Deaths of Drug overdose in Ohio

Xinyu Gao (1828056)

Department of Statistics, University of Washington

gaoxinyu@uw.edu

Abstract

Deaths caused by drug overdose is an issue that concerns the public, and drug abuse is sometimes correlated with HIV. In this study, we will apply spatial lag model (model 2) to obviate spatial dependence among residuals from model 1 detected by Moran's I test (Table 4), and to find how strong the relationship between the death and this virus is in Ohio state. Model 2 requires the construction of spatial connectivity and spatial weight matrix, which are simply illustrated by a small area "ABHPS" (Fig 7.1). By fitting model 2, we can arrive at the conclusion that positive relationship between death rate due to drug abuse and HIV rate does exist in Ohio (see in Table 6).

Key words: Moran's I test, Spatial dependence, Spatial Lag Model

Introduction

In a health report of Ohio state in 2018, health behavior is an important index measured by the count of deaths caused by overdose drug and some other factors, and drug overdose accounts for almost 20% percentage of the weight, which indicates its role when measuring the health behavior. In another report in nationwide, the count of deaths due to drug overdose increases dramatically in recent years, rising from 36010 cases to 70237 during the past 10 years.

Many researches have involved the the reasons about this kind of death, and a common cause is related to HIV. From these researches ([5] [6]), people with HIV are often exposed to opioid medications during their HIV care experience; others may continue to use illicit opioids despite their disease status. In either situation, there may be a heightened risk for nonfatal or fatal overdose.

Although many works have been done about the relationship between drug overdose death and HIV, their fields of interest are based on biological and behavioral factors, while our study of interest is to find the association of them in a geographical scale (in this study, Ohio state and its 88 counties are our subjects). At the same time, introducing spatial methods to this academic area is also a promising direction due to the development of spatial data collecting.

Data description and map visualization

We download the health-related data from http://www.countyhealthrankings.org /ranking /data/OH and saved as “2018 County Health Rankings Ohio Data”, and downloaded the shape-file data from https://canvas.uw.edu /courses/1254570/pages /lecture-notes-and-exercises, and saved as “ohio_map”. Then all of our coding are based on R with version of 3.5.1.

First, we read the ohio map by function “readShapePoly”, and we text the order of county ranking by their first character (see in **Fig 1**)

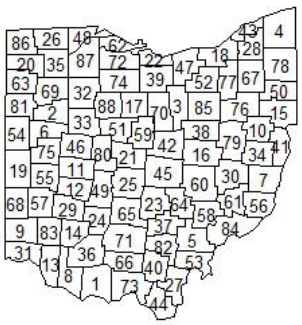

Fig 1: A county map of Ohio with order

We can see the counties are ordered from 1 to 88, so there are obviously 88 counties in Ohio state, and our studying area is restricted to these 88 counties, and our data collecting will focus on the health-related data within these counties. One should note that the county data collected from the website are ordered by FIPs (Federal information Processing) code (from the smallest to the largest and see in **Table 1**), while FIPs in the shape files are ordered differently, so we must match them before modeling. Fortunately, “match” function works effectively.

Table 1: Ohio counties with their FIPs and order

Ohio County	FIPs	Order
Adams	39001	1
Allen	39003	2
Ashland	39005	3
Ashtabula	39007	4
Athens	39009	5
...	...	...

Then we will take a look at our data. Since we want to find a relationship between the number of death due to overdose drug and the HIV prevalence, we will simply view them in the level of county and draw the histogram (Fig 2.1 and Fig 2.2) first.

