Spatial Cluster Detection

Matthew Quinn

May 16th, 2019

Abstract

Many research questions in public health, surveillance, and related fields concern data involving a spatial component. For such investigations, one common concern is being able to address dependence across observations that arises due to their physical proximity to one another. When studying an infectious disease outcome, for instance, it might be of concern that individuals that live near one another would not exhibit independent outcomes due to the fact that they may come in contact with one another. In such cases, detecting spatial clusters exhibiting such dependence is a useful step to take. In this report, we discuss two such approaches for spatial cluster detection: the Getis-Ord G_i^* and Local Moran's I_i. We discuss their properties and implement them in a data example involving drug-related mortality rates in counties throughout the United States.

Contents

1 Introduction

In areas like public health, geography, and surveillance, researchers frequently encounter data-driven investigations that involve a spatial component. For instance, public health officials may want monitor and study the geographical prevalence of diseases, ranging from cancers to HIV, so that resources can be deployed to areas where screenings or treatment would be of the largest benefit [9, 15]. More recently, the resurgence of diseases like measles necessitates surveillance in order for health organizations to help minimize the diseases' impacts [10]. In such cases, it is often going to be of interest to identify regions where individuals or observations exhibit dependence across one another based on physical proximity. For instance, if several cases of measles appear in a given town, we may expect neighboring towns to soon exhibit an increase in measles cases as well. The physical proximity between towns lends itself to frequent contact between individuals across municipal lines, thereby potentially spreading measles. Such a scenario, where we observe dependence in the prevalence of measles in a group of neighboring cities, may be referred to as a "spatial cluster", also sometimes called a "hotspot". In trying to identify where such clusters exist in a given study region, we perform what is called "spatial cluster detection".

More broadly, in performing spatial cluster detection, we are considering that there are possibly underlying phenomena through which physical proximity can induce dependence across our observations. For instance, as mentioned, those who live or work near each other may be more prone to spreading an infectious disease due to frequent contact. Alternatively, those living in the same neighborhood may share many underlying socioeconomic factors with one another that may influence the risk of a particular cancer. Residents of the same county or state will also often be subject to the same culture and rule of law, which in turn may influence behaviors and risk factors for diseases. All of these are underlying factors that may cause individuals to exhibit outcomes dependent on those physically located around them.

With this in mind, it is worth noting that spatial cluster detection may serve multiple purposes and that these purposes may differ depending on the goal of a particular investigation. It is possible for cluster detection to be the goal in itself, allowing health organizations to respond to regions and their resident populations most in need of assistance during outbreaks. In such an investigation, we may not care as much about controlling for other mediating factors. For instance, if an organization wants to identify spatial cluster across the United States where measles is growing in incidence, then it may not matter that access to health care could be a risk factor, which would differ by region. It just matters that the organization can deploy resources to those clusters exhibiting high or growing measles incidence to begin with. However, spatial cluster detection could also potentially be used as a step in exploratory data analysis towards constructing more complex models for an outcome of interest, in which case understanding the underlying spatial dependence structure is of great importance. In such an investigation, it may be of more interest to first control for mediating factors that can be measured, such as health care access when looking at an outcome of measles incidence, since this may account for some spatial dependence. Similarly, if an investigation aims to uncover new risk factors for a given disease by utilizing spatial clusters, then we would likewise want to account for already-known risk factors [17]. In the context of model-building, it is worth noting that cluster detection itself would not necessarily be used to inform the specific covariance structure for a spatial model so much as to identify regions where dependence may be present. For discussion of spatial modeling itself, please refer to [16].

In Section 2, we discuss some of the different categories of approaches for spatial cluster detection and establish notation. In Section 3, we present two popular and foundational approaches for the detection of spatial clusters in detail: the Getis-Ord G_i^* and the Local Moran's I_i. In Section 4, we go through an example of using cluster detection on county-level outcomes of drug-related mortality

in the wake of the growing opioid crisis. In Section 5, we conclude the report.

2 Setting

There exist a large variety of methods available for detecting spatial clusters. While we will only discuss a couple in detail in this report, it is important to be aware of the variety of approaches available. In Chapter 8 of "Spatial Epidemiology: Methods and Applications", Wakefield, Kelsall, and Morris provide an overview of about a dozen common approaches [4]. As the chapter explains, these can be described as falling into four main categories. The first, specifically for count data, includes some relatively naive approaches which entail partitioning a region and using methods like Pearson's χ^2 statistic or Potthoff and Whittinghill's method. A second category of approaches consists of autocorrelation measures, which we will discuss in more detail shortly. The third consists of moving window and scan statistics, which can be thought of as overlaying a circle on the study region, testing for an excess number of counts within the circle, and then moving the circle over to repeat the test in a new subregion. The final category includes methods for actually modeling the residual risk or odds for an outcome, namely residual after accounting for various covariates.

2.1 Autocorrelation Methods

The two methods for detecting spatial clusters discussed in this report will both be autocorrelation methods. Autocorrelation measures provide an indication of the amount of dependence between observations within a given region, using a single summary statistic. Two main categories of autocorrelation statistics include global and local statistics. Global autocorrelation statistics largely came into fruition in the 1950s with measures like Moran's I and Geary's C [11, 6]. These measures produce one statistic for the entire study region at hand. In other words, if one was considering cancer incidence by county throughout the United States, one statistic for the entire U.S. would be produced to indicate spatial dependence. For the purpose of cluster detection, this is not particularly useful for a couple reasons. One is that even if there is no significant spatial dependence across the entire country, this is not to say spatial dependence does not exist in subregions of the U.S., but a global measure would fail to account for this. Secondly, even if significant dependence is found, there is no formal indication of whether or not particular subregions are driving this significant spatial dependence to arise, as opposed to the dependence being somewhat uniform across the country.

These issues with global autocorrelation measures largely lead to the creation of local autocorrelation methods in the 1990s, including measures such as the Getis-Ord G_i^* and Local Moran's I_i [7, 1]. These local measures consider subregions of the overall study region and, for each subregion, assess the evidence for spatial dependence. Doing so for subregions accounts for the aforementioned issues with global measures by looking at smaller areas individually. In particular, we typically will calculate these local autocorrelation measures with respect to a subregion corresponding to each observation in our sample, hence the indexing by i in G_i^* and I_i. We formalize this a bit more by first establishing notation.

