (alliteration not intended).
2. See Moran 1950.

1.2.2 Distance Approach 1

Clearly, I is sensitive to the choice of distance measure and the corresponding weight factor w_{ij}. If we thought of distance simply as adjacency, then the distance matrix becomes quite sparse, where w_{ij} is unity if locations i and j are adjacent, zero otherwise. However, given how our data are defined and reasonable modeling preferences, we will treat distance as a continuous variable. Taking the longitude/latitude coordinates of i and j, we will compute great-circle distance d_{ij} and then take the inverse: [3] thus, we define w_{ij} as

$$w_{ij} = \begin{cases} \frac{1}{d_{ij}} & i \neq j \\ 0 & \text{o.w.} \end{cases} \tag{3}$$

1.2.3 Distance Approach 2

Following the example of Oden 1995, we will adjust distance weight w_{ij} by population weight. In particular,

$$w_{ij}^{P} = \begin{cases} e^{-4\left(\frac{d_{ij}}{k_i}\right)^2} & i \neq j \\ 0 & \text{o.w.} \end{cases} \tag{4}$$

where d_{ij} is defined as before, $k_i = d_{im_i}$ and $m_i = \max\{j : u_{j(i)} \leq \lambda\}$. $u_{(j)i}$ is the total population of geographic unit i and all of its j neighbors. λ is effectively a measure of spatial clustering: larger (respectively smaller) values of λ give more weight to larger (respectively smaller) population clusters.[4]

1.2.4 Global vs Local

With either (3) or (4), I is a global measure of closely related variables are over all distances. However, it stands to reason that direction and strength of the correlation between units i and j would be different if $d_{ij} = 50$ miles versus $d_{ij} = 500$. For this reason, we will vary how d_{ij} is calculated via

$$\hat{d_{ij}} = \begin{cases} d_{ij} & \text{if } d_{ij} \leq \bar{d} \\ 0 & \text{o.w.} \end{cases} \tag{5}$$

where $\bar{d}$ is the maximum distance specified by the user. For the purposes of this project, $\bar{d}_k$ will be the k^{th} decile of the distance distribution (across all locations i and j).

1.3 Proposal

Given our Medicaid & Medicare Schedule D data on Opioid claims at the county level, we will:

1. Construct generic functions that compute (1), (7), (3), (4), and (5) given latitude/longitude coordinates and some variable of interest.
2. For each year in our sample, we will:
   (a) Determine Moran's I and test-stat/p-value for each:
      i. Distance approach; and,
      ii. Distance decile $\bar{d}_k$
   (b) Plot the results.
3. Determine Moran's I for both the Opioid prescriptions and unemployment rates. In the case of Opioid, determining regions sharing the same Moran's I. Repeat the analysis for the case of the employment rates. Given a Moran's I during a time period, determine whether the two types of regions overlap. Most interestingly, how do unemployment and opioid prescription interact over time (Hollingsworth, Ruhm, and Simon 2017).
4. Analyze and summarize our findings.

3. Depending on the instructor's request, we can either directly implement great circle distance functions from `geosphere` or code from scratch some geodesic distance function by hand; we chose to code distance by hand.

4. In our implementation of (3), we ending up changing how d_{ij}/k_i was exponentiated: cf (12). We made this change since (3), left unchanged, assigned essentially zero weight to most observations.

2 Additional Methods

2.1 New Measures

After the initial proposal, we decided to incorporate two new measures of spatial auto-correlation: Geary's C and Getis Ord's G statistic.

2.1.1 Geary's C

Geary's C is defined as

$$C = \frac{(N-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{2\Big(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\Big)\sum_{i=1}^{n}(x_i - \bar{x})} \tag{6}$$

Whereas (1)'s cross-product terms (in the numerator) are based on *deviations from the mean*, Geary's C cross-products involve differences between observed values per location pair. By design, $C \in [0, 2]$ and has essentially the opposite interpretation as I: $C \to 0$ indicates perfect *positive* autocorrelation, $C \to 2$ indicates perfect *negative* autocorrelation, and $C = 1$ indicates random clustering. C's significance testing involves: E[C] and Var[C] and has a z-score of:

$$z_C = \frac{C - \mathbb{E}[C]}{\sqrt{\mathbb{V}\text{ar}[C]}} \tag{7}$$

where

$$\mathbb{E}[C] = -\frac{1}{N-1} \quad \text{and } \mathbb{V}\text{ar}[C] = \frac{(2S_1 + S_2)(N-1) - 4S_0^2}{2(n-1)S_0^2}$$

and furthermore

$$S_0 = \sum_{i}^{N}\sum_{j}^{N} w_{ij} \quad S_1 = \frac{\sum_{i}^{N}\sum_{j}^{N}(w_{ij} + w_{ji})^2}{2} \quad \text{and } S_2 = \sum_{i}^{N}(w_{i\cdot} + w_{\cdot i})^2$$

2.1.2 Getis Ord's G

Getis Ord's G is distinct from (1) and (6) as G is used for "hot spot" analysis. While I and C provide a measure of clustering, neither can describe the *kind* of clustering presence. For example, consider two regions A and B each with n sub-regions. If all A_n (sub)-regions were all high income areas whereas all B_n regions were low income areas, then $I(A) \approx I(B) \approx 1$ and $C(A) \approx C(B) \approx 0$. In contrast, $G(A) \neq G(B)$ as G is meant to distinguish between regions with a varying degrees of spatial concentration of values *and* differences in levels of values. G large (respectively small) signifies high (respectively small) values clustering together. When G is used in its general form, as defined in (8), the idea is to see how much much G varies from its expected value.