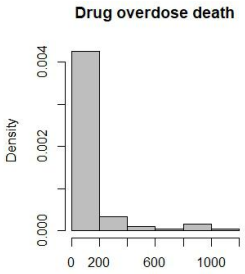

Fig 2.1: Hist of Drug overdose death

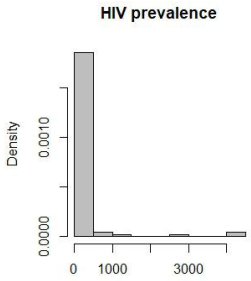

Fig 2.2: Hist of HIV prevalence

From the histogram, it is clear that most of the counties are limited in a range with regard to both variables, with 0-200 deaths and 0-500 cases respectively. We can view these individual data in a county level by mapping the counts in each county (Fig 3.1 and Fig 3.2)

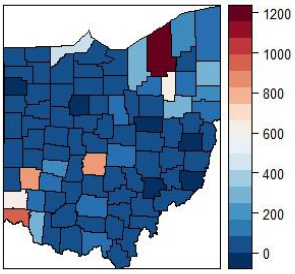

Fig 3.1: Map of Drug overdose death

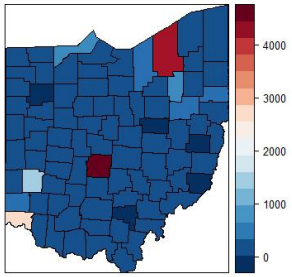

Fig 3.2: Map of HIV prevalence

Mapping the counts in a county level provides us with a better knowledge of which counties have significantly more counts or cases and which counties have few. It is clear, from the map, that the county at the right top of Ohio has dramatically more deaths and HIV cases, and we can easily find which county it is from the order map (Fig 1): Cuyahoga, with 1154 deaths and 4330 cases of HIV respectively. We can also find which counties “stands out” in the map: the center (Franklin) and the left corner (Hamilton).

Methods

Linear Regression Model

In the following part, we will detect the relationship between the death (drug abused) and the HIV prevalence in Ohio by linear regression model based on the assumption that there is no spatial dependence of the counts of death among 88 counties.

First, population should be taken into consideration because we have evidence to believe that the death (or detection of HIV) in a county might be dramatically high if the population of this county is significantly larger than other. Similarly, a county with small size of population is less likely to have such number of deaths or the high prevalence of HIV.

It is, therefore, essential to introduce population and treat it as the denominator, that is, the death and the prevalence of HIV will be divided by population. So next we will find the relationship between the rate of death (due to overdose drug) and the rate of HIV in Ohio.

The mapping of two rates are as follows:

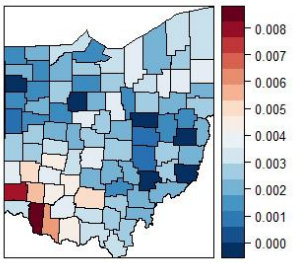

Fig 4.1: Rate of death (drug) in Ohio

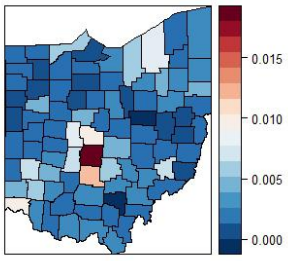

Fig 4.2: Rate of HIV in Ohio

The maps of two rates are very different from that of counts (or cases), which is mainly because we focus on the rate, rather than the counts, and each county can be compared in the same level.

We denote Y as the rate of death and X as the rate of HIV, and build a linear regression

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

Fitting (1), and the summary of OLS fit is as Table 6.

Spatial dependence

Mapping the residuals from OLS fit (see in Fig 5.1), we find an obvious spatial clustering of residuals, and from the TA plot (Fig 5.2) we can find that the model violate the assumptions of equal mean and constant variance, based on which we deduce that there exist the spatial dependence among residuals.

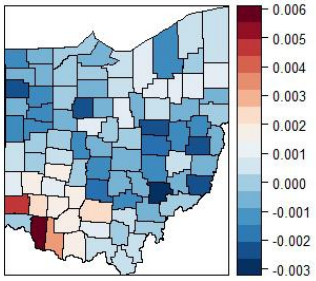

Fig 5.1: Residual (OLS) plot

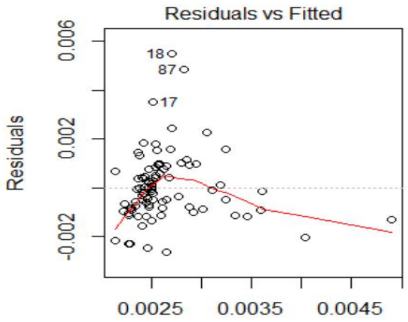

Fig 5.2: TA plot

Before checking the spatial dependence, we need to define spatial neighbors and introduce spatial weight matrix calculated by connectivity matrix. In order to clarify its mechanic more clear and straightforward, we will define the matrix within a small area.