2.2 Notation

For the following methods, we will consider a sample size n where observations are indexed by i and/or j. We consider a positive random variable, Y, that represents a given outcome of interest (e.g. counts associated with the number of incident cases at a location, the incidence rate of a given disease in a county, etc.). This outcome has observed values $y_i, i = 1, \ldots, n$. Each observation is associated with a location, which we will denote as $\ell_i, i = 1, \ldots, n$. For many research problem,

location will be in a geographical coordinate system like longitude and latitude, but essentially any system that allows for a distance between observations to be calculated (e.g. Cartesian coordinates, polar coordinates, etc.) works for the methods we discuss below. For the particular autocorrelation methods we consider here, we imagine the total study region divided into n overlapping regions. The i^{th} region is centered at observation i's location, ℓ_i, and consists of a disc of radius d surrounding ℓ_i. We let W be a spatial weights matrix where w_{ij} represents the spatial weight between observations i and j. Typically, this weight will be reflective of the distance between points. For instance, we may end up assigning greater weights to pairs of points that are located more closely together than pairs that reflect points relatively far apart. As such, we can think of w_{ij} as a function of d, $w_{ij}(d)$. While we will frequently denote w_{ij} and resulting quantities as functions of d in subsequent paragraphs, it is common to drop this notation when d is treated as fixed. We also use $W_i^*(d) = \sum_{j=1}^n w_{ij}(d)$ to denote the sum of all spatial weights for pairs involving observation i.

Other notation follows typical statistical convention. For instance, we would use $E[Y]$ if we were interested in denoting the expected value of the outcome, Y. Likewise, $\bar{y}$ will denote the sample mean and s_y^2 the sample variance based on the observed $y_i, i = 1, \ldots, n$.

3 Methods

We now review the Getis-Ord G_i^* and Local Moran's I_i in more detail [7, 1].

3.1 Getis-Ord G_i^*

Before defining the actual G_i^* statistic, it is important to first choose a spatial weights matrix, W, as defined in Section 2.2. Following the choice of the original paper, we will set $w_{ij} = 1$ if the distance between observations i and j (i.e. the distance between locations ℓ_i and ℓ_j) is no more than d and $w_{ij} = 0$ otherwise. While we will use this convention here for some simplicity, we also mention an extension to nonbinary weights in Section 3.3.2.

We now define the G_i^* statistic to be:

$$G_i^*(d) = \frac{\sum_{j=1}^n w_{ij}(d) y_j}{\sum_{j=1}^n y_j} \tag{1}$$

where we note that the G_i^* can also be viewed as a function of a chosen distance, d. Note that only those observed values within distance d of observation i contribute to the numerator, $\sum_{j=1}^n w_{ij}(d) y_j$, when we use the binary spatial weights as described. Intuitively then, the G_i^* is essentially a measure of the concentration of values captured within distance d of observation i. We recall that this statistic is calculated for each $i = 1, \ldots, n$. While one can choose differing values of d, it will often be kept constant. In the case where d is kept constant, $w_{ij}(d)$ may just be denoted as w_{ij} and $G_i^*(d)$ is typically just denoted as G_i^*, as we often refer to it here.

We want to take this statistic and make a formal assessment of the evidence for spatial dependence in a particular region. To do so, Getis and Ord use hypothesis testing. They propose a null hypothesis, H_0, under which there is spatial independence. This is to imply that any physical grouping of values similar to one another would be due entirely to chance, instead of any underlying phenomena as those discussed in Section 1. To this extent, we can imagine that the observed values, $\{y_i\}_{i=1}^n$, and their corresponding locations, $\{\ell_i\}_{i=1}^n$, are just one permutation of $n!$ possible permutations of the observations across the locations. Under H_0, we can therefore consider each permutation of the observed values over the set of locations to be equally likely. Under this null

assumption, one can then derive that:

$$E[G_i^*(d)] = \frac{1}{n} W_i^*(d) = \frac{1}{n} \sum_{j=1}^{n} w_{ij}(d) \tag{2}$$

$$Var(G_i^*(d)) = \frac{W_i^*(d)(n - W_i^*(d))s_y^2}{n^2(n-1)(\bar{y})^2} \tag{3}$$

where, as a reminder, $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ and $s_y^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j)^2 - (\bar{y})^2$. For the full derivation of these moments, please refer to [7].

A particularly nice property of G_i^* is that under certain conditions, it asymptotically approaches normality for increasingly large n. The needed conditions for this to occur are that:

1. The chosen distance d cannot be too small such that $W_i^*(d) \to 0$. Intuitively, we do not want to select a subregion of such a small size that effectively no or very few other observations are captured.

2. The chosen distance d cannot be too large such that $W_i^*(d) \to 1$. Intuitively, we do not want to select a subregion of such a large size that effectively almost all other observations are captured.

If either of these cases were to occur, then G_i^* would approach 0 or 1, which are its bounding values when using the binary spatial weights as described previously.

Taking this into account, we can then standardize G_i^* to obtain:

$$Z_i^*(d) = \frac{G_i^*(d) - E[G_i^*(d)]}{\sqrt{\text{Var}(G_i^*(d))}} \tag{4}$$

which will then approach a standard normal asymptotically under the appropriate conditions. Again, we may denote this simply as Z_i^* if the chosen d is fixed. One can then use the p-value from a z-test for evidence against the null of spatial independence or choose a significance level to assess the presence of statistically significant spatial dependence. Lastly, we note that a positive Z_i^* indicates clustering of high values in an area while a negative Z_i^* indicates clustering of low values. This stands in contrast to the interpretation of Local Moran's I_i in Section 3.2.

3.2 Local Moran's I_i

Local Moran's I_i follows a very similar setup to that of the G_i^*. However, we will standardize the observed outcomes of interest instead of considering raw values. Thus, Let $v_i = \frac{y_i - \bar{y}}{\sqrt{s_y^2}}$ for $i = 1, \ldots, n$. We can again take spatial weights to be binary, as described in Section 3.1 for consistency in comparing the two approaches. However, other options for weighting schemes also exist, as described in Section 3.3.2. We will take, by convention, that $w_{ii}(d) = 0$ for all i and for any d, regardless of the chosen weighting scheme. The Local Moran's I_i is then:

$$I_i(d) = v_i \sum_{j=1}^{n} w_{ij}(d) v_j \tag{5}$$

Following the same null hypothesis, H_0, of spatial independence as presented in Section 3.1, we can find that, under the set of permutations of observed values across locations:

$$E[I_i(d)] = -\frac{\sum_{j=1}^{n} w_{ij}(d)}{n-1} \tag{6}$$

$$\begin{aligned} Var(I_i(d)) = &\frac{1}{n-1}\left[(n-b)\sum_{j=1}^{n} w_{ij}(d)^2\right] \\ &+ \frac{1}{(n-1)(n-2)}\left[(2b-n)\sum_{k=1}^{n}\sum_{j=1}^{n} w_{ik}(d)w_{ij}(d)\right] - \frac{1}{(n-1)^2}\sum_{j=1}^{n} w_{ij}(d)^2 \end{aligned} \tag{7}$$

where $b = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i^4}{(\frac{1}{n}\sum_{i=1}^{n} v_i^2)^2}$. For the full derivations, please refer to [1]. As with prior convention, it is also common to drop notation indicating that quantities are functions of d when d is considered fixed.