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j} \; \forall j \neq i \tag{8}$$

G's z-score is computed by

$$z_G = \frac{G - \mathbb{E}[G]}{\sqrt{V[G]}}$$

where

$$\mathbb{E}[G] = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}{n(n-1)} \; \forall j \neq i \quad \text{and } V[G] = \mathbb{E}[G^2] - \mathbb{E}[G]^2$$

and furthermore:

$$\mathbb{E}[G^2] = \frac{A+B}{C} \qquad D_2 = -\left[2nS_1 - (n+3)S_2 + 6W^2\right]$$

$$A = D_0\Big(\sum_{i=1}^n x_i^2\Big)^2 + D_1\sum_{i=1}^n x_i^4 + D_2\Big(\sum_{i=1}^n x_i\Big)^2\sum_{i=1}^n x_i^2 \qquad D_3 = 4(n-1)S_1 - 2(n+1)S_2 + 8W^2$$

$$B = D_3\sum_{i=1}^n x_i\sum_{i=1}^n x_i^3 + D_4\Big(\sum_{i=1}^n x_i\Big)^4 \qquad D_4 = S_1 - S_2 + W^2$$

$$C = \Big[\Big(\sum_{i=1}^n x_i\Big)^2 - \sum_{i=1}^n x_i^2\Big]^2 n(n-1)(n-2)(n-3) \qquad W = \sum_{i=1}^n\sum_{\substack{j=1\\ j\neq i}}^n w_{ij}$$

$$D_0 = (n^2-3n+3)S_1 - nS_2 + 3W^2 \qquad S_1 = \frac{1}{2}\sum_{i=1}^n\sum_{\substack{j=1\\ j\neq i}}^n (w_{ij}+w_{ji})^2$$

$$D_1 = -\left[(n^2-n)S_1 - 2nS_2 + 6W^2\right] \qquad S_2 = \sum_{i=1}^n\Big(\sum_{\substack{j=1\\ j\neq i}}^n w_{ij} + \sum_{j=1}^n w_{ij}\Big)^2$$

2.2 Local Versions

The previous section, Global vs Local, discusses how we take global measures I, C, and G and vary the number of observations via capping the allowable maximum distances between geographies. For example, if we were to set $\bar{d} = 500$ miles, then our estimate of I computed from weight matrix W (adjusted for $\bar{d} = 500$) would indicate the degree of spatial autocorrelation across all geographic pairs (i, j) that are no more than 500 miles apart. Moreover, if $\bar{d}$ were much larger (respectively smaller), then our corresponding I would reflect the degree of spatial autocorrelation across a larger (respectively smaller) set of of larger (respectively smaller) *ranged* distance pairs. However, irrespective how $\bar{d}$ is parameterized, I, C and G will only produce one globalized measure.

For this reason, we develop and implement localized versions of each of the three autocorrelation measures; i.e. if a sample of geographies has n observations, then any of the following localized autocorrelation functions will yield $n \times 1$ estimates. The choice of $\bar{d}$ is still important but now takes on a different role. For instance, if we were to estimate the localized Moran's I for location i, as in (9), we still need to determine which (and the number of) neighbors to compare i against. Because each of the localized measures are implemented for only one year across all n geographies, $\bar{d}$ must be set such that for all locations i, i has at *least two* neighbors.[5]

2.2.1 Local Moran

Define Local Moran for location i as:

$$I_i = \frac{x_i - \bar{x}}{S_i^2}\sum_{\substack{j=1\\ j\neq i}}^n w_{ij}(x_j - \bar{x}) \quad \text{given } S_i^2 = \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^n w_{ij} - \bar{x}^2 \tag{9}$$

I_i, unlike I, can take on values outside ± 1, while higher (respectively smaller) values of I_i continue to correspond to stronger positive (respectively negative) spatial autocorrelation. Under the null hypothesis (e.g. randomization), the expected value and variance of I_i are given by:

$$\mathbb{E}[I_i] = \frac{1}{n-1}\sum_{\substack{j=1\\ i\neq j}}^n w_{ij} \quad \text{and } V[I_i] = \frac{n-b_2}{n-1}\sum_{\substack{j=1\\ j\neq i}}^n w_{ij}^2 - \frac{(2b_2-n)}{(n-1)(n-2)}\sum_{\substack{k=1\\ k\neq i}}^n\sum_{\substack{k=1\\ k\neq i}}^n w_{ik}w_{ik} - \Big(\mathbb{E}[I_i]\Big)^2$$

where $b_2 = n\sum_{\substack{i=1\\ i\neq j}}^n (x_i - \bar{x})^4 \times \Big(\sum_{\substack{i=1\\ i\neq j}}^n (x_i - \bar{x})^2\Big)^{-2}$.

5. Our code could be changed to set $\bar{d}$ in such a way to $m \geq 2$ neighbors. If $m < 2$, then the total number of geographies used in localized measure computations will be less than 3 and estimated means/variances will be undefined.

2.2.2 Local Geary's C

The local version of Geary's C, denoted as C_i, is defined as:[6]

$$C_i = \frac{1}{2}\sum_{j=1}^{n} w_{ij}(z_i - z_j)^2 \tag{10}$$

Equation 10 is presented as a standardized z-score. For this particular function, we perform the standard normal transformation of our data. The formula differs from Anselin 1995 by adjusting the test-statistic by a factor of 2; under the null hypothesis, $\mathbb{E}[C_i] = 1$. In this way, there is much stronger degree of interpretive similarity between C and C_i.

