For office use only	Team Control Number	For office use only
T1 ________________	**1910381**	F1 ________________
T2 ________________		F2 ________________
T3 ________________	Problem Chosen	F3 ________________
T4 ________________	**C**	F4 ________________

2019
MCM/ICM
Summary Sheet

How to counter the opioid crisis?

Summary

The United States is experiencing a nationwide crisis in the use of synthetic and non-synthetic opioids to treat and manage pain (legal, prescription use).

First, in order to determine the changes in the number of drug cases in all counties of each state the counties in each state, we divided all in the five states is divided into three groups according to severity of drug usage. Since the number of drug cases in each county will be affected by other counties and the varying limitations, they will increase or decrease with a certain probability. Basing on the analysis above, we carry out differential equation model (Drug Spread Model) to describe the change of the number of counties in three groups. After that we replace the derivative with discrete difference, applying least sum of square method to calculate unknown parameters in the model. We observe figures of the functions and obtained when the drug cases started in the states, the trend of the number of counties in various states in the future and the threshold level of drug identification. Similarly, we chose heroin as the research object and obtained the features of heroin transmission between states and the time when heroin cases began to occur in each state, thus judging the earliest heroin cases in Ohio.

Then we study the relationship between socioeconomic indicator data and drug cases data. After cleaning the data, we applied LASSO regression based on cross validation and finally obtained 13 factors. From the 13 contact factors, we selected two indicators that have a significant impact on the number of drug cases in the county each year: the number of men who did not have a junior high school diploma at the age of 25 and the number of unmarried women with children. Based on the Drug Spread Model, we add a linear combination of two index functions to each parameter to achieve the purpose of correction. We still applied the same method as the one in solving Drug Spread Model to find the optimal parameters and finally the modified coefficients are substituted into Drug Spread Model and we obtained Modified Drug Spread Model.

Finally, combined with the above analysis we have identified a strategy of increasing funding for education to deal with the drug crisis. When the indicator of reaches the level of 80

Keywords: LASSO; Cross Validation; Differential equation;Numerical fitting model

How to counter the opioid crisis?

January 29, 2019

Summary

The United States is experiencing a nationwide crisis in the use of synthetic and non-synthetic opioids to treat and manage pain (legal, prescription use).

First, in order to determine the changes in the number of drug cases in all counties of each state the counties in each state, we divided all in the five states is divided into three groups according to severity of drug usage. Since the number of drug cases in each county will be affected by other counties and the varying limitations, they will increase or decrease with a certain probability. Basing on the analysis above, we carry out differential equation model (Drug Spread Model) to describe the change of the number of counties in three groups. After that we replace the derivative with discrete difference, applying least sum of square method to calculate unknown parameters in the model. We observe figures of the functions and obtained when the drug cases started in the states, the trend of the number of counties in various states in the future and the threshold level of drug identification. Similarly, we chose heroin as the research object and obtained the features of heroin transmission between states and the time when heroin cases began to occur in each state, thus judging the earliest heroin cases in Ohio.

Then we study the relationship between socioeconomic indicator data and drug cases data. After cleaning the data, we applied LASSO regression based on cross validation and finally obtained 13 factors. From the 13 contact factors, we selected two indicators that have a significant impact on the number of drug cases in the county each year: the number of men who did not have a junior high school diploma at the age of 25 and the number of unmarried women with children. Based on the Drug Spread Model, we add a linear combination of two index functions to each parameter to achieve the purpose of correction. We still applied the same method as the one in solving Drug Spread Model to find the optimal parameters and finally the modified coefficients are substituted into Drug Spread Model and we obtained Modified Drug Spread Model.

Finally, combined with the above analysis we have identified a strategy of increasing funding for education to deal with the drug crisis. When the indicator of reaches the level of 80

Keywords: LASSO; Cross Validation; Differential equation;Numerical fitting model

Contents

1 Introduction

The United States is experiencing a nationwide crisis in the use of synthetic and non-synthetic opioids to treat and manage pain (legal, prescription use). Try to solve the following problem：

- Part 1： Establish a mathematical model to characterize the spread of drug cases in and between the five states, and use the model to determine when and where to reach drug identification thresholds level?
- Part 2:Is drug’ s use or trends-in-use associated with any of the U.S. Census socio-economic data provided? If so, modify the model in Part 1 to include any important factors from this data set.
- Part 3: Develop a viable strategy to deal with the drug crisis and then use your model to test the feasibility of the strategy

1.1 Assumptions

- Assuming that drug incidents are the discovery of drug addicts
- Assuming that the impact is positive

2 Drug spread model

2.1 Preliminary analysis

2.1.1 Visualization of Original Data

Since most of the detailed data of synthetic opioid and heroin incidents for each county are missing, we first analyzed the number of total drug reports in all counties only.

First, we visualize the data in ”MCM_NFLIS_Data.xlsx” to make a preliminary observation of the changes in the data. We use the latitude and longitude data in Google Map and use Google Map API interface [1] to connect to Google Maps data server in Matlab get the latitude and longitude information of each county in ”MCM_NFLIS_Data.xlsx” and then realize the data visualization through Matlab programming. (The code is shown in the appendix.) The result of 2016 is shown in the figure below, other result that between 2010-2015 are included in Appendix C . In order to understand the time-dependent changes of the ”TotalDrugReportsCounty” data in 2010-2017 more intuitively, we made a dynamic change video using Matlab. The Matlab code for video production is shown in Appendix D.