It is then possible to again use a z-statistic:

$$Z_i(d) = \frac{I_i(d) - E[I_i(d)]}{\sqrt{\text{Var}(I_i(d))}} \tag{8}$$

in order to obtain a p-value for evidence against the null of spatial independence or choose a significance level and assess the presence of statistically significant spatial dependence. We note that a positive I_i suggests clustering of similar values while a negative I_i suggests clustering of dissimilar values, such as a high observed value surrounded by relatively low values. This stands in contrast to the interpretation of the standardized G_i^* in Section 3.1.

However, Luc Anselin, who developed this method, notes that the exact distribution of the Local Moran's I_i is not known and that approximate normality may not be an appropriate assumption, such that using a z-test may not be appropriate either. With these concerns, Anselin suggests using a non-parametric permutation test, which is discussed in more detail in Section 3.3.3.

3.3 Extensions of the Methods

It is worth noting several extensions to the approaches involving the Getis-Ord G_i^* and the Local Moran's I_i.

3.3.1 Classes of Autocorrelation Statistics

Both of the methods presented here are part of larger classes of autocorrelation statistics. In the case of the G_i^*, Getis and Ord present an analogous G_i statistic that operates in a nearly identical manner. The only difference is that for the G_i, the i^{th} observation does not contribute to its own statistic. In this case, we could imagine setting $w_{ii} = 0 \forall i$ in the G_i^* statistic and adjusting the subsequent mean and variance of G_i^* to decrease n to $n-1$ due to the exclusion of one datum point. Note that this is somewhat similar to how the Local Moran's I_i has $w_{ii} = 0$. However, the i^{th} standardized observation, v_i, still contributes to its own I_i statistic despite this. Therefore, while both the G_i and I_i statistics set $w_{ii} = 0 \forall i$, it is possibly more appropriate to compare the G_i^* to the I_i since both allow for contributions of the i^{th} observation to its own statistic. This also seems to be convention, as Anselin compares his Local Moran's I_i to the G_i^* specifically in [1]. Additionally, Getis and Ord present a global autocorrelation G statistic, which they discuss in [7].

The Local Moran's I_i is part of a class of approaches called local indicators of spatial association (LISAs) [1]. A LISA is defined to be a statistic such that:

1. For each observation, it provides an indication of the extent of spatial clustering of similar values around that observation.

2. The sum of LISAs for all observations is proportional to a global indicator of spatial association.

Essentially, LISAs allow for the decomposition of a global autocorrelation measure into local autocorrelation measures, something that did not necessarily hold for the G_i^*. Anselin argues that this is useful primarily for the detection of outliers when global association is actually present. That is, the null hypothesis for the tests discussed in Section 3 was spatial independence, which necessitates that there is no global autocorrelation present. In reality, it is possible for global autocorrelation to be present and for local areas to exhibit patterns of spatial dependence different from the global pattern. In other words, even with global spatial dependence present, there can be heterogeneity in local spatial dependence across the observations. However, it is difficult to derive formal tests for extreme pockets of local autocorrelation under such global autocorrelation. Thus, Anselin argues that LISAs, even in the presence of global autocorrelation, allow one to compare local autocorrelation statistics to the global autocorrelation measure in order to assess relatively extreme local autocorrelation patterns. This does not replace the desire for a more formal testing scheme in the presence of global autocorrelation, but may still be a useful tool for investigation [1].

3.3.2 Choice of Spatial Weights

In this report, we have chosen binary spatial weights as described in Section 3.1 for the sake of consistency across tests, for the purposes of a straightforward application in Section 4, and because it is the scheme first introduced in [7]. However, it is not necessary to use binary spatial weights as presented here, though it may be a common choice. At that, w_{ij} need not even truly be a function of a distance d. For instance, in an example he presents on international conflict, Anselin uses binary weights where $w_{ij} = 1$ if countries i and j share a common border and $w_{ij} = 0$ if not. Anselin also suggests row-standardization of the spatial weights matrix, W, for some possible improvement in interpretation, though it is not required [1]. However, one could also potentially consider nonbinary weights. For example, one could weight by a quantity proportional to the inverse distance between points, as is one option explored in an example by Getis and Ord in their follow-up 1995 paper, in which they discuss nonbinary weights [14].

3.3.3 Permutation Tests

As referenced in Section 3.2, instead of using z-tests for either statistic at the risk of not meeting the assumption of asymptotic normality, it may be more appropriate to use a non-parametric permutation test. Anselin suggests a conditional permutation test, in which the value y_i at location ℓ_i is kept fixed but the remaining $n-1$ observed values are permuted over the remaining $n-1$ locations. That is to say, we would ideally perform every such possible permutation, calculate the statistic (either I_i or G_i^*) each time, and then compare the observed statistic to this empirical distribution obtained through permutations. Since it is computationally unreasonable to perform $(n-1)!$ permutations in many cases, it will often be more appropriate to randomly sample a subset of such permutations instead of performing every one.

3.3.4 Multiple Testing

As mentioned previously, typically we will want to calculate G_i^* or I_i for each observation in our data set, consequently then performing n hypothesis tests in total. If one is concerned about restricting the type I error rate, then this presents a multiple testing problem. Thus, if one desires the probability

of at least one type I error across all tests to be no more than α, then Anselin suggests using a Bonferroni correction, $\alpha^* = \frac{\alpha}{n}$, or a Šidák correction, $\alpha^* = 1 - (1 - \alpha)^{1/n}$, for determining a significant result within each test individually [1]. Since the tests in this context are positively correlated, both of these corrections offer relatively conservative thresholds.

4 Data Application

4.1 Background and Data

In recent years, drug abuse and the opioid crisis have garnered national attention for the rise in drug-related deaths throughout the United States. With these tragic deaths and related drug incidents on the rise, the U.S. Department of Health and Human Services has been deploying resources to hamper these trends [13]. It seems that cluster detection could offer some benefit in identifying regions exhibiting clustering of relatively high rates of drug-related deaths. Identifying these regions would allow health organizations to more effectively target resources to those most affected. Additionally, by looking at these data over time, one could also identify some trends in any geographic spread of drug-related deaths. To investigate this, we consider several data sets on drug-related deaths from the CDC.

The data for deaths associated with drug use were gathered from the CDC's Multiple Cause of Death repository by selecting results to be given by county, selecting a particular individual year (2013 through 2017), and selecting drug-induced causes for the multiple cause of death [5]. We choose to also obtain suppressed and zero values in the request. All other options were kept at default. The requested data were exported to a tab delimited text file. This was then repeated for remaining the years in 2013-2017. Some pruning of the files in a text editor were necessary to remove metadata exported to the bottom of each file from the repository. Since we are just interested in identifying those areas with the most need for resources, we simply consider the crude drug-related mortality rate per 100,000 individuals rather than adjusting for any other risk factors. We display the map with this information by county for 2017 in Figure 1.