2.2.3 Local Getis-Ord G

The localized Getis-Ord G_i is better suited to "hot spot" analysis compared to its global counterpart G.

$$G_i = \frac{\sum_{j=1}^{n} w_{ij}x_j - \bar{x}\sum_{j=1}^{n} w_{ij}}{S\sqrt{\frac{1}{n-1}\left(n\sum_{j=1}^{n} w_{ij}^2 - \left(\sum_{j=1}^{n} w_{ij}\right)^2\right)}} \tag{11}$$

given

$$S = \sqrt{\frac{1}{n}\sum_{j=1}^{n} x_j^2 - (\bar{x})^2}$$

Equation 11 is a standardized z-score. Getis and Ord 2010 asserts that under particular conditions (e.g. $n \to \infty$), this version of G_i is distributed such that $G_i \sim \mathcal{N}(0, 1)$. For this reason, we compute the one tailed p.test for each G_i for hypothesis testing.[7]

3 Results

3.1 Data Work

One of the first judgment calls with regard to the data was deciding the geographic unit. Due to data limitations—e.g. our study would have been far more worthwhile if it had been performed at the zip-code or census track level—we choose US counties (their FIPS code in particular) as our geographic identifier. This data limitation proved to be something of a blessing in disguise. Given n unique geographies, computing the distance matrix $D_{n\times n}$ requires $\frac{n(n-1)}{2}$ computations and so is $\mathcal{O}(n^2)$. A number of key data changes were made relative to what was presented in the Data Description section. In particular, we

1. **Used `housingData` for county coordinates**.
   Median geographical coordinates of U.S. county centroids were present in this new library within the `geoCounty` dataframe. Counties present in `geoCounty` were marked by US census FIPS codes and had other identifiers that made later map construction/visualizations easier.

2. **Restricted our attention to OPR and Unemployment Rates.**
   The main variable of interest in our analysis was the Medicare Part D Opioid Presecriping Rates (OPR hereafter). OPR is defined as the number of Opioid Claims divided by Overall Claims and multiplied by 100, and for our purposes, is listed at the county level for each year in $t \in \{2013, \ldots, 2017\}$. Population data come from the U.S. Deptartment of Agriculture (USDA).[8] Finally, we narrow our spatial analysis of labor data to only include unemployment rates (UR hereafter) at the county level for each year in our sample; data come from the USDA.[9]

6. Please refer to:
https://www.biomedware.com/files/documentation/spacestat/Statistics/Gearys_C/Geary_s_C_statistic.htm.
7. Two tailed testing would be less useful, given the hot-cold spot interpretation depending on the sign of G_i.
8. See the following for the download:
https://www.ers.usda.gov/webdocs/DataFiles/48747/PopulationEstimates.xls?v=2561.3
9. See: https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/.

Another challenge that we encountered was merging our OPR data with unemployment data in such a way to keep as many observations as possible. Looking first to our OPR data, of the unique 3,027 FIPS counties, 3,005 were present across all five years of our sample. However, when we merge unemployment data, the only 1,163 counties (out of a reduced 3,024) remain across all years.

Year	OPR	OPR+UR
2013	3,018	3,015
2014	3,018	3,015
2015	3,017	3,014
2016	3,017	1,189
2017	3,017	3,011

Table 1: FIPS counts

Table 1 summarizes the number of observations per year for each dataset (e.g. OPR only and OPR+UR), whereas Figures 1 and 2 depict the number of shared observations per dataset.

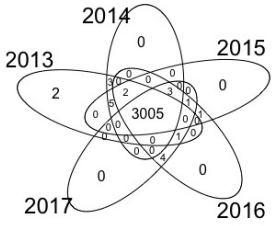

Figure 1: OPR data only

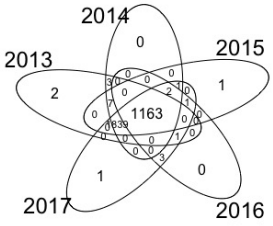

Figure 2: OPR +UR data

Going forward, our analysis will be based on two main datasets: `OPR_test_data` has OPR rates for each of the 3,005 consistent counties for all 5 years ($3,005 \times 5 = 15,025$ observations); and, `test_data_unemployment_new` which is the unbalanced panel of OPR+UR data.[10] `OPR_test_data` is used to test out our time varying functions whereas `test_data_unemployment_new` is used when correlating OPR and UR data for a given year (e.g. 2013).

3.2 Plots

3.2.1 `OPR_test_data` Work

Figure 3 was created using `OPR_test_data`'s 3,005 counties and depicts the distribution of distances between counties.[11]. The smallest (largest) distance between counties was 4.82 (2832.58) miles, between Arlington County Virginia and the District of Columbia (Washington County Maine and San Mateo County California) . The 25th, median, and 75th distance percentiles were 471.40, 756.90, and 1,115.60 miles, respectively. Computing the distance matrix D often was the most computationally intensive task in our work; for example, creating the distance matrix for `OPR_test_data`'s 3,005 counties took approximately 93 seconds. For this reason, we computed only one distance matrix for all analyses/plots which incorporated the same geographies.

10. This dataset also has county populations per year.
11. Note that Figure 3 depicts the distribution of D_{ij} $i \neq j$ such that $D_{ij} \neq 0$.