Fig 6: Adams and its neighbors

The mark A, B, H, P, and S in the map (Fig 6) represent the five counties:

Marks	A	B	H	P	S
County	Adams	Brown	Highland	Pike	Scioto

Imagine that an area (assume its name is "ABHPS") only has these five counties, and we define counties share common boundary as "neighbors".

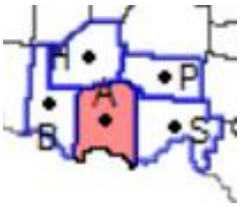

Fig 7.1: area "ABHPS"

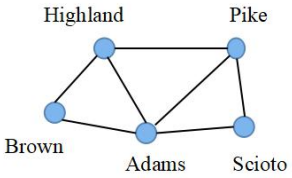

Fig 7.2: Connectivity map

Then we will define the connectivity matrix by the following principle :

$$C_{ij} = \begin{cases} 1, \ if \ S_i \ and \ S_j \ are \ neighbors \\ 0, otherwise \end{cases} \qquad (*)$$

(note: we assume the connectivity of county i and itself is 0, i.e. Cii=0).

Representing this connectivity relationship by solid lines (shown in **Fig 7.2**), and we will display the connectivity matrix as follows:

Table 2: The connectivity matrix of area "ABHPS"

	Adams	Brown	Highland	Pike	Scioto	# of neighbours
Adams	0	1	1	1	1	4
Brown	1	0	1	0	0	2
Highland	1	1	0	1	0	3
Pike	1	0	1	0	1	3
Scioto	1	0	0	1	0	2

For a certain county i, the sum of connectivity weight is 1, and we assign the equal weights to the neighbors of this county because we want to combine information about the connected observations. The weight matrix table will make it easier to see the building of "spatial lag" term:

Table 3: The weight matrix of area “ABHPS”

	Adams	Brown	Highland	Pike	Scioto
Adams	0	1/4	1/4	1/4	1/4
Brown	1/2	0	1/2	0	0
Highland	1/3	1/3	0	1/3	0
Pike	1/3	0	1/3	0	1/3
Scioto	1/2	0	0	1/2	0

What is the average value of “Drug overdose death” of Adams’ neighbors? From the Table 3, we can calculate it by the average number of death of “Brown”, “Highland”, “Pike”, and “Scioto”, i.e.

$$\overline{D}_{Adams(ne)} = \omega_{Adams} D_{Adams(ne)}$$

where $\overline{D}_{Adams(ne)}$ represents the average death of Adams’ neighbors, which is a “spatial lag” term, and ω_{Adams} represents the weights of Adams’ neighbors.

Moran’s I test

When building Moran’s I test, a strong clustering can be detected if a high value of Moran’s I appears, and this indicates that the residuals are similar with neighbors. The following table shows the results of Moran’s test on the residuals from the OLS fit.

$$I = \frac{n}{W} \frac{\sum_i \sum_j w_{ij}(r_i - \bar{r})(r_j - \bar{r})}{\sum_i (r_i - \bar{r})^2}$$

where $W = \sum_i \sum_j w_{ij}$ is the sum of weights of index i and j

By Moran’s I statistic, we can perform a test and obtain the results as follows:

Table 4: Moran’s I test

Moran I statistic	0.43729
Standard deviate	7.0118
p-value	1.176e-12

From the Table 4, Moran’s I statistic for these residuals is about 0.43, and the p-value is rather small, which rejects the null hypothesis of spatial independence, or specifically, this implies that there is a strong evidence that the death rate (due to overdose drug) of a county is highly dependent of that of its neighbors. Hence, we will take the spatial dependence into account and fit the model by introducing a spatial term in the following part.