Since we need to be able to calculate distances between counties, we select points within each county as representative of that county's location. Specifically, we take the U.S. Census Bureau's centers of population for each county from 2010 as this representative point [3]. Additionally, for our implementation, we will need to choose a distance d which will serve as the radius around each county's center in assessing the presence of spatial clustering. In practice, this choice would likely be based off of subject-matter knowledge and may vary depending on the region of interest. For our purposes here, we will somewhat arbitrarily choose a radius of $d = 75$ miles for all observations. After some trial and error, this appears to be on the order of the smallest radius we can choose such that counties towards the West still have other counties within distance d of themselves. We display our data for 2017 again in Figure 2, this time with the centers of population for each county indicated by a yellow dot and an example of a 75 mile radius around Bronx County, NY.

Visually, we note some potential clustering of high crude drug-related mortality rates in areas such as Maine, Northern Michigan, and Appalachia (from southern Ohio towards Tennessee). Additionally, some areas of potentially low crude drug-related mortality rates are in Pennsylvania into Connecticut. We will be interested in seeing whether the Getis-Ord G_i^* or the Local Moran's I_i detect any of these regions as spatial clusters.

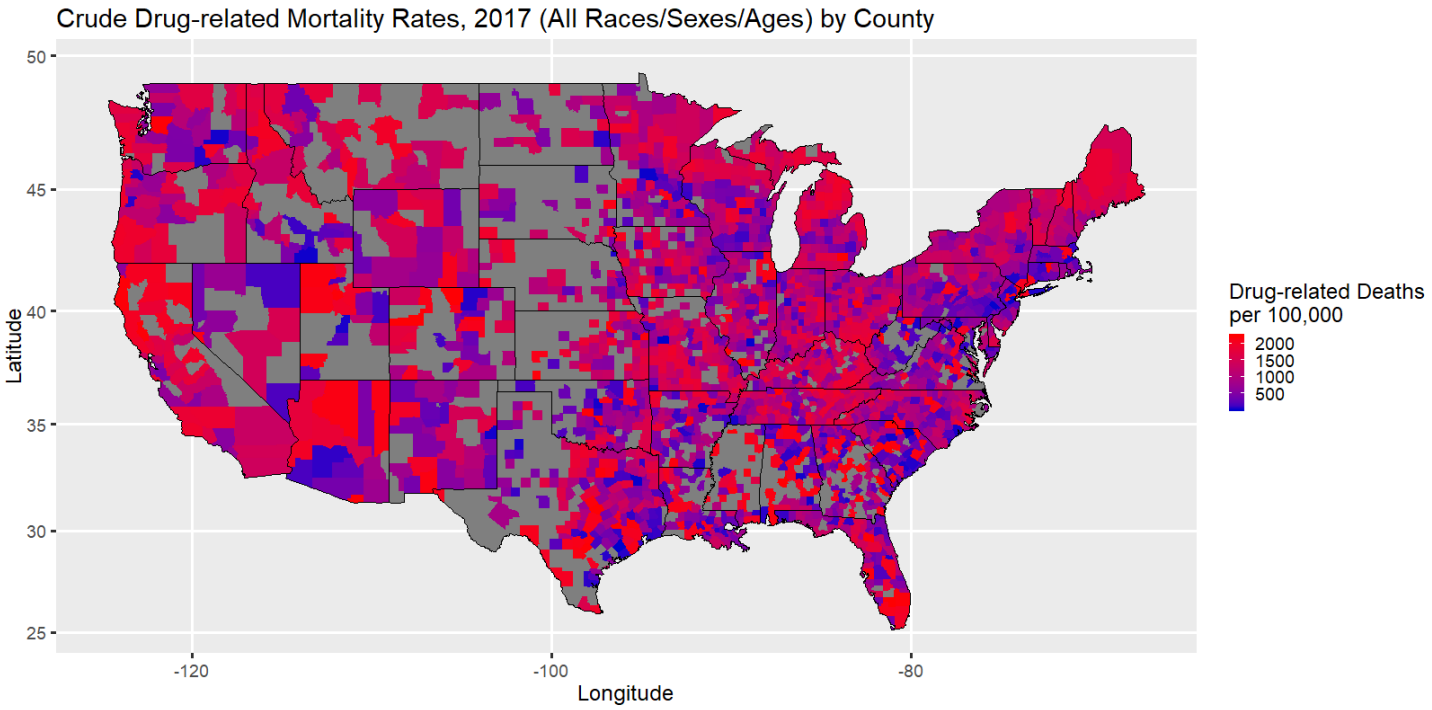

Figure 1: Crude drug-related mortality rates per 100,000 individuals by county in 2017

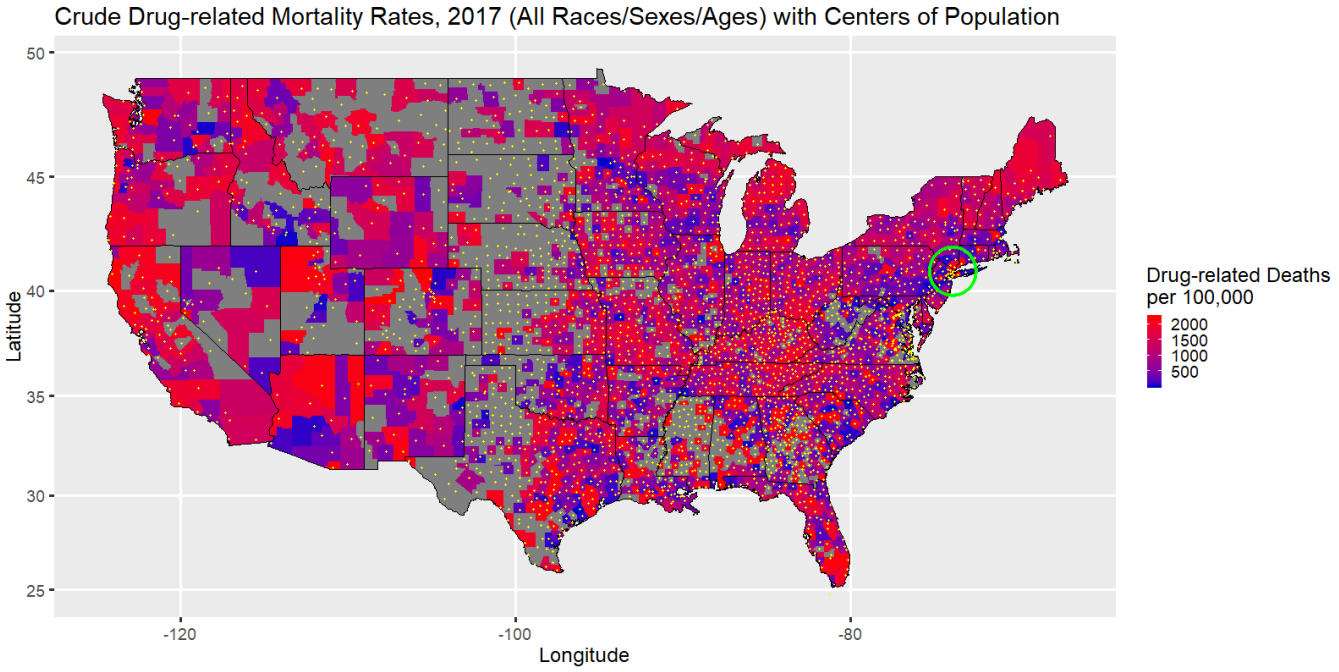

Figure 2: Crude drug-related mortality rates per 100,000 individuals by county in 2017 with 2010 centers of population and a 75 mile radius around Bronx County, NY