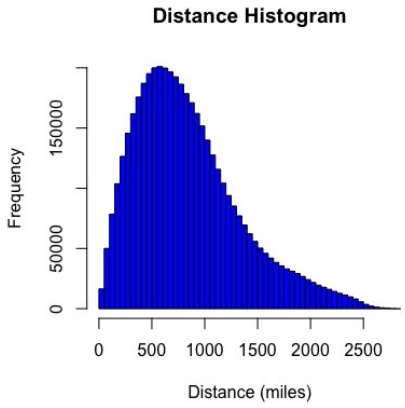

Figure 3

The next plot, Figure 4, depicts how Moran's I changes over different $\bar{d}$ and over time for OPR data only. We choose a sequence of $\bar{d}$ according to an exponential sequence defined here.[12] Figure 4 includes a "zoomed-in" panel as a better visualization of $I(\bar{d})$ for smaller $\bar{d}$. Both Figure 4 and Table 2 indicate that over time: (1) I in expectation falls (more spatial randomness); (2) I becomes less dispersed; and, (3) I reaches its max around 24.7 miles for all years.

Year	Max	Mean	Sd
2013	0.431	0.139	0.0943
2014	0.361	0.116	0.0753
2015	0.355	0.107	0.0689
2016	0.373	0.103	0.0691
2017	0.345	0.0945	0.0640

Table 2: Moran I (OPR) Summary

Figure 5 shows how Moran's I (OPR) is changes if we use population adjusted weights using 2013 data only.[13] We used the same sequence of distance bandwidths as with Figure 4. Our choice for population parameter λ followed from using the 10th/50th/90th percentiles of 2013 county population data. Similar to the non-population adjusted $I(d)$'s, $I(d;\lambda)$ maxes out for $\bar{d} \approx 25$ miles across all λ. Moreover, I seems to be (almost) monotonically increasing in λ. $\lambda_{0.5}$ and $\lambda_{0.9}$ produce essentially indistinguishable Is.

12. We set the minimum distance at 20 miles, the max $\bar{d} = 2832.58$ (i.e. the max distance of D), and the exponential tuning parameter $\theta = 2$. We used this kind of exponential "grid" to compute more instances of I for smaller $\bar{d}$ since I is likely to vary more for small $\Delta\bar{d}$ when $\bar{d}$ is small relative to the same $\Delta\bar{d}$ for larger $\bar{d}$.

13. The data used for Figure 5 came from `test_data_unemployment_new`.

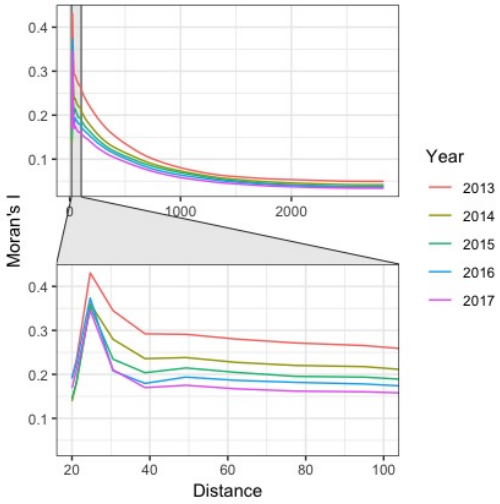

Figure 4: Moran I varying $\bar{d}$ over time

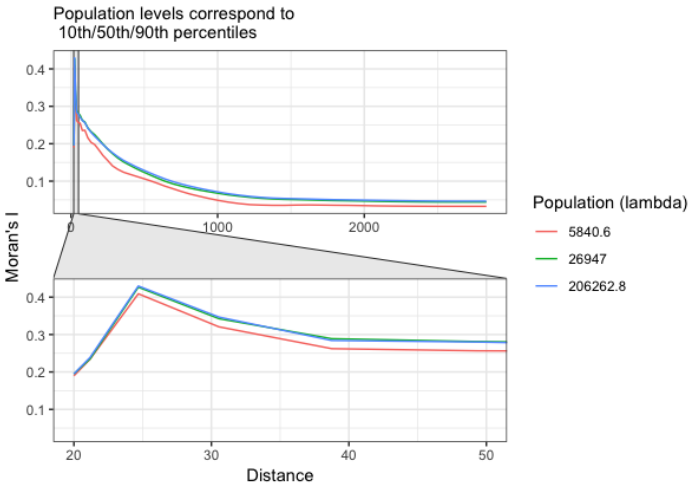

Figure 5: Moran I varying $\bar{d}$ w/ λ.

Local Moran's I_i—as with virtually all local versions of our spatial autocorrelation statistic—took a considerably long time to calculate. Our computation of `OPR_test_data`'s OPR observations from 2013 required 38 minutes to run. We choose a $\bar{d} = 120$ miles. In plain English, this means that for each county i, we calculated (9) given i and $j \neq i$ counties that were 120 miles away.

County	**State**	**OPR**	**Local Moran I**
Clay County	Georgia	8.1	1,296
Shelby County	Alabama	9.56	1,234
Hardin County	Ohio	4.02	484
Sebastian County	Arkansas	6.25	451
Franklin County	Arkansas	5.66	259
⋮	⋮	⋮	⋮
Highland County	Virginia	6.57	-245
Perry County	Ohio	5.23	-273
Okmulgee County	Oklahoma	6.66	-312
Osceola County	Michigan	8.36	-840
Christian County	Missouri	9.18	-937