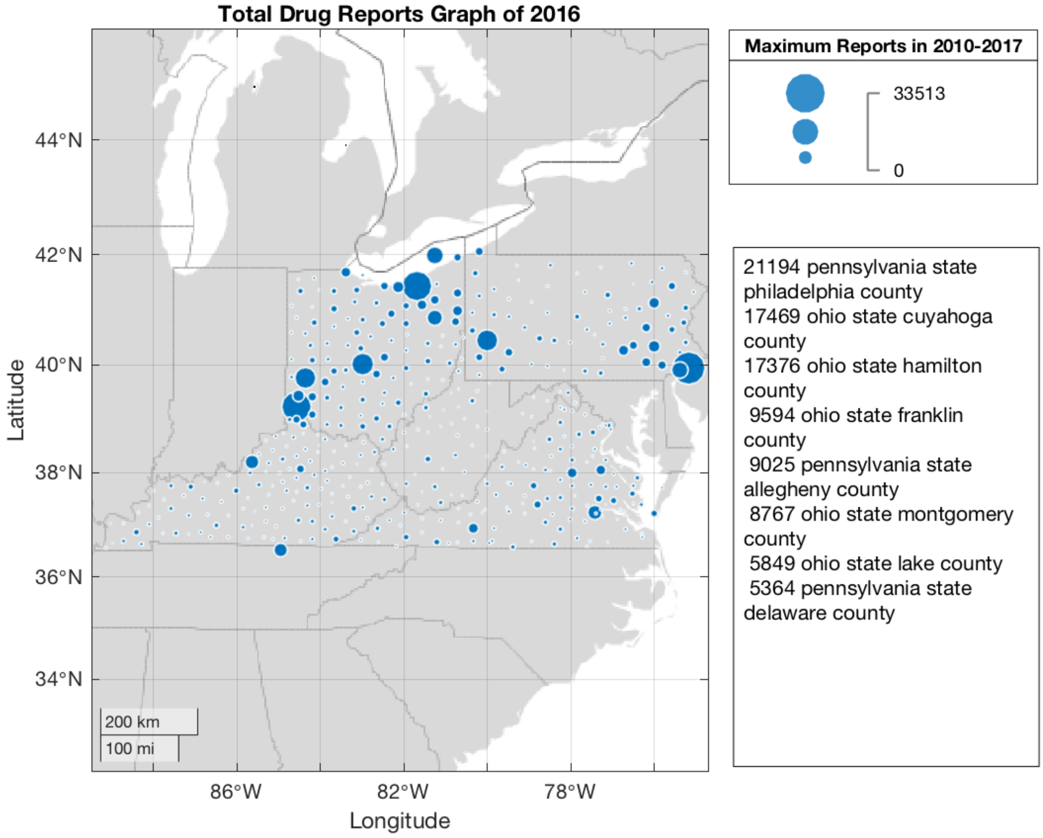

Total Drug Reports Graph of 2016
Maximum Reports in 2010-2017
33513
0
21194 pennsylvania state philadelphia county
17469 ohio state cuyahoga county
17376 ohio state hamilton county
9594 ohio state franklin county
9025 pennsylvania state allegheny county
8767 ohio state montgomery county
5849 ohio state lake county
5364 pennsylvania state delaware county
44°N
42°N
40°N
38°N
36°N
34°N
Latitude
200 km
100 mi
86°W
82°W
78°W
Longitude

Figure 1: Total Drug Reports County in 2016

2.1.2 Preliminary analysis

Based on the result of visualization, we can see that the number of Total Drug Reports in the four counties —-Pennsylvania State Philadelphia County , Ohio State Hamilton County , Pennsylvania State Allegheny County , Ohio State Montgomery County —-was more than 5000 in 2010-2017. The number of Total Drug Reports in Ohio State Cuyahoga County in 2010-2017 was more than 5000 in 6 years and its growth was the fastest of all counties. The statistical results of the total number of Total Drug Reports over 5000 are shown in the table below.

2.2 Model construction

2.2.1 Modeling

In general, when a drug identification occurs in a county, it may affect the number of such cases in other counties. (The number of such cases may increase or decrease.) It is easy to understand that the possibility of the occurrence of such cases is closely related to the position of other county—-First, when two counties are located in the same state, the one of them can easily affect the other one. (Of course, for a specific county, a county in another state can affect it as well, but in less strength.) Second, the shorter the distance between a county and another

Table 1: results of the total number of Total Drug Reports over 5000

Value	Count
Pennsylvania State Philadelphia County	8
Ohio State Hamilton County	8
Pennsylvania State Allegheny County	8
Ohio State Montgomery County	8
Pennsylvania State Bucks County	3
Kentucky State Jefferson County	2
Ohio State Cuyahoga County	6
Ohio State Franklin County	6
Ohio State Lake County	2

county, the greater the mutual influence they directly receive, the greater the possibility drug identifications occur at the same time in these two counties. We create matrix $G = (g_{ij})_{5\times 5}$ to describe the impact level between two counties, where g_{ij} means the impact level of drug cases occurring in counties in the state j towards counties in state i. The farther the distance, the lower the level of influence. We define

$$g_{ij} = \begin{cases} \frac{\frac{1}{d_{ij}}}{\sum\limits_{i,j}\frac{1}{d_{ij}}} i \neq j \\ 1 i = j \end{cases} \tag{1}$$

Apparently, matrix G is a symmetric matrix.

At the same time the possibility of a county with a drug case affecting the number of cases in other counties is also related to the number of drug cases in the county. Therefore we first divide the counties in each of the five states into three categories according to the number of cases in the county/the total number of cases in five states

We divide all counties in these five states into three groups:

1. In these five states, there are some counties where no drug cases occur at specific times. We divide these counties into a separate group, denoted as group A.
2. All counties different from Group A are identified with drug cases at that time, but the severity of them differ from each other. We assume that the rest counties can be divided into two groups according to the severity. We define drug infected level m, which is the ratio between the number of cases in the county and the total number of cases in five states. The remaining counties are divided by the average of drug infected level m : the counties below the average(0.0023) belong to one group, which is the moderate drug-using group, denoted Group B. The counties above the average belong to one group, which is the severe drug-using group, denoted Group C.

For a county s in state i:

1. If county s belongs to Group A, the number of cases in s will not influence the number of cases in other counties.
2. If county s belongs to Group B, the number of cases in s will only influence the number of cases in counties that belong to Group A.

3. If county s belongs to Group B, the number of cases in s will only influence the number of cases in counties that belong to Group A.

To simplify the problem, we need to make the following notations:

Table 2: results of the total number of Total Drug Reports over 5000

Value	Count
t	Time
$N_{Ci}(t)$, $N_{Ci}(t)$, $N_{Ci}(t)$	The number of three types of counties at time t in state i
$delta_B$	The removal rate of Group B counties
$delta_C$	The removal rate of Group C counties
C_A	The probability that Group C counties are converted into Group A counties, on the condition that they are removed from Group C
C_B	The probability that Group Ccounties are converted into Group B counties, on the condition that they are removed from Group C
b	The possibility that a Group A county coverts into a Group B county due to the influence of B.
c_1	The possibility that a Group B county covert into a Group C county due to the influence of C.
c_21	The possibility that a Group A county coverts into a Group B county due to the influence of C.
c_22	The possibility that a Group A county coverts into a Group C county due to the influence of C.
$T_{ABi}(t)$	The number of Group A counties converted into Group B counties in state i at time t.
$T_{AC}(t)$	The number of Group A counties converted into Group B counties in state i at time t.
$T_{BC}(t)$	The number of Group A counties converted into Group B counties in state i at time t.

In addition to being affected by counties with a large number of other cases, the number of cases in a county may be reduced due to restrictions imposed by national policies. This probability is related to the number of cases itself. For the three types of counties we have defined we can see that we have a basis for dividing the three types of counties. We can think that the probability of the number of counties in any one of the counties being converted to another type of counties is a fixed number.

Now we analyze the change of the number of counties in Group A, B, C separately.