4.2 Implementation

A fair amount of data wrangling was performed in `R` to make data more consistent with one another across the multiple sources we are using. In addition to the data on multiple deaths and centers of population, we also require `map_data("county")` data from the `ggplot2` package in `R` for generating maps. This involved matching different naming conventions (e.g. Louisiana's counties are sometimes called "parishes") and accounting for duplicate names (e.g. some states have cities that are "county-equivalents" but share a name with an already-existing county), among other issues. After the data wrangling, I implement the approaches discussed here mostly from scratch, including calculating the statistics and performing z-tests. I also use some techniques like parallelization for calculating the distances between the thousands of counties. Some online material was referenced and/or modified in order to generate maps and perform computation like calculating distance based on outright longitude/latitude coordinates. These references are provided in the code. The code for this project is hosted on `https://github.com/matthewquinn1/BST245` and a bit more detail about it is included in Appendix B. It is worth noting that the `spdep` package also provides implementations for running the methods discussed here [2].

4.3 Results

We will discuss the results for 2017 in detail and then, for those interested, provide some plots for 2013-2016 in the Appendix A in order to demonstrate the changes in clusters over time. The 2017 drug-related deaths data set contains at least some information (e.g. name, population) on a total of 3,110 counties. However, after accounting for those that have a missing crude drug-related mortality rate or missing coordinates, only 2,280 have the data necessary in order to apply the Getis Ord G_i^* and/or Local Moran's I_i. Most of these are due to missing crude mortality rates, often as a result of a very small population in a county, leading the Census Bureau to suppress the observed rates or to mark them as unreliable. Only 44 of these 830 incomplete counties are incomplete due to lacking geographical information, which largely arises because of mismatches between the census data and CDC data (e.g. differences in spellings, differences in naming conventions, handling of "county-equivalents"). I made an effort to account for some such cases but, due to time-constraints, I am not able to check every remaining one as they need to be handled on a case-by-case basis and often require some background research.

With these 2,280 counties with complete data and a nominal overall type I error rate of $\alpha = 0.05$, our Bonferroni-adjusted significance level is $\alpha^* = \frac{0.05}{2280} \approx 0.0000219$ and the Šidák adjusted level is $\alpha^* = 1 - (1 - 0.05)^{1/2280} \approx 0.0000225$. Both effectively yield the same results but we use the Šidák correction in this case because both will be conservative, but the Šidák correction less so.

Performing the z-tests for each statistic, we find a total of 76 counties across the U.S. whose p-value falls below α^* for at least one of the tests. More specifically, 11 are significant under the Getis-Ord G_i^* and 72 are significant under the Local Moran's I_i, indicating 7 counties are significant under both. It is worth noting that even though Local Moran's I_i appears to be considerably more liberal with identifying significant clustering, many of those counties significant under Local Moran's I_i, but not the G_i^*, still have very small p-values under the G_i^*. In fact, the median p-value under the G_i^* approach for those found significant under Local Moran's I_i, but not the G_i^*, is about 0.00056. By comparison, the median p-value corresponding to the G_i^* for the whole data set is around 0.3. For the sake of clean formatting, we display the results only for the first 10 of these 76 counties alphabetically in Table 1.

Referring back to our discussion of interpretations in Section 3, we can make some general comments. Places like Bell County and Butte County, with positive z-statistics in both cases,

County	G_i^*	z-stat	p-value	I_i	z-stat	p-value
Bath County, KY	0.03	4.30	0.00002	20.34	2.97	0.00297
Bedford County, PA	0.01	-4.44	0.00001	23.86	4.62	< 0.00001
Bell County, KY	0.02	3.80	0.00014	29.22	4.67	< 0.00001
Berks County, PA	0.01	-4.80	< 0.00001	38.78	6.80	< 0.00001
Butte County, CA	0.01	3.45	0.00057	12.89	4.88	< 0.00001
Calhoun County, AL	0.01	-2.98	0.00293	-26.57	-5.14	< 0.00001
Campbell County, TN	0.02	3.75	0.00018	31.86	5.15	< 0.00001
Canadian County, OK	0.01	-2.19	0.02854	-17.97	-4.51	0.00001
Carbon County, PA	0.01	-4.38	0.00001	-6.10	-1.07	0.28560
Carter County, KY	0.03	3.85	0.00012	34.17	5.33	< 0.00001

Table 1: Counties and their respective results from each G_i^* and I_i.

Number of Statistically Significant Counties	States	Number of Statistically Significant Counties	States
18	Pennsylvania	4	Louisiana, Utah
13	Kentucky	3	California, South Carolina, Tennessee
7	Minnesota	2	Connecticut, Maryland, Mississippi, Texas, West Virginia, Wisconsin
5	Michigan	1	Alabama, Arkansas, Delaware, Oklahoma

Table 2: States organized by their number of counties found to exhibit statistically significant clustering under at least one approach.

exhibit clustering of similar and high drug-related mortality rates with nearby counties. Places like Bedford County and Berks County, with a negative G_i^* z-statistic instead, exhibit clustering of similar and low drug-related mortality rates nearby. Places like Canadian County and Carbon County, with negative z-statistics in both cases exhibit clustering of dissimilar and low drug-related mortality rates. This suggests these counties have relatively high drug-related mortality rates but are surrounded by counties with relatively low rates. Finally, while not seen in these ten counties, a county with a positive G_i^* z-statistic but negative I_i z-statistic would be one with a low drug-related mortality rate surrounded by those with high rates.

Tallying results across states, we find the following results presented in Table 2. We note that Pennsylvania and Minnesota have many counties exhibiting clustering with exceptionally low crude drug-related mortality rates, while states like Kentucky and Michigan exhibit large amounts of clustering with relatively high crude drug-related mortality rates. Thus, if a public health organization wanted to allocate resources for addressing drug abuse to where its most needed, it appears that many regions in places like Kentucky and Michigan would be wise choices. Likewise, officials could further investigate regions of Pennsylvania and Minnesota to explore as to what factors may contribute to them exhibiting such low drug-related mortality rates.