Table 3: Extreme Local Moran's

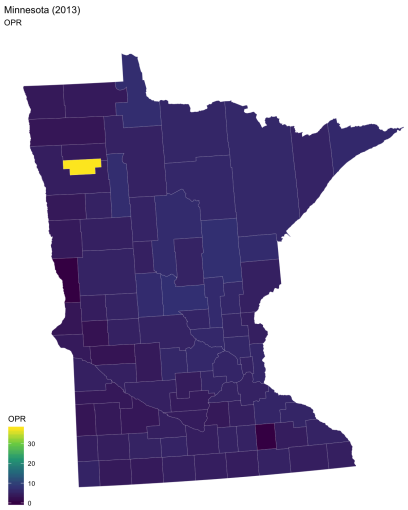

Figure 6: OPR data only

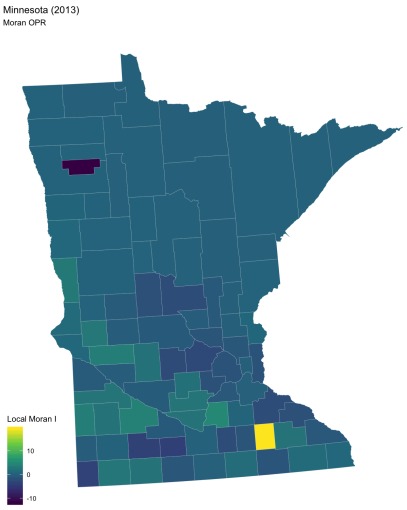

Figure 7: Local Moran OPR data

Coupling Table 2 and Figure 8 findings together, the presence of I_i outliers makes visualizing our results challenging. Figure 8 partially remedies this concern by only color coding counties i such that $|I_i| < 10$. However, even this modification is insufficient to meaningfully distinguish local variation in I_i. For this reason, Figures 6 and 7 were included. Note that the bright yellow county in Figure 6 corresponds to Red Lake County. In 2013, Red Lake County's OPR was 37.61. Since most of Red Lake County's neighbors had much lower OPR, Red Lake County's I_i score was very low (the lowest in the state) at around -12; Figure 7 indicates this disparity as Red Lake County is filled in with a very dark hue to indicate strong negative spatial assortment. Conversely, the bright yellow county in Figure 7 is Dodge County. Dodge County had OPR of ≈ 0, and since most of Dodge County's neighbors also had low OPR, Dodge County's I_i score was very positive at around 19.

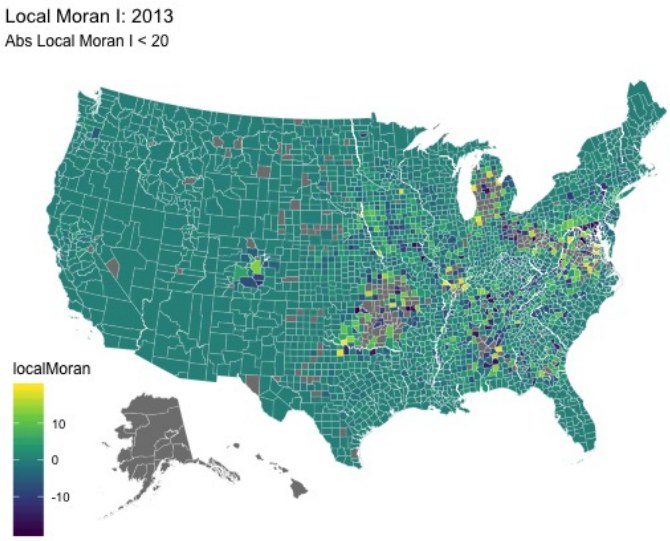

Figure 8: I_i OPR Scores for US

3.2.2 `test_data_unemployment_new` Work

- **Global Geary's C**

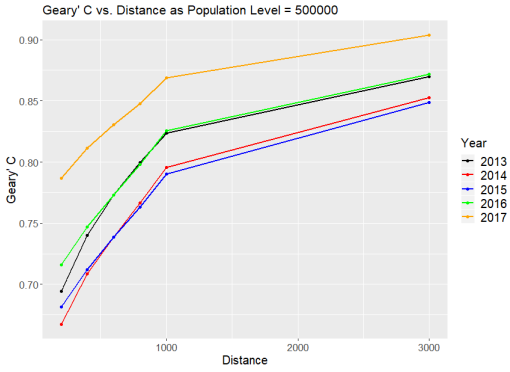

Figure 9: Geary's C for Opioid Prescription Rate

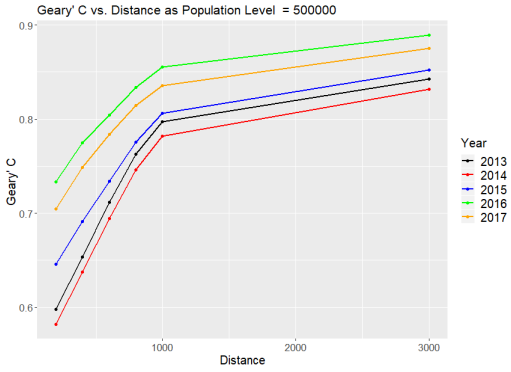

Figure 10: Geary's C for Unemployment Rate

The two plots above show a relationship between the global Geary's C and $\bar{d}$ given $\lambda = 500{,}000$ for the opioid prescription rate and unemployment rate. On average, the strongest positive spatial correlation occurred in Year 2017 for the OPR and Year 2016 for the Unemployment Rate. In each year, the Geary's C is monotonically increasing in $\bar{d}$. Recall that $C \to 0$ signifies increasingly strong positive spatial autocorrelation whereas $C \to 1$ signifies no spatial autocorrelation. Since $I \propto 1 - C$, this explains the visual contrast between Figures 4 and Figure 9 when it comes to OPR data. Curiously, while 4 indicates that OPR become less spatially concentrated over time (almost in a uniform way), 9 indicates no such clear pattern.