Spatial Lag Model

Since it has come to the conclusion verified by small p-value of Moran's I test that there exists spatial dependence among residuals from OLS fit (without spatial term), we will divide the residuals ε into two parts: 1. a part does not have spatial dependence； 2. a part has spatial effect and caused by the responses of its neighbors Wy, where W is the spatial weight matrix. The part 2 looks similar with the construction of auto-regression in time series, and we assume it to be spatial lag term. The spatial lag model is shown as (2)

$$Y = \rho Wy + \beta_0 + \beta_1 X + \varepsilon \tag{2}$$

where ρ is the parameter of spatial lag term and β are the coefficients of the predictors. In R, we can fit this model by "lagsarlm" function, and get the fitted parameters and coefficients (see in Table 6). Meanwhile, we can also extract the residuals from model (2). Drawing a map of residuals and performing a Moran's I test will help us verify that the spatial dependence of residuals has been eliminated, which will all be covered in section of results.

Results

Followed by the previous section, we will first check the residuals from spatial lag model (2) by mapping and clarify carefully again by Moran's I test.

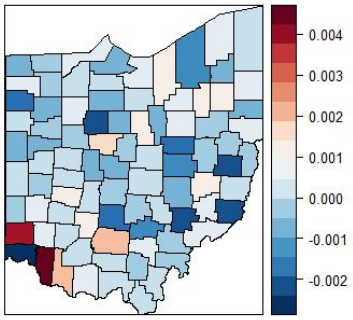

Fig 8: Residuals from spatial lag model

There seems no spatial clustering among residuals from the residual map (Fig 8), and double check by Moran's I test.

Table 5: Moran's I test

Moran I statistic	-0.11163
p-value	0.9411

Moran's I statistic for these residuals is about -0.11163, and the p-value is very large, which means that we have eliminated the spatial dependence of the residuals and shows that spatial lag model works very well.

Then we will display the summary of two fitted model: linear regression fit without spatial lag term and fit with that term in Table 6.

Table 6: The summary of fits

The summary of fit without spatial lag term from model (1)				
	Estimate	Std. Error	t value	p-value
(Intercept)	0.002147	0.0002319	9.261	1.45e-14***
Rate of HIV	0.152688	0.0549939	2.776	0.00674**
Significance Codes: '***' 0.001 '**'0.01 '*' 0.05				
F statistic: 7.709 on 1 and 86 DF, p-value: 0.006743				
AIC: **-901.353**				
The summary of fit with spatial lag term from model (2)				
	Estimate	Std. Error	t value	p-value
(Intercept)	0.00073106	0.000308	2.3738	0.0176
Rate of HIV	**0.0811719**	0.043577	1.8627	0.0625
Rho	0.61952	/	/	1.576e-08
AIC: **-931.310**				

Observations from Table 6 have all the information we want to acquire to compare and further evaluate two models, and the estimates of coefficients can tells us how strong a relationship between the response and the predictors.

First, there does exist spatial dependence among residuals from model (1) based on the Moran's I test, which indicates that the assumption of independent error in OLS fit is violated, then model (2) eliminates this effect so that it can be reasonable fitted by OLS, which shows a better application of model (2) when spatial dependence appears. Model (2) performs better not just lies in the solve of spatial dependence, but in lower AIC value (-931.310 by model (2) less than -901.353 by model (1)).

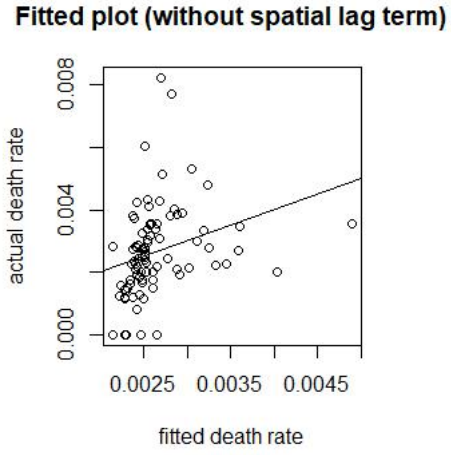

Fig 9.1: Fitted plot (without lag term)