In Appendix A, we provide plots and discussion regarding the counties found to exhibit statistically significant spatial clustering under either approach from 2013-2017, using the same approach as described here.

5 Conclusion

In many investigations involving spatial data, it will often be of interest to check for the presence of spatial clustering. A number of different phenomena can lead to dependence across observations based on their physical proximity to one another. Checking for such dependence can have implications when it comes in a variety of different scenarios - from public health organizations attempting to uncover where resources are best used, to a researcher who is planning to go on to construct a more formal spatial model. In this report, we have discussed two approaches for spatial cluster detection, the Getis-Ord G_i^* and the Local Moran's I_i. While both measures are comparable to one another, they differ in some aspects, such as their interpretations, and can yield different results in practice. We saw these approaches implemented with a data example exploring drug-related mortality rates in U.S. counties. We were able to assess the presence of spatial clustering in various counties, and in the appendix, also discuss this over time.

While the two approaches discussed here are common ones, there exist many other approaches that can be implemented as well [4, 17]. While the choice of a specific approach will often be very context-dependent, spatial cluster detection more broadly serves a role for many different types of investigations. At least some thought should be given to spatial cluster detection for virtually any research question that involves spatial data.

References

[1] Luc Anselin. Local indicators of spatial association - LISA. *Geographical Analysis*, 27(2):93–115, 1995.

[2] Roger Bivand and David W. S. Wong. Comparing implementations of global and local indicators of spatial association. *TEST*, 27(3):716–748, 2018.

[3] United States Census Bureau. Centers of population for the 2010 census. https://www.census.gov/geographies/reference-files/2010/geo/2010-centers-population.html, 2010.

[4] Paul Elliott, Jon Wakefield, Nicola Best, and David Briggs. *Spatial Epidemiology: Methods and Applications*. Oxford University Press, 2001.

[5] Centers for Disease Control and Prevention. Multiple cause of death, 1999-2017 request. https://wonder.cdc.gov/mcd-icd10.html, 2018.

[6] R. C. Geary. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3):115–146, 1954.

[7] Arthur Getis and J. K. Ord. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3):189–206, 1992.

[8] Sharon Greene, Eric Peterson, Deborah Kapell, Annie Fine, and Martin Kulldorff. Daily reportable disease spatiotemporal cluster detection, New York City, New York, USA, 2014-2015. *Dispatch*, 22, 2016.

[9] BA Hixson, SB Omer, C del Rio, and PM Frew. Spatial clustering of HIV prevalence in Atlanta, Georgia and population characteristics associated with case concentrations. *Journal of urban health: bulletin of the New York Academy of Medicine*, 88(1):129–1441, 2011.

[10] Jake Hutchison. Measles 'cluster' found in three northern california counties. *The Mercury News*, 2019.

[11] P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[12] Daniel Neill. *Detection of Spatial and Spatio-Temporal Clusters*. PhD thesis, Carnegie Mellon University, 2006.

[13] National Institute of Drug Abuse. Opioid overdose crisis. https://www.drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis, 2019.

[14] J. K. Ord and Arthur Getis. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27(4):286–306, 1995.

[15] RL Sherman, KA Henry, SL Tannenbaum, DJ Feaster, E Kobetz, and DJ Lee. Applying spatial analysis tools in public health: An example using SaTScan to detect geographic targets for colorectal cancer screening interventions. *Preventing Chronic Disease*, 11, 2014.

[16] Daphne Tsoucas and Zack McCaw. Spatial modeling and prediction. *BST 245: Analysis of Multivariate and Longitudinal Data*, 2016.

[17] Jonathan Wakefield and Albert Kim. A Bayesian model for cluster detection. *Biostatistics*, 14(4):752–765, September 2013.

A Additional Plots, 2013-2017

In this appendix, we provide some plots indicating counties that demonstrate statistically significant clustering from the years 2013-2017. Each plot reflects the counties that were found to be significant under at least one of the z-tests associated with the Getis-Ord G_i^* or Local Moran's I_i. Please note that some small discrepancies are possible between the 2017 map and the table presented in Section 4.3 because of mismatches between `ggplot2`'s county data and those data from the CDC. We provide some comments after the maps.

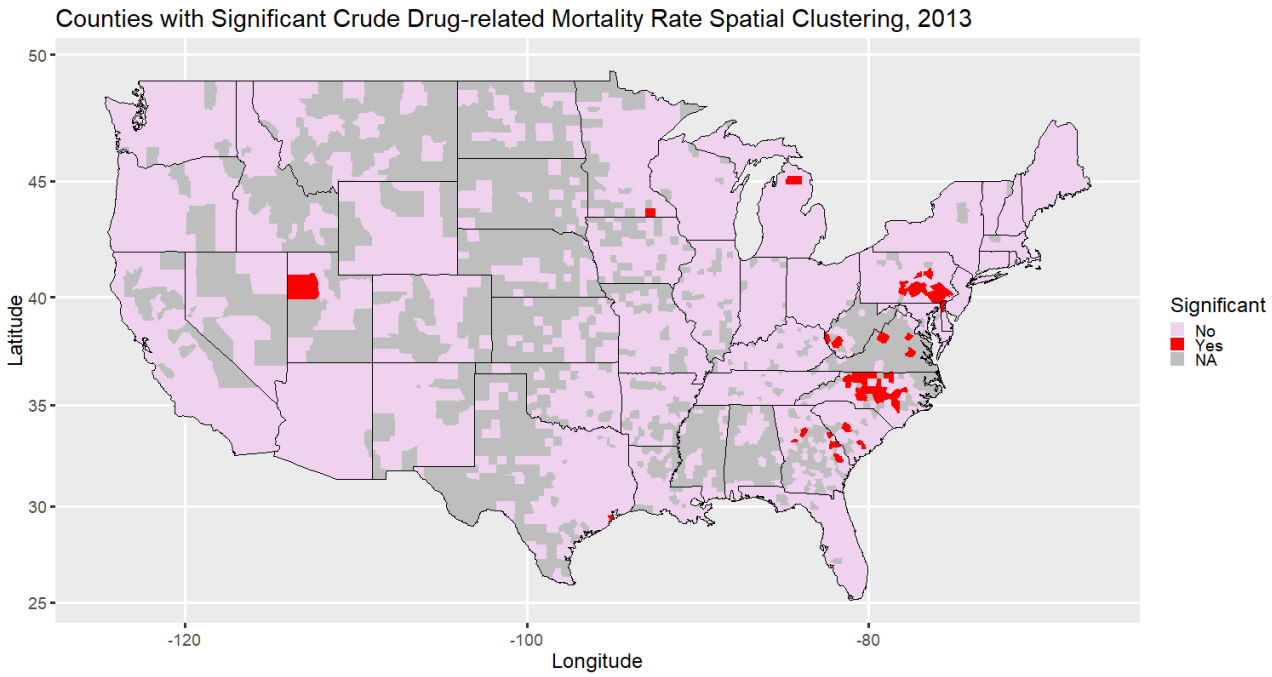

Figure 3: Counties found to have statistically significant spatial clustering within a 75 mile radius in 2013, according to at least one of the Getis-Ord G_i^* or Local Moran's I_i.