- **Hot Spot Detection?**

We decided to test out G_i over the general G stat since G_i is used for "hot spot" analysis. In our OPR context, our G_i estimates can be used to identify counties that, along with their immediate neighbors, are characterized by a high prevalence of OPR relative to the larger region. In short, these counties could be thought of as OPR "hot spots." Likewise, G_i can be used to detect "cold spots," or counties that, along with their immediate neighbors, have a low prevalence of OPR rates relative to the larger region. County i's neighbors were defined as being no more than $\bar{d} = 100$ miles away.

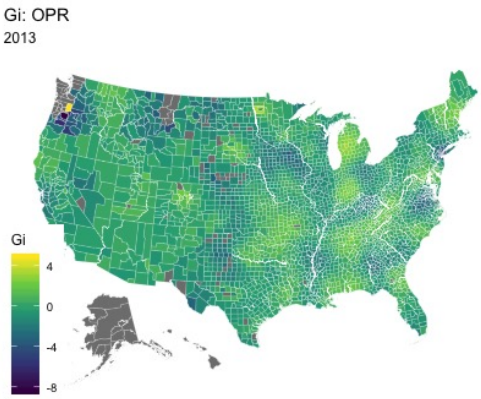

Figure 11: G_i OPR

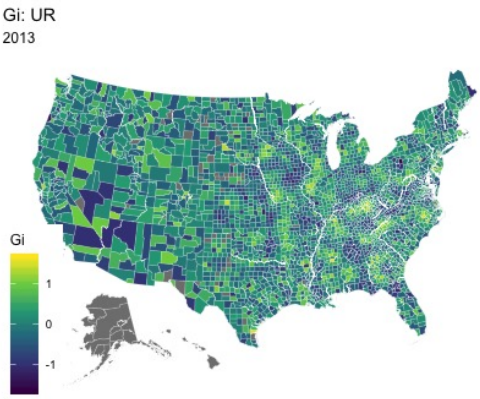

Figure 12: G_i UR

Figures 11 and 12 depict G_i scores for 2013 OPR and unemployment rates (e.g. UR), respectively. Recall that the lighter colors in this context correspond to "hotter" spots whereas darker colors indicated "colder" spots. One striking comparison between Figures 11 and 12 is the difference in visual smoothness; by this, we mean that the color gradient for OPR G_i is less "choppy" relative to UR G_i. If the reader will indulge in something of a topographic analogy, OPR extreme (i.e. very hot or very cold) spots tend to clump together at a regional level so that plateaus and depressions have a gradual ascent/descent. In contrast, UR G_i scores might be conceptualized as particularly steep mountains and plunging slot canyons interspersed amongst uneven terrain.

- **Changes over time?**

Arguably the most policy-oriented research question our work can answer is *are counties heating up over time* as it relates to changes in OPR.[14] To answer this question, we compute G_i scores using OPR data across 2013 and 2014. With these results, we perform two back of the envelope analyses:

1. **Changes to Spot Score Distribution**. We restrict our selection of G_i scores to those with p.values robust to $\alpha = 10\%$ and compare the distribution of G_i scores between 2013 and 2014. Note that the composition of (significant G_i score) counties across periods changes. However, since we are primarily interested at seeing how the distribution changes, we press forth (mostly) undeterred. Looking to Figure 13, 2013 G_is has a larger density for $G_i > 0$ relative to 2014's distribution. Indeed, whereas approximately 5% of counties where "hot spots" in 2014 (i.e. $G_i > 0$), 2013's corresponding figure was close to 7%; thus, number of hot spots (at least in relative terms) fell from 2014 compared to 2013.

2. **Changes in Average Spot Score**. On average between 2013 and 2014, did OPR hot spots heat up? Alternatively, how did county spot scores change on average? To answer both of these questions, we turn the information presented in Table 4.[15] We decompose county G_i scores into one of four self-explanatory categories: Did county i heat up or cool down between 2013 and 2014 and did it start out as hot or cold? Table 4's results suggest that yes indeed American's "hot" OPR counties did indeed heat up from 2013 relative to 2014 but that on balance, most counties on average cooled down. Interestingly enough, once we adjust for the number of counties,[16] cold counties that heat up essentially offset hot counties cooling down. Furthermore, since both the number of colder counties cooling down and the magnitude of that average heat change exceeds that of hot counties heating up, this leads to our net cooling.

14.

15. Future versions of this work will need to improve on which counties to compare. For instance, should we only compare counties that had significant G_i scores in 2013, in 2014, or in both? As of now, we included **all** G_i scores irrespective of significance.

16. Future versions of this project would adjust by population and/or population density rather than county count.

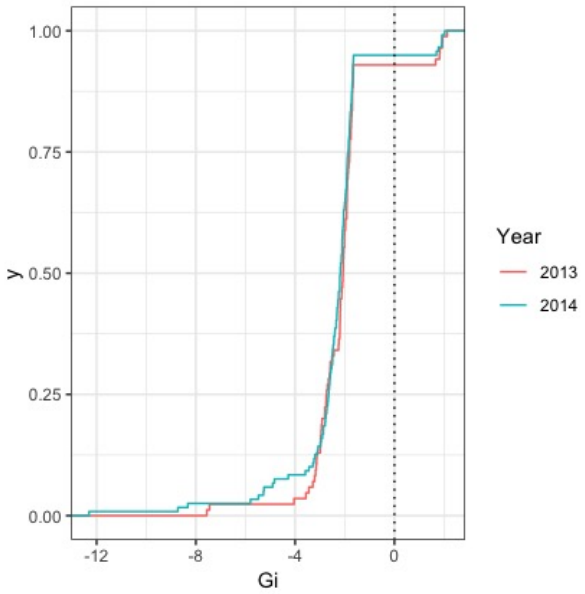

Figure 13: Empirical CDF of G_i (OPR) Scores

Direction	Count	Mean Change	Share	Adjusted Mean
Cold getting Colder	511	-0.514	0.182	-0.0937
Cold getting Hotter	994	0.647	0.355	0.230
Hot getting Colder	927	-0.699	0.331	-0.231
Hot getting Hotter	370	0.292	0.132	0.0386
Net Change				-0.057