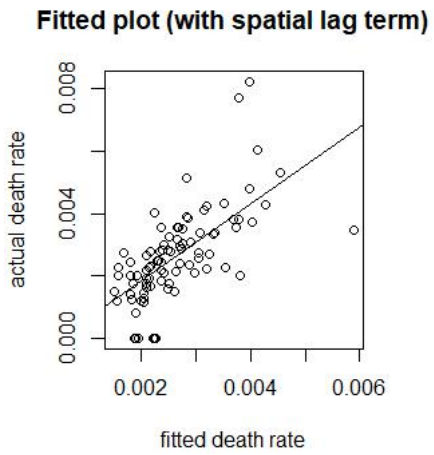

Fig 9.2: Fitted plot (with spatial lag term)

Furthermore, two models can be compared by their fitted plot (see in Fig 9.1 and Fig 9.2). and model (2) fits obviously better, and all of above reinforces our belief that the spatial lag term in model (2) really adds some significant information to the distribution of the response.

Then we turn to the results of model (2). First, the coefficient of spatial lag term is about 0.62, which shows a significant role of neighbors in fitting the model. Then we can get another two conclusions based on the estimates of intercept and slope conclusions: 1. the averaging death rate in a county of Ohio caused by overdose drug is 0.073% when there is no HIV cases; 2. there exists a positive relationship between death rate (due to drug abuse) and HIV rate in Ohio: if HIV rate increases by 10%, the averaging death rate will rise by about 1%.

Discussion

Our study have found evidence to support our hypothesis that there exists a positive relationship between death rate (due to drug abuse) and HIV rate. This finding strengthens the prior reports which suggested this connection but limited by the lack of spatial analysis. While our study do not find evidence of effect of other factors on the death rate, such as the number of condoms used, unemployment rate, median income, etc., future efforts will be applied to detect these effects.

We can find an effective application of spatial lag model (i.e. model (2)) when a linear regression model (1) comes to a failure due to spatial dependence of residuals. We should note that the failure results in the unconvinced estimates of coefficients, which means that the relationship we have built by model (1) between the death rate (drug abuse) and HIV rate is less efficient than that of model (2), although HIV rate is more significant in model (1) than in model (2).

However, our model is still subject to some limitations. First, our analysis is limited by insufficient number of data because our study is restricted in Ohio state and there are 88 counties, we can probably conduct our study on a smaller scale (divide area based on zip code). Besides that, the prediction is highly dependent of the predictor and the response of nearby areas, so the lack of information nearby may decrease the confidence of predicted value. Then, the way we define neighbors makes the spatial connectivity matrix simple and relatively easy to realize, but in reality, counties do not share a common boundary but very close in geographically may still have an nontrivial effect of the response on each other, and the same issue with the weights: when studying a certain county, we might assign zero weight to a county actually contributes most just because of no common boundary. It is, hopefully, resolved by redefining the neighbors or by using another set of weights, for example, the weights can be determined by the Gaussian function on the distance between the geographical centers.

References

[1] Michael D. Ward, Kristian Skrede Gleditsch. An Introduction to Spatial Regression Models in the Social Sciences.

[2] https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates

[3] Bivand, R., Pebesma, E., and Gomez-Rubio, V. (2013). Applied Spatial Data Analysis with R, 2nd Edition. Springer, New York

[4] Wakefield, J. (2007). Disease mapping and spatial regression with count data. Biostatistics, 8, 158–183

[5] Traci C. Green, Samuel K. McGowan (2012). HIV infection and risk of overdose: a systematic review and meta-analysis. AIDS. 26(4):403–417, FEB 2012

[6] Eskild A, Magnus P, Samuelsen SO, Sohlberg C, Kittelsen P. Differences in mortality rates and causes of death between HIV positive and HIV negative intravenous drug users. Int J Epidemiol. 1993;22:315–320.