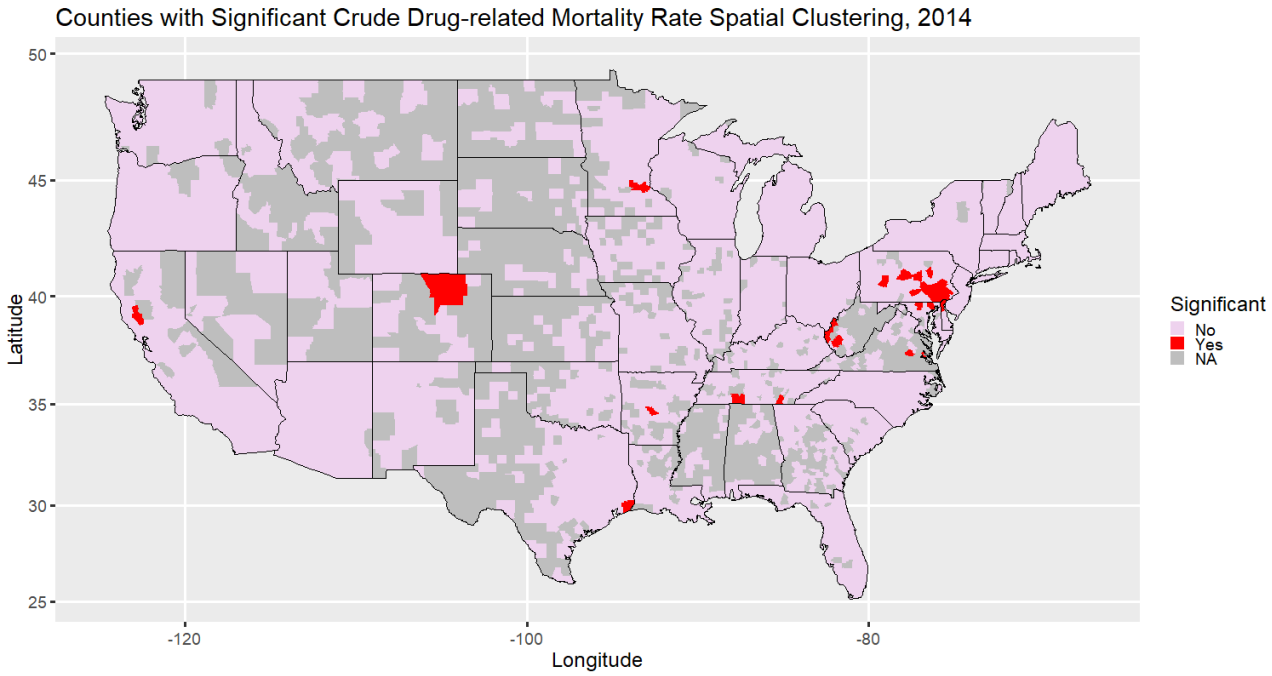

Figure 4: Counties found to have statistically significant spatial clustering within a 75 mile radius in 2014, according to at least one of the Getis-Ord G_i^* or Local Moran's I_i.

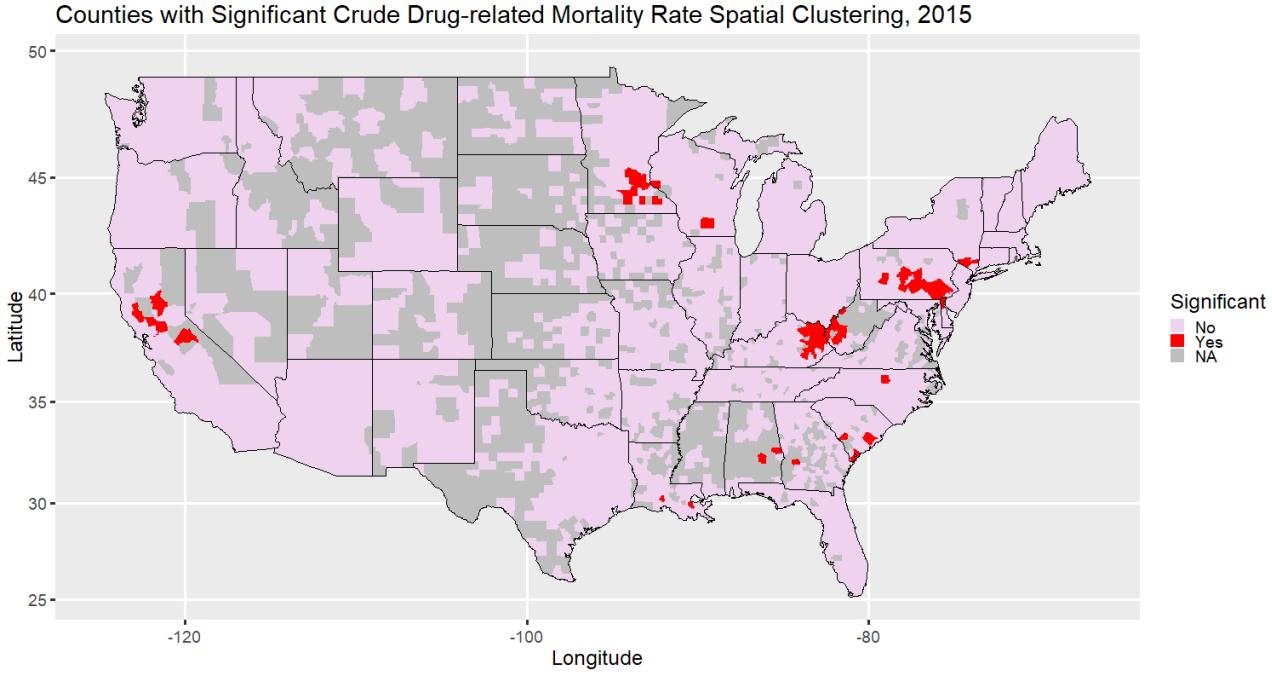

Figure 5: Counties found to have statistically significant spatial clustering within a 75 mile radius in 2015, according to at least one of the Getis-Ord G_i^* or Local Moran's I_i.