Table 4: Change in G_i

Finally, Figure 14 gives a summary level view of G_i OPR scores changed between 2013-2014. Note that while the color scheme has changed, brighter colors still correspond to higher levels, and in this instance, higher levels of change. As with both before, counties shaded gray did not have G_i scores for either 2013 or 2014.

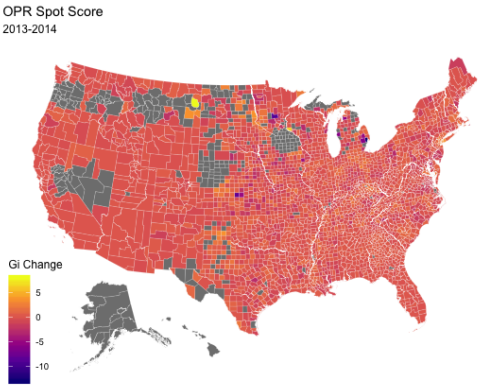

Figure 14: ΔG_i OPR Map

Once again, having a relatively larger dispersion of ΔG_i hinders visual distinctions. Suffice to say, Michigan has a number of darker spots (cooling down) whereas states in the Northeast are more uniformly light (heating up); indeed, Table 6 presents the average ΔG_i between 2013 and 2014 and confirms that Michigan has the lowest average ΔG_i whereas Northeastern states tend to be getting hotter.

4 Appendix

4.1 Code Developed

Table 5: Function Summaries

Function	Arguments	Output
`gcd.hf`	4, $n \times 1$ vectors with longitude and longitude coordinates.	Haversine great circle distance in miles.
`distance.matrix`	A $n \times 3$ dataframe; the first column is a list of geographic identifiers (e.g. place name) and the last two columns list longitude/latitude coordinates per each identifier.	$n \times n$ symmetric distance matrix D; d_{ij} is the haversine distance between locations i and $j, d_{ii} = 0$.
`weight_distance_matrix`	Distance matrix D, distance bandwidth parameter $d_{max} > 0$, $n \times 1$ population vector, population clustering parameter λ, and options to population-weight distances.	Weight matrix $W_{n \times n}$. See additional details.
`moranI`	Spatially lagged variable $y_{n \times 1}$, weight matrix $W_{n \times n}$, weight scaling option (default is false), and p.value options.	Moran's index I (as in (1)) and (one or two-sided) p.value. If weight scaling option is set to `TRUE` (default is `FALSE`), then W is normalized such that $\sum_i W_{ij} = 1$ (which lowers the magnitude of I).
`moran_time_dist`	Spatially lagged and time-varying $y_{(n \times t) \times 1}$, a column vector of time variables ($(n \times t) \times 1$), distance matrix $D_{n \times n}$, a sequence of distance bandwidths $d_{max}(d \times 1)$, and a year sequence $(t \times 1)$.	Moran's index I and p.value (default set to two.sided) for each distance bandwidth and year in the year sequence. Note that weight-scaling is defaulted to `TRUE`. Output is a $(d \times t) \times 4$ dataframe. See additional details.
`MoranI_pop`	Spatially lagged variable $y_{n \times 1}$, population vector $p_{n \times 1}$, distance matrix $D_{n \times n}$, a sequence of distance bandwidths $(d \times 1)$, and sequence of population clustering-parameters $\lambda_{l \times 1}$.	Moran's index I using population adjusted weights for each distance bandwidth parameter and population clustering-parameter. Note that the weight scaling option and the p.value are defaulted to `FALSE` and two.sided, respectively. Output is $(d \times l) \times 4$ dataframe.
`LocalMoran`	Spatially lagged variable $y_{n \times 1}$, distance matrix $D_{n \times n}$, distance bandwidth $d_{max} > 0$, weight scaling option (default is still `FALSE`), and p.value options.	Local Moran's I_i (as in (9)) for each n locations and corresponding p.value (default here is set to two-sided). Output is formatted as a list of two lists: the first list contains I_i and the second list contains associated p.values.

Continued on next page

Table 5– continued from previous page

Function	Arguments	Output
`Getis_Ord`	Spatially lagged variable $y_{n\times 1}$, weight matrix $W_{n\times n}$, and p.values (default is set to one.sided).	G statistic (from (8)), Spot type which is merely $G - \mathbb{E}[G]$ (if this value is positive, return Hot, o.w. Cold), and one.sided p.value. Note that the output is formatted as a list containing the aforementioned items. See additional details.
`Getis_Ord_local_z`	Spatially lagged variable $y_{n\times 1}$, distance matrix $D_{n\times n}$, and distance bandwidth $d_{max} > 0$.	G_i (as in (11)), one sided p.value, and spot status ($G_i < 0$ returns Cold, o.w. Hot). Unlike, `Getis_Ord`, this output is formatted as a dataframe.
`find_global_Geary_C`	Year from test data and distance bandwidth $d_{max} > 0$.	C statistic as in (6). See additional details.
`local_GC`	Spatially lagged variable $y_{n\times 1}$, distance matrix $D_{n\times n}$, and distance bandwidth $d_{max} > 0$.	$n \times 2$ dataframe with local Geary's c_i as in (10) and corresponding two tailed p.value.
`grid_spacing`	Start/end points a/b, number of grid points n, and exponential tuning parameter θ (if $\theta = 1$, then output is a sequence of linearly spaced points).	$x_{n\times 1}$ vector such that: $x_i = a + (b-a)\left(\frac{i-1}{n-1}\right)^{\theta}$

4.2 Additional Function Details

1. `weight_distance_matrix`.
   Weight matrix W is calculated in one of two ways.