$T_{ABi}(t)$is composed by two parts：

1. The number of Group A counties in the state I that are converted into Group B counties by the influence of Group B counties in all 5 states

2. The number of Class$A$ counties in the state I that are converted into Group B counties by the influence of Group B counties in all 5 states

$$T_{ABi}(t) = \sum_{j=1}^{5} g_{ij} N_{Bj}(t) N_{Ai}(t) b + \sum_{j=1}^{5} g_{ij} N_{Cj}(t) N_{Ai}(t) c_{21} \tag{2}$$

$T_{AC}(t)$is the number of Group A counties in the state I that are converted into Group B counties by the influence of Group B counties in all 5 states

$$T_{ACi}(t) = \sum_{j=1}^{5} g_{ij} N_{Cj}(t) N_{Ai}(t) c_{22} \tag{3}$$

$T_{BC}(t)$is the number of Group B counties in the state I that are converted into Group C counties by the influence of Group C counties in all 5 states

$$T_{BCi}(t) = \sum_{j=1}^{5} g_{ij} N_{Cj}(t) N_{Bi}(t) c_1 \tag{4}$$

$\delta_B N_{Bi}(t)$,$C_A \delta_C N_{Ci}(t)$are the number of Group B and Group C counties converted into Group A counties after they reduce the number of cases , that is, the net increase in the number of Class A counties. Consider the increase in the number of counties from t to $t+\Delta t$:

$$N_{Ai}(t+\Delta t) - N_{Ai}(t) = (\delta_B N_{Bi}(t) - T_{ABi}(t) - T_{ACi}(t) + C_A \delta_C N_{Ci}(t))\Delta t \tag{5}$$

Then we get the differential equations:

$$\frac{dN_{Ai}(t)}{dt} = \delta_B N_{Bi}(t) - T_{ABi}(t) - T_{ACi}(t) + C_A \delta_C N_{Ci}(t) \tag{6}$$

Similarly, we get the growth rate of Group B, C counties

$$\begin{cases} \frac{dN_{Ai}(t)}{dt} = \delta_B N_{Bi}(t) - T_{ABi}(t) - T_{ACi}(t) + C_A \delta_C N_{Ci}(t) \\ \frac{dN_{Bi}(t)}{dt} = T_{ABi}(t) - \delta_B N_{Bi}(t) - T_{BCi}(t) + C_B \delta_C N_{Ci}(t) \\ \frac{dN_{Ci}(t)}{dt} = T_{BCi}(t) - \delta_C N_{Ci}(t) + T_{ACi}(t) \end{cases} \tag{7}$$

where

$$C_A + C_B = 1 \tag{8}$$

Finally, we get the Drug Spread Model

$$\begin{cases} \frac{dN_{Ai}(t)}{dt} = \delta_B N_{Bi}(t) - T_{ABi}(t) - T_{ACi}(t) + C_A \delta_C N_{Ci}(t) \\ \frac{dN_{Bi}(t)}{dt} = T_{ABi}(t) - \delta_B N_{Bi}(t) - T_{BCi}(t) + C_B \delta_C N_{Ci}(t) \\ \frac{dN_{Ci}(t)}{dt} = T_{BCi}(t) - \delta_C N_{Ci}(t) + T_{ACi}(t) \end{cases} \tag{9}$$

2.2.2 Obtaining parameters

From the first document we can get the number of counties in Group A, B and C in each state in 2010-2017. So we let $\Delta t = 1$ and convert equation (1)(2) into discrete forms:

$$\begin{array}{l} N_{Ai}(t+1) - N_{Ai}(t) \approx \delta_B N_{Bi}(t) - T_{ABi}(t) - T_{ACi}(t) + C_A \delta_C N_{Ci}(t) \triangleq f_{i1} \\ N_{Bi}(t+1) - N_{Bi}(t) \approx T_{ABi}(t) - \delta_B N_{Bi}(t) - T_{BCi}(t) + C_B \delta_C N_{Ci}(t) \triangleq f_{i2} \\ N_{Ci}(t+1) - N_{Ci}(t) \approx T_{BCi}(t) - \delta_C N_{Ci}(t) + T_{ACi}(t) \triangleq f_{i3} \\ C_A + C_B = 1\, i = 1,2,3,4,5 \end{array} \tag{10}$$

We denote,

$$r_{ij} = \begin{cases} [f_{k1} - (N_{Ak}(t+1) - N_{Ak}(t))]_{t=t_j} i = 3k+1 \\ [f_{k2} - (N_{Bk}(t+1) - N_{Bk}(t))]_{t=t_j} i = 3k+2 \quad i = 1...15, j = 1...7 \\ [f_{k3} - (N_{Ck}(t+1) - N_{Ck}(t))]_{t=t_j} i = 3k+3 \end{cases} \tag{11}$$

The r_{ij} is called the residual which characterizes the error of the numerical fitting, where t_j is the value of the independent variable t.

To compute the parameters $\delta_B, \delta_C, b, c_1, c_{21}, c_{22}, C_A$, we use the least squares criterion to determine the value of the parameter so that the sum of the squares of the residuals is minimal.

$$R = \sum_{i=1}^{15} \sum_{j=1}^{7} {r_{ij}}^2 = \min \tag{12}$$

Since the values of all the parameters are in $[0, 1]$, the parameters can be obtained by solving the following constraint optimization problem:

$$\begin{aligned} &\min R = \sum_{i=1}^{15} \sum_{j=1}^{7} {r_{ij}}^2 \\ &s.t.\ \delta_B, \delta_C, b, c_1, c_{21}, c_{22}, C_A \in [0, 1] \end{aligned} \tag{13}$$

Using Matlab function "fmincon()" , we can compute the estimated value of all parameters:

$$\begin{cases} \delta_B = 0.0365 \\ \delta_C = 0.0277 \\ b = 0.0031 \\ C_1 = 0.0042 \\ C_{21} = 1.0566 \times 10^{-8} \\ C_{22} = 1.1231 \times 10^{-9} \\ C_B = 1 \\ C_A = 0 \end{cases} \tag{14}$$

In this model, each parameters has its practical meaning, so we can recognize the way and features of drug cases "spreading" between different states.

In this model, each parameter has its practical meaning, so we can recognize the way and features of drug cases "spreading" between different states.

For the counties in Group C, there is a possibility of $\delta_C = 0.0277$ that counties are transformed into other types of counties. More accurately, there is a possibility of $\delta_C \times C_B = 0.0277$ that Group C counties are transformed into Group B counties, which cannot be directly converted into Group A counties, however. For Group B counties, there is a possibility of $\delta_B = 0.0365$ that they are transformed into Group A counties.

The number of counties in each of the three categories of states is related to the number of counties in the three states of the five states since the increase in the number of cases per county is affected by the number of cases in other counties. For example, for Kentucky, the number of Group A counties converted into Group B counties is

$$\begin{aligned} T_{AB1}(t) &= \sum_{j=1}^{5} g_{1j} N_{Bj}(t) N_{A1}(t) b + \sum_{j=1}^{5} g_{1j} N_{Cj}(t) N_{A1}(t) c_{21} \\ &= \sum_{j=1}^{5} g_{1j} N_{Bj}(t) N_{A1}(t) \times 0.0031 + \sum_{j=1}^{5} g_{1j} N_{Cj}(t) N_{A1}(t) \times 1.0566 \times 10^{-8} \end{aligned} \tag{15}$$

The number of Group A counties converted into Group C counties is

$$T_{AC1}(t) = \sum_{j=1}^{5} g_{1j} N_{Cj}(t) N_{A1}(t) \times 1.1231 \times 10^{-9} \tag{16}$$

The number of Group B counties converted into Group C counties is

$$T_{BC1}(t) = \sum_{j=1}^{5} g_{1j} N_{Cj}(t) N_{B1}(t) \times 0.0042 \tag{17}$$

2.3 Solution of model

2.3.1 Trend of future

We have obtained the values of the given 7 parameters approximated by the data in "MCM_NFLIS_Data.x Now we get the linear ordinary differential equations. We regard the data in 2010 as the initial value condition, and use the function "dsolve()" in Matlab to solve the differential equations, so that we get a set of functions of the number and time of each of the three groups in each state (Due to the limitation of the paper' s space, the solutions is shown in the appendix.). We let $t \in [0, 50]$,and draw the figures of the functions.

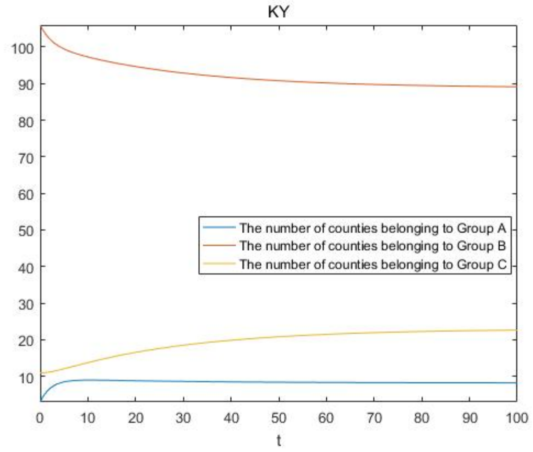


Figure 2: KY 1

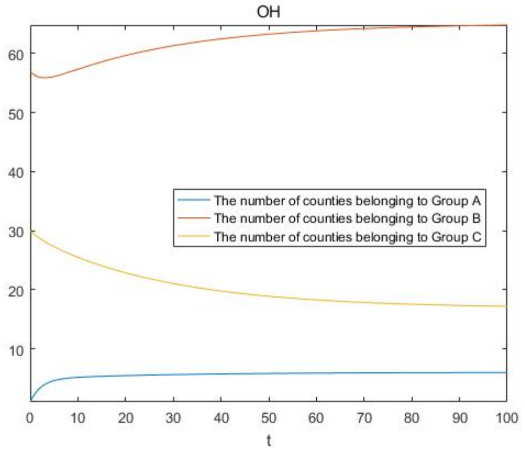


Figure 3: OH 1

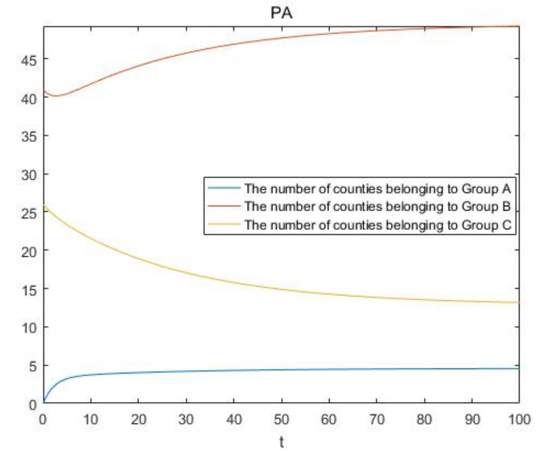


Figure 4: PA 1

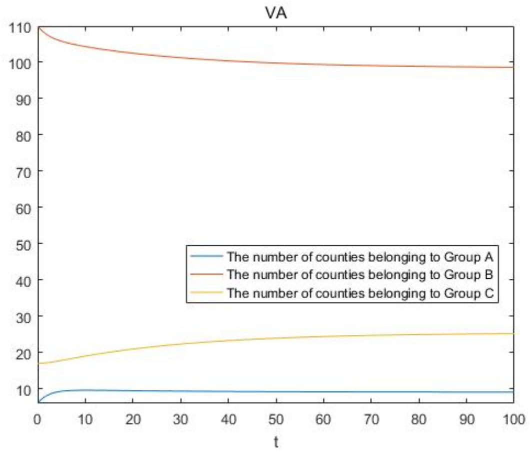


Figure 5: VA 1

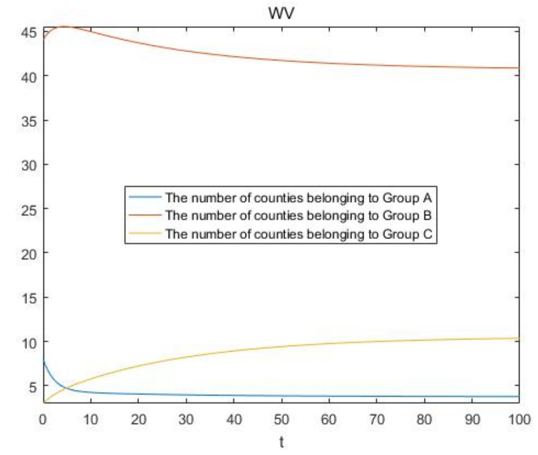


Figure 6: WV 1

From the five figures above, we can find that

1. In the next 40 years the number of Group C counties in Kentucky, Virginia and West Virginia has a certain increase about 10%; The number of Group A counties remains basically the same. The number of Group B counties has decreased. As the number of counties in each state unchanged, the decrease in the number of Group B counties approximately equals to the increase in the number of Group C counties. For Ohio and Pennsylvania, the number of Group A counties remains basically the same, but the number of Group C counties is decreasing and the number of Group B counties is increasing.

   We can roughly understand that without policy interventions, some of the severe drug-using counties in Ohio and Pennsylvania will be converted to moderate drug-using counties. Some of the moderate drug-using counties in Kentucky, Virginia and West Virginia will be transformed into severe drug-using counties. Therefore, the government should focus on drug cases in Kentucky Virginia and West Virginia.

2. The number of counties in the three groups of five states tends to be stable over time. Since the number of counties in the three categories is directly related to the number of

drug cases in each county, we have reason to believe that when the number of three groups of counties in a state becomes stable, the drug identification has reached the threshold level. From the above five figures we found that Virginia was the first to reach a stable state and it reached a steady state in the 60th year. Therefore, we can consider that in the 60th year —that is, the year 2070—the drug identification in Virginia reached a threshold level.

2.3.2 Finding where Heroin started

In Section 2.2.2 , We have calculated the number of Group A, B and C counties in each state as a function of time according to the number and proportion of all drug cases. Next we counted the number of heroin cases in each county in the first document and the number of counties A, B and C classified by number and proportion of heroin cases in each state in 2010-2017.Here we get the optimal value of all parameters:

$$\begin{cases} \delta_B = 0.0698 \\ \delta_C = 0.2259 \\ b = 0.0024 \\ C_1 = 0.0028 \\ C_{21} = 0.0066 \\ C_{22} = 2.1404 \times 10^{-6} \\ C_B = 1 \\ C_A = 0 \end{cases}$$

Then we use Matlab function to solve the corresponding ordinary differential equations to obtain the function of the number and time of each county in each of the three groups of each state. We let $\in [-10, 20]$, and get the figures of the functions:

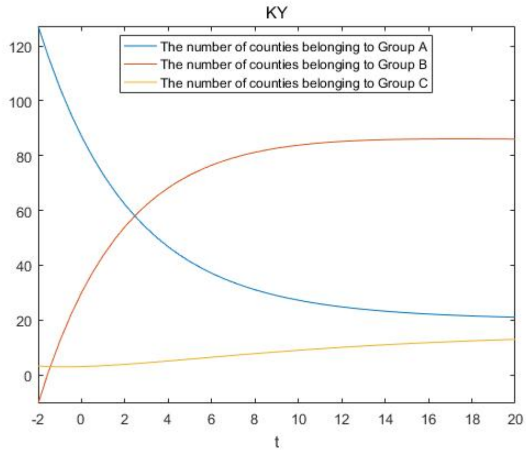


Figure 7: KY 2

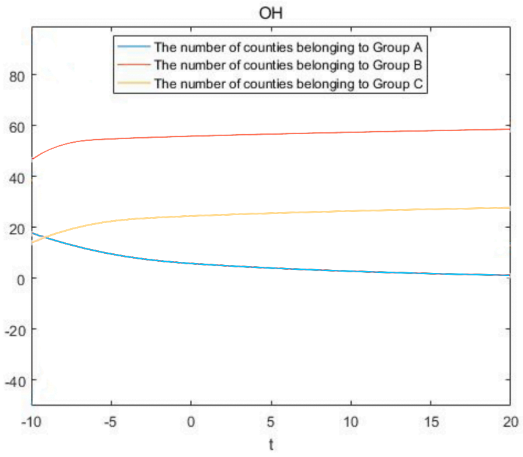


Figure 8: OH 2

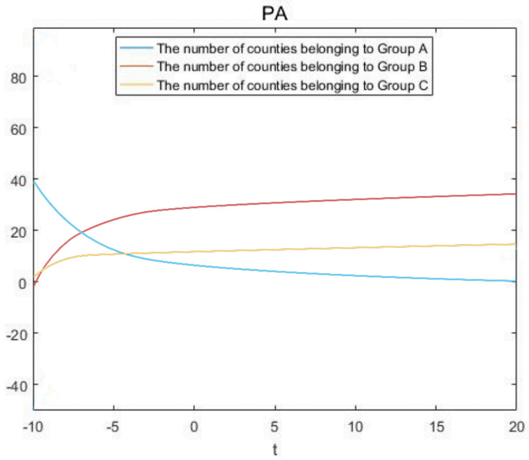


Figure 9: PA 2

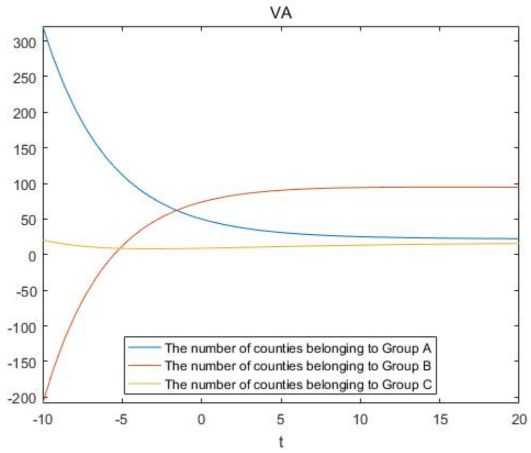


Figure 10: VA 2

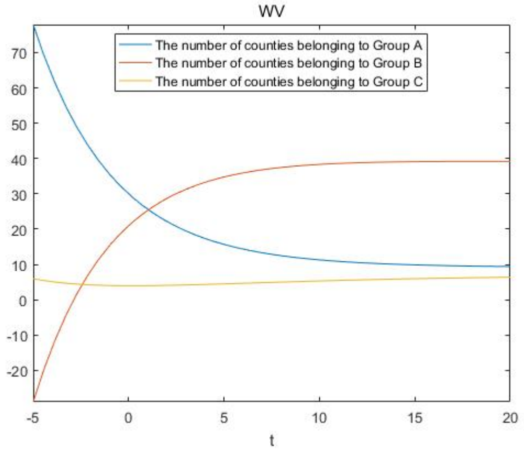


Figure 11: WV 2

Set 2010 at $t = 0$ the image shows that the timing of heroin identification cases in Kentucky Pennsylvania Virginia and West Virginia is about -2, -10, -5, -3, while for Ohio, at $t = -10$ the three groups of counties still have a relatively large number so it can be considered that Ohio first began to appear heroin cases. At the same time we can also see in the image that the number of counties in the three states of the five states tends to be stable over time. From the analysis in section 1.2.1 there is a threshold level for drug identification and even we can directly conclude that In the seventh year—that is, the year 2017— drug identification in Virginia reached a threshold level.

3 Selection of socio-economic factors

Since the given data is spatio-temporal data the data dimension is extremely large, we should reduce the dimensionality of the data and extract useful data.

3.1 Preliminary cleaning and screening variables (Take "ACS_10_5YR_DP02.zip" data as an example)

3.1.1 Eliminate duplicate social data indicators

In the socio-economic indicator data of "ACS_10_5YR_DP02_with_ann.csv" , there are 464*596 data (the number of counties* the number of social and economic indicators). According to the data given by the topic we know the abbreviations of each variable as follows.

Table 3: Meanings of abbreviations

Abbreviations	Meanings
HC_01/HC_02/HC_03/HC_04	Estimate/Margin of Error/Percent/Percent Margin of Error
VC_XX	Types of economic data

It can be seen that there is only a dimensional difference between HC_01 and HC_03 ,and HC_02 and HC_04 ,which are repeated social data indicators. Only two of them can be selected. The data gap entries in HC_03 and HC_04 are more than HC_01 and HC_02 , so we only considers HC_01 and HC_02 and the data volume drops to 464*298.

3.1.2 Processing missing indicators

According to "ACS_10_5YR_DP02.txt" , we know that the distribution and types of missing data in the socioeconomic indicator data are as follows.

Table 4: Add caption

Data Type	Possible Missing Data Type	Handle Way
HC_01	(X)	Delete
HC_02	(X) ********	Delete\Ignore\Ignore
HC_03	(X) \-	Delete\Replace with 0
HC_04	(X) **	Delete\Replace with 100

- (X) data It refers to an entirely missing column of data. So we directly eliminate this type of data.
- *****/*** data It refers to individual missing data. This error only occurs with HC_02. Since we do not use this data, it is ignored.
- \- data It refers to individual missing data. The reason for this is that the amount of data is small so we replace it directly with 0.
- ** data It refers to individual missing data. The reason for the occurrence is that the sample observation value or the sample observation value is too small to calculate the standard error. The data standard error is unknown and may be taken as 100 which is classified as a large error indicator.

3.1.3 Excluding the counties that did not appear in the data

In the previous Part A, the number of counties is 445. In “ACS_10_5YR_DP02_with_ann.csv” the number of counties is 464. Since we analyze the socio-economic indicator data below we will use the data of each county in each year as a sample to carry out LASSO regression to ensure the county sample. The dimensions are the same and we have removed the counties that did not appear in the first question data here.

3.1.4 Eliminate indicators with large errors

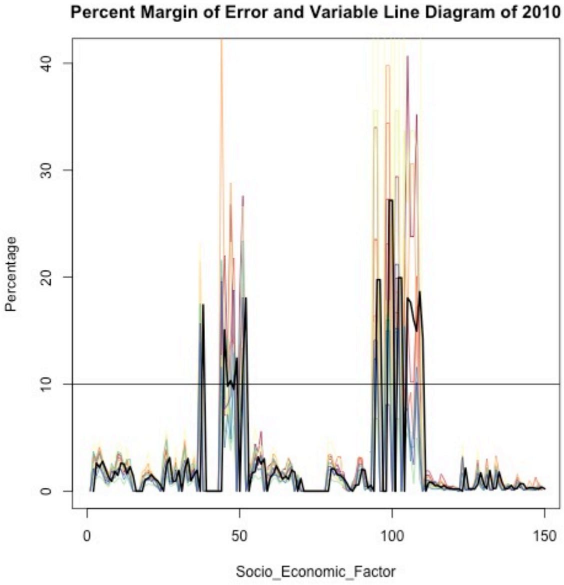


Figure 12: Percent Margin Error Data and Variable Graph 2010 <strong>[1]</strong>

Based on the percentage error data in the socioeconomic indicator data, we plot the upper graph for all 149 indicator variables. According to the image we can see that the error percentage of most social and economic indicators is below 10%. In order to ensure the accuracy of the subsequent screening variables we removed the indicator with a large percentage error (>10%) and the data volume decreased. The data volume drops to 464*298.

3.2 Dimension reduction and variables filtering by LASSO regression (Take “ACS_10_5YR_DP02.zip” data as an example)

After our initial screening of the data the total amount of data is still large, although the data dimension has declined. Moreover the socioeconomic indicator data is spatiotemporal data and there is likely to be serious multicollinearity between the data. Therefore in order to screen out important socio-economic indicators we need to further refine the data.

[1] Caption: The color of the same color line is the error percentage distribution line of each social and economic indicator of a single county. The black line is the error percentage distribution line of all the social and economic indicators of all counties. The horizontal line of 10% is the threshold line for large error indicators.

3.2.1 Multicollinearity test

First we test the multicollinearity of the data.

- **Calculate the rank of the matrix**
  Using R software we can calculate the rank of the socioeconomic indicator data matrix X and get R(X)=98. Since the dimension of the socioeconomic indicator data matrix X is 445*131 the socioeconomic indicator data matrix is not full rank, indicating that X_i can be represented by other linear combinations of X_j that is there may be collinearity.

- **Calculate condition number**
  Since matrix X may have collinearity but not complete linear correlation, we calculate the condition number kappa(X).

  - If k<100 indicates that the degree of collinearity is small;
  - If 100<k<1000 there is more multi-colinearity;
  - If k>1000 there is severe multicollinearity.

  We calculate the condition number of X as Inf » 1000 and consider that X has very serious multicollinearity. Therefore we need to build a model that meets the following two requirements.

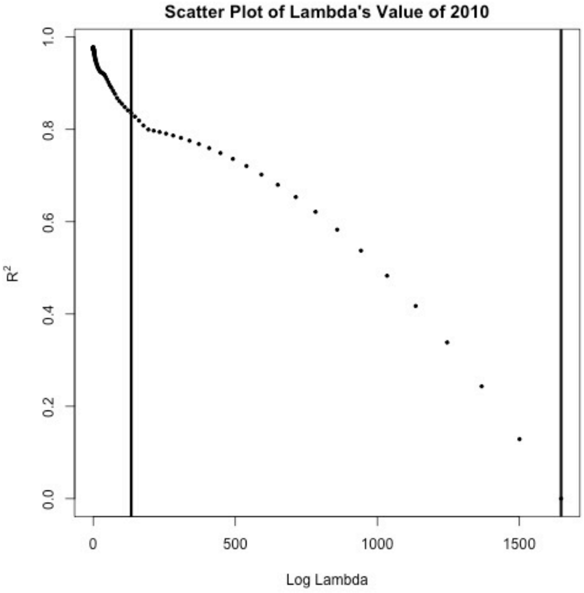
  1. Eliminate the multicollinearity of data
  2. Screen out the factors related to the high drug abuse rate in the socioeconomic indicator data.

3.2.2 Cross validation and Lasso Regression

We use the "glmnet" package in R to regress the data. The parameter λ normalized in the glmnet package is in $[0, 1]$ and the standardization formula is as follows:

$$\hat{\lambda} = -log(\lambda), \lambda \in [0, 1] \tag{18}$$

In order to estimate the parameters of the model using LASSO first we select 100 values of λ and build 100 LASSO regression models. Where the distribution of λ is as shown in the figure:



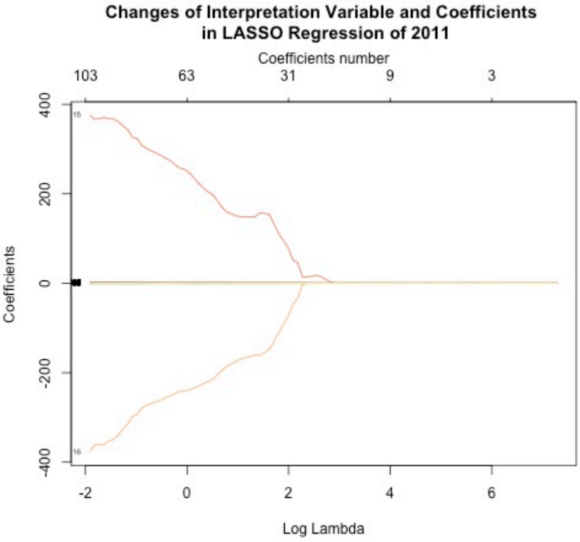


Figure 13: Selected Lambda Corresponding RegressnR2nGraph of 2010

Figure 14: Changes of Interpretation Variable and Coefficients in LASSO Regression of 2010

By taking the 100 sets of λ values we selected into the LASSO model we get 100 sets of regression equations. To understand our model intuitively we have drawn Figure 3.2. Each curve in the graph represents the trajectory of the coefficient of each independent variable. The ordinate is the value of the coefficient the lower abscissa is $\log(\lambda)$ and the upper abscissa is the number of non-zero coefficients in the model at this time. We can see that the non-zero coefficient at the beginning of λ is smaller as the value of λ becomes larger and eventually all coefficients converge to zero.

The selection of parameter λ for the LASSO regression model is important when selecting the LASSO parameter λ artificially:

1. If λ is chosen too large, all parameters w will be minimized resulting in under-fitting;

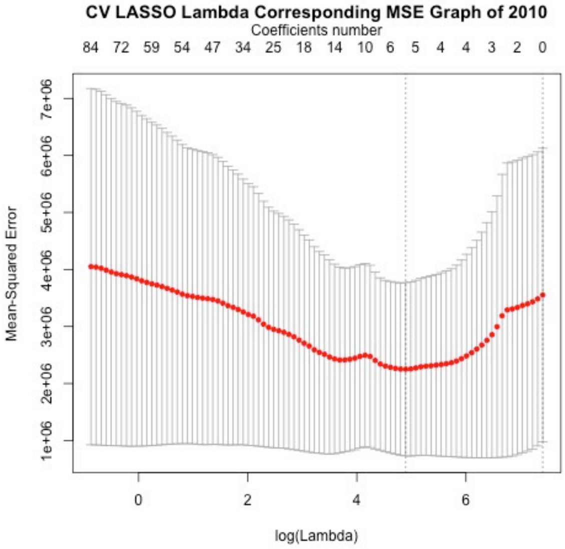
2. If λ is chosen too small, it will lead to improper resolution of the overfitting problem;

3. Artificially selecting the λ value is subjective and not easy to operate.

To solve these problems we cross validation to choose the optimal λ value.

3.2.3 Selecting λ by using 10-fold cross-validation

We use 10 cross-validation to perform cross-check on the 100 λ values in the previous section. The results are shown in the following figure.



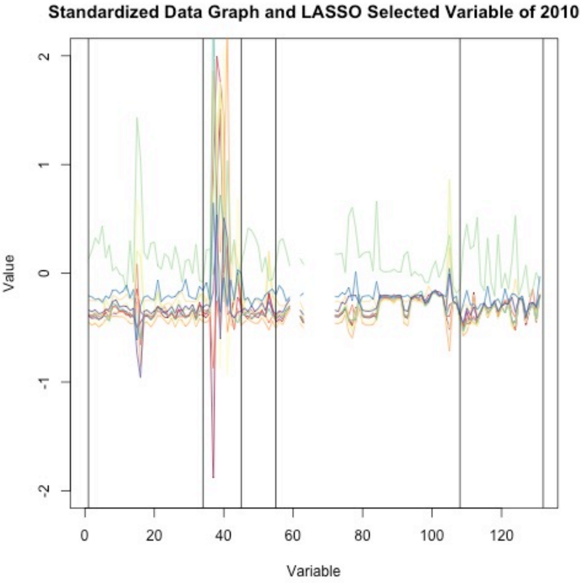


Figure 15: CV LASSO Lambda Corresponding MSE Graph of 2010

Figure 16: Standard Data LASSO Variable of 2010

In Figure 15, the red scatter is the scatter plot for the cross-check. The horizontal axis is logλ and the vertical axis is the mean square error. The upper and lower bounds of the standard deviation for each point are also drawn. The top word count of the graph represents the number of non-zero coefficients and the two vertical dashed lines are the selected λ after the cross-check. Among them a dotted line refers to the value of the interactive test that minimizes the mean square error (MSE) λ and the other dotted line is the λ value that is one standard deviation from the minimum mean square error.

3.2.4 Result analysis and prediction

We select the λ value with the smallest mean square error (MSE) from 100 λ by cross validation and then get the corresponding regression model. The λ value selected here selects 59 explanatory variables and the coefficient positions of each are shown in Figure 16. The coefficients of each explanatory variable are shown in the Appendix.

3.2.5 filtering variables

For the variables obtained from the application of LASSO regression in 2010-2016, we obtained all explanatory variables that occurred more than 3 times in 7 years as shown in the following table.

4 Modifying Drug Spread Model

4.1 Modified Drug Spread Model

In Drug Spread Model we constructed eight parameters to characterize the changes in each county of each state and assumed that they have no relationship with time and then the optimal parameters were obtained from the actual discrete data of 2010-2018. However, in fact, due to the different objective conditions such as national policy each year, these parameters may change with time. Therefore the model needs to be modified, which means correction functions $\varepsilon_i(t)$ are added based on the original parameter data (since$C_A + C_B = 1$, we only consider the modification of the 7 parameters) the modification of the Drug Spread Model is transformed into solving the specific form of the correction function $\varepsilon_i(t)$.

In past section we got a series of indicators that have a significant impact on the number of drug cases in the county. We selected two indicators that have a significant impact on the number of drug cases in the county each year: the number of men who did not have a junior high school diploma at the age of 25 (Estimate; EDUCATIONAL ATTAINMENT - High school graduate (includes equivalency) Estimate; EDUCATIONAL ATTAINMENT - 9th to 12th grade no diploma)and the number of unmarried women with children. Moreover it is available from section 2.1 that in a statistical point of view, the number of drug cases in each county is linear with the two indicators and it can be considered that the two indicators are independent of each other.

We denote $M_{A1}(t)$ ratio of the sum of the first indicator(the number of men who did not have a junior high school diploma at the age of 25) of all Group A counties to the sum of the first indicator of all counties at time t.

Similarly defined variable$M_{A2}(t)$, $M_{B1}(t)$, $M_{B2}(t)$, $M_{C1}(t)$, $M_{C2}(t)$.

First ,we convert the discrete variables into continuous variables: $M_{A1}(t)$, $M_{A2}(t)$, $M_{B1}(t)$, $M_{B2}(t)$, $M_{C1}(t)$, $M_{C2}(t)$ are discrete variables, We first use the known set of discrete values to do Newton interpolation on t separately to get the corresponding continuous function $M_{A1}(t)$, $M_{A2}(t)$, $M_{B1}(t)$, $M_{B2}(t)$, $M_{C1}(t)$, $M_{C2}(t)$.(All functions are listed in the Appendix.)

$$\begin{array}{l}
{\delta_B}' = 0.0365 + k_{11}M_{B1}(t) + k_{12}M_{B2}(t)\\
{\delta_C}' = 0.0277 + k_{21}M_{C1}(t) + k_{22}M_{C2}(t)\\
b' = 0.0031 + k_{31}M_{A1}(t) + k_{32}M_{A2}(t)\\
{C_1}' = 0.0042 + k_{41}M_{B1}(t) + k_{42}M_{B2}(t)\\
{C_{21}}' = 1.0566 \times 10^{-8} + k_{51}M_{A1}(t) + k_{52}M_{A2}(t)\\
{C_{22}}' = 1.1231 \times 10^{-9} + k_{61}M_{A1}(t) + k_{62}M_{A2}(t)\\
{C_B}' = 1 + k_{71}M_{C1}(t) + k_{72}M_{C2}(t)
\end{array}$$

4.2 Computing parameters

In past section , we use the numerical differentiation to replace the derivative value at that moment approximate equation (1)(2) to discrete form and then solve the optimization problem of inequality constraint to get the value of the parameter. In order to find the value of a set of parameters we can also do the above processing on the modified model YYYY. It is worth noting that the optimization problem with the sum of squared residuals as the objective function is unconstrained. Solving an unconstrained optimization problem yields a set of optimal parameters and the parameters are corrected to

$$\begin{aligned}
&{\delta_B}' = 0.0365 + 0.0209M_{B1}(t) - 0.0076M_{B2}(t)\\
&{\delta_C}' = 0.0277 + 1.1326M_{C1}(t) + 0.0967M_{C2}(t)\\
&b' = 0.0031 + 0.9846M_{A1}(t) + 1.2254M_{A2}(t)\\
&{C_1}' = 0.0042 - 0.0014M_{B1}(t) + 2.8972 \times 10^{-4} \times M_{B2}(t)\\
&{C_{21}}' = 1.0566 \times 10^{-8} + 2.0984 \times 10^{-4} \times M_{A1}(t) + 0.0045M_{A2}(t)\\
&{C_{22}}' = 1.1231 \times 10^{-9} + 3.1876 \times 10^{-4} \times M_{A1}(t) - 5.0832 \times 10^{-4} \times M_{A2}(t)\\
&{C_B}' = 1 - 0.0038M_{C1}(t) - 0.0142M_{C2}(t)
\end{aligned}$$

The Modified Drug Spread Model is:

$$\begin{aligned}
&T_{ABi}(t) = \sum_{j=1}^{5} g_{ij} N_{Bj}(t) N_{Ai}(t) b' + \sum_{j=1}^{5} g_{ij} N_{Cj}(t) N_{Ai}(t) {c_{21}}'\\
&T_{ACi}(t) = \sum_{j=1}^{5} g_{ij} N_{Cj}(t) N_{Ai}(t) {c_{22}}'\\
&T_{BCi}(t) = \sum_{j=1}^{5} g_{ij} N_{Cj}(t) N_{Bi}(t) {c_1}'\\
&\tfrac{dN_{Ai}(t)}{dt} = {\delta_B}' N_{Bi}(t) - T_{ABi}(t) - T_{ACi}(t) + {C_A}'{\delta_C}' N_{Ci}(t)\\
&\tfrac{dN_{Bi}(t)}{dt} = T_{ABi}(t) - {\delta_B}' N_{Bi}(t) - T_{BCi}(t) + {C_B}'{\delta_C}' N_{Ci}(t)\\
&\tfrac{dN_{Ci}(t)}{dt} = T_{BCi}(t) - {\delta_C}' N_{Ci}(t) + T_{ACi}(t)\\
&{C_A}' + {C_B}' = 1\\
&i = 1, 2, 3, 4, 5
\end{aligned}$$

5 Strategies to deal with the drug crisis

Based on our research on socioeconomic data and the number of drug cases we have obtained 13 factors that may be closely related to the number of drug cases. If you want to reduce the rate of drug abuse you can try to consider the development of relevant policies and laws and regulations from some aspects. On the other hand considering the two indicators involved in the revised model we have finally determined a strategy that is easy to implement increase the investment in education.

5.1 Use model to test the feasibility of strategies

Suppose the government invests a certain amount each year in 2010-2017 which reduces the number of people who have no junior high school diplomas over 25 years old in each county to 80% ,that is in Modified Drug Spread Model , $M_{A1}(t)$, $M_{A2}(t)$, $M_{B1}(t)$, $M_{B2}(t)$ reduce to 80% . Then according to section 2.2.1, the parameters of the model are corrected to

$${\delta_B}'', {\delta_C}'', b'', {C_1}'', {C_{21}}'', {C_{22}}'', {C_B}''$$

. Substituting into the system of ordinary differential equations, we find the function of the number of different counties in five states over time under the policy implementation and compared it with that before the implementation. It can be clearly found that after the implementation of the policy the number of counties in each of Group B and C in almost every integer time in each state has a different degree of decline compared to the implementation of the policy that is the total number of drug cases, compared to the decline before the implementation of the policy, which also illustrates the feasibility of our proposed policy.

$f(k)=\sum_{i=1}^{5}\sum_{t=0}^{7}N_{Bi}(t)'''+N_{Ci}(t)'''-N_{Bi}(t)-N_{Ci}(t)$ reflects the change before and after policy implantation.We only need to solve the following optimization problem:

$$\max f(k)=\sum_{i=1}^{5}\sum_{t=0}^{7}N_{Bi}(t)'''+N_{Ci}(t)'''-N_{Bi}(t)-N_{Ci}(t)$$
$$s.t.\begin{cases} 0<k<1 \\ f(k)<0 \end{cases}$$

And we get $k=0.854$. That is to say, in 2010-2017 if the government increased investment in education the number of people in each county who did not have a junior high school diploma over 25 years old would be reduced to the original k times. As long as $k\in(0,0.854)$, the policy is feasible.

6 Memo

MEMORANDUM TO: The Chief Administrator, DEA/NFLIS Database FROM: MCM Team DATE: January 29, 2019 SUBJECT: Significant Insights and Results from Modeling 在我们建模过程中我们发现了许多有趣的现象。We found many interesting phenomena in our modeling process.

1. Based on our statistical results we can see that the distribution of drug incidents in five states is mainly concentrated near the Great Lakes (Ohio) and the US East Coast (Pennsylvania). In particular, Ohio State Cuyahoga County, the fastest growing county, has more than 5,000 Total Drug Reports in six years, and is the fastest growing County in all counties. Total Drug Reports increased 4.74 times between 2011 and 2012. As shown in the following figure.

Table 5: Add caption

Value	Count
Pennsylvania State PHILADELPHIA County	8
Ohio State HAMILTON County	8
Pennsylvania State ALLEGHENY County	8
Ohio State MONTGOMERY County	8
Pennsylvania State BUCKS County	3
Kentucky State JEFFERSON County	2
Ohio State CUYAHOGA County	6
Ohio State FRANKLIN County	6
Ohio State LAKE County	2

2. According to our Drug Spread Model, we divided the drug abuse situation in each county into basic non-existence (A) existence (B) and serious existence (C) level 3 and simulated the states according to mathematical models. The law of time and space changes in the number of cases in all counties. We finally the following four conclusions

   - a) Grade A counties have a probability of 0.0277 and are converted to Grade B counties and it is almost impossible to directly convert to Grade A counties; Grade B counties have a 0.0365 chance to be converted to Grade A counties.

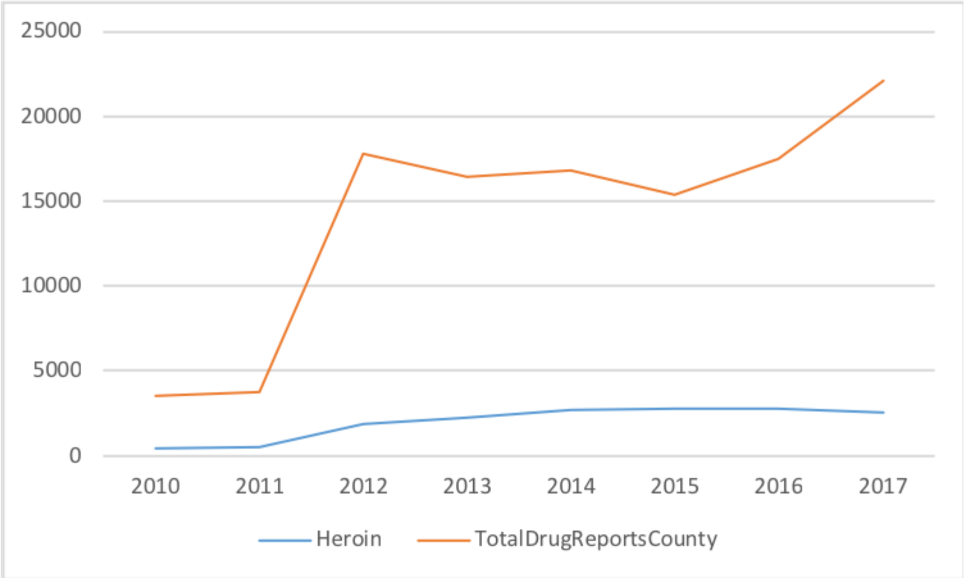


Figure 17: Reports Graph of Ohio State Cuyahoga County

- b) If no policy intervention is added, some heavily drug-using counties in Ohio and Pennsylvania will be converted to moderate drug-using counties in the future and some moderate drug-drug counties in Kentucky, Virginia and West Virginia will be converted to severe drug-using counties. Therefore, the government should focus on drug cases in Kentucky Virginia and West Virginia.
- c) The number of counties in the three states of the five states tends to be stable over time, while Virginia is the first to reach a steady state, that is in the 60th year that is the drug identification in Virginia reached the threshold level in 2070.
- d) It is worth mentioning that we have analyzed the number of heroin cases and found that the first case of heroin in