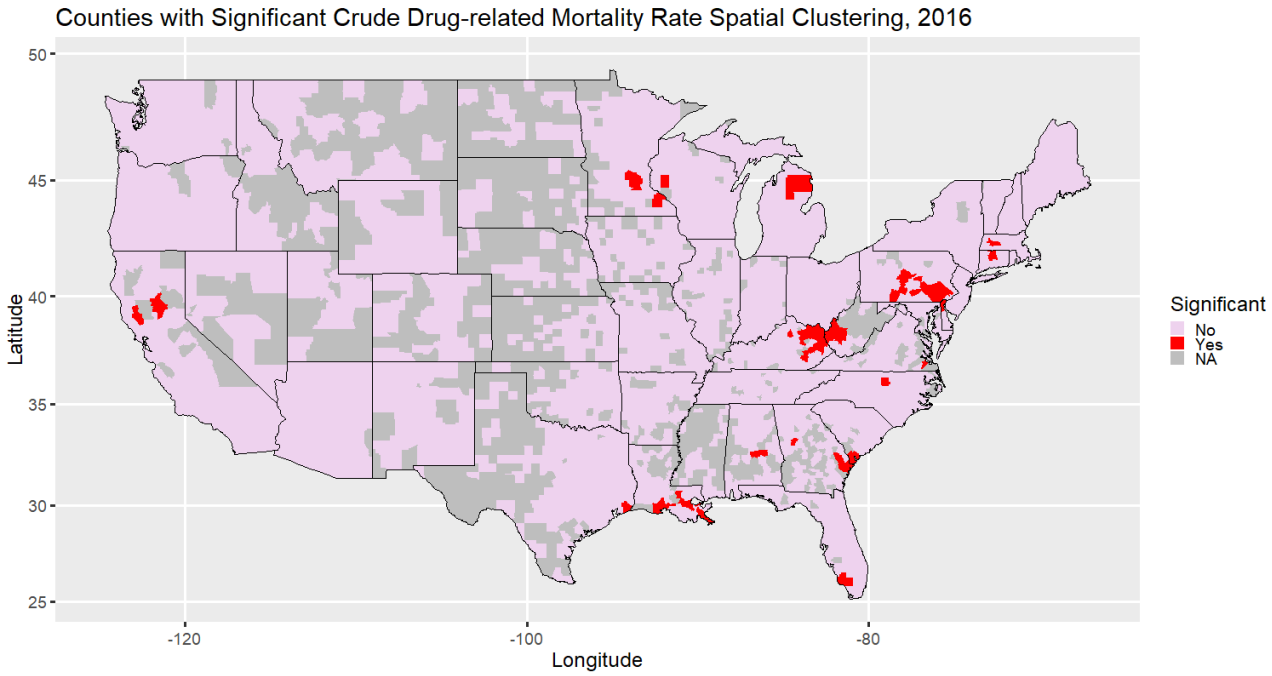

Figure 6: Counties found to have statistically significant spatial clustering within a 75 mile radius in 2016, according to at least one of the Getis-Ord G_i^* or Local Moran's I_i.

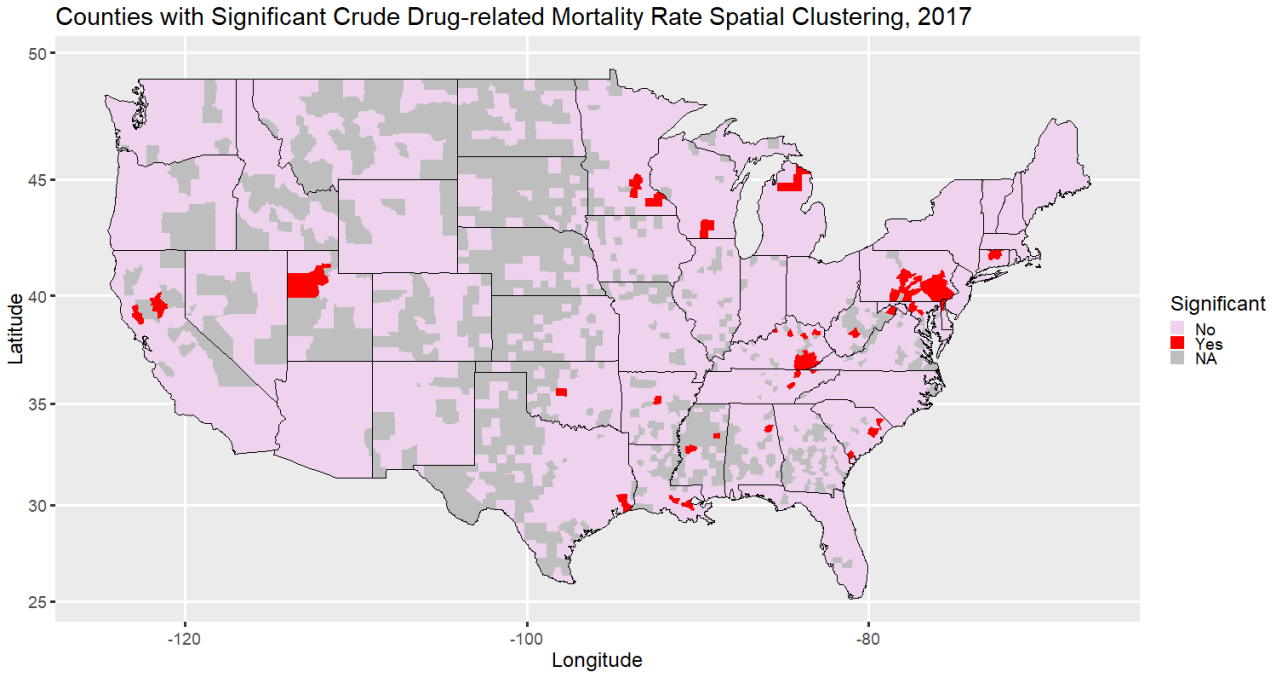

Figure 7: Counties found to have statistically significant spatial clustering within a 75 mile radius in 2017, according to at least one of the Getis-Ord G_i^* or Local Moran's I_i.

A couple characteristics stand out from these maps. One is that the evidence for spatial clustering in Pennsylvania has been fairly resilient over time, while other areas have exhibited evidence for spatial clustering coming and going. This lends more to the idea that it would be interesting to further investigate which underlying factors contribute to this spatial clustering of low drug-related mortality rates in Pennsylvania.

It is also worth noting that the evidence for statistically significant spatial clustering in eastern Kentucky and northern Michigan is fairly recent, mostly appearing in 2016 and 2017. These are areas suffering from relatively high crude drug-related mortality rates such that more resources may be needed to hamper further increases in drug abuse mortality around these areas. Similar may be true for southern Louisiana and some counties in northern California, but the amount of clustering seems to be smaller relative to Kentucky and Michigan.

More formal methods for spatiotemporal clustering exist and some discussion of them can be find in sources such as [12, 8].

B Additional Notes on Code

As mentioned in Section 4.2, the code used to run the analyses here is provided on `https://github.com/matthewquinn1/BST245`. Please note that this is more so a workflow specific to this project than a general package. However, there are general functions for operations, like computing the distance matrix and performing the G_i^* and I_i computation with their respective z-tests, that can be sourced and easily reused. Much of the data cleaning and mapping is specific to the data used here, though, so those operations may be more difficult to extract.