   (a) **Weights are not population adjusted**. For a distance bandwidth $d_{max} > 0$ then

$$w_{ij} = \begin{cases} \frac{1}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \text{ or } d_{max} < d_{ij}. \end{cases} \tag{12}$$

   (b) **Weights are population adjusted.**

$$w_{ij} = \begin{cases} \frac{\alpha_{ik}}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \text{ or } d_{max} < d_{ij}. \end{cases} \tag{13}$$

   Parameter α_{ik} and index k require some explanation. Let $\mathcal{S}_i$ denote the set of location i's distances to all locations (including itself) sorted in increasing order. Rather than keeping track of each elements of this sorted distance list, $\mathcal{S}_i$ simply keeps track of the sorted-list's elements indices (from the ith row of D). A typical element of S_i, k, will correspond to the sth closest location to i. For example $S_i(1) = 1$ since i is the index of the closet location to i.
   Let $RP_i(k)$ denote the regional population of i's closest neighbors. Population-clustering parameter $\lambda > 0$ is the upper bound on large $RP_i(k)$. Now let $\mu_i > 0$ represent the population of i. Index k and correspondingly α_{ik} are determined via the following algorithm:

Algorithm 1 Determining α_{ik}

if $\mu_i \geq \lambda$ **then return** $\alpha_{ik} = d_{ik}$
$RP_i = \mu_i$
for ($s \in S_i$ $s = 2, \ldots, n$) {
  $k = \mathcal{S}_i(s)$
  $RP_i = \mu_k + RP_i$
  if $RP_i \geq \lambda$ **then return** $\alpha_{ik} = d_{ik}$
}

2. `moran_time_dist`.
   Note that this function exists primarily to facilitate computing I for different $d_{\max}$ and across different time periods and to make graphing much easier.

3. The functions `find_global_Geary_C` and `plot_OPR` were implemented with specific test data in mind. In principle, these functions could be reworked to accommodate more generalized datasets.

4. `Getis_Ord`
   One issue that we had with computing G was when it came to determining $V(G)$. In particular, each time we tried to calculate $V(G)$, we consistently ran into the case where $\mathbb{E}[G^2] < \mathbb{E}[G]^2$, such that we ended up with negative variance. This is bad. Most likely, some of the place holder variables (e.g. A, B and/or C) were incorrectly coded in the first instance.

4.3 Testing

All relevant files/outputs from this project may be found at the following GitHub repository: https://github.com/davidmterneriu/S610_Final_Project.git

Testing Resources

- Our actual and cleaned data is found in `data_folder.zip`.
- `test_file.R` contains generic implementations of our functions/codes.
  `find_global_Geary_C and a testthat for it.R` includes outstanding test material.

4.4 Assorted Results

State	Mean Change
New York	0.4166
North Dakota	0.4144
Montana	0.347
Rhode Island	0.3428
Delaware	0.2556
New Jersey	0.2414
Massachusetts	0.228
Georgia	0.1724
Kentucky	0.1707
Nevada	0.1345
Tennessee	0.1051
Texas	0.0957
Alabama	0.0914
South Carolina	0.0695
New Hampshire	0.0401
Ohio	0.0301
Florida	0.0164
Arkansas	0.0152
Illinois	0.0141
Vermont	0.0138
Louisiana	0.0101
Wyoming	-0.0156
Utah	-0.0204
Arizona	-0.0264
Washington	-0.0328
Oklahoma	-0.0357
Oregon	-0.0395
Colorado	-0.0482
New Mexico	-0.0706
California	-0.0709
West Virginia	-0.0712
Mississippi	-0.0812
Indiana	-0.1195
Idaho	-0.1314
Missouri	-0.1573
South Dakota	-0.1629
North Carolina	-0.1721
Minnesota	-0.1782
Wisconsin	-0.1786
Maryland	-0.2411
Kansas	-0.2956
Connecticut	-0.3199
Pennsylvania	-0.3351
Nebraska	-0.3389
Virginia	-0.361
Maine	-0.4438
Iowa	-0.4675
Michigan	-0.7448

Table 6: Mean ΔG_i (OPR) change 2013-2014

References

Anselin, Luc. 1995. "Local indicators of spatial association—LISA." *Geographical analysis* 27 (2): 93–115.

Getis, Arthur, and J Keith Ord. 2010. "The analysis of spatial association by use of distance statistics." In *Perspectives on Spatial Data Analysis,* 127–145. Springer.

Hollingsworth, Alex, Christopher J Ruhm, and Kosali Simon. 2017. "Macroeconomic conditions and opioid abuse." *Journal of health economics* 56:222–233.

Moran, Patrick AP. 1950. "Notes on continuous stochastic phenomena." *Biometrika* 37 (1/2): 17–23.

Oden, Neal. 1995. "Adjusting Moran's I for population density." *Statistics in Medicine* 14 (1): 17–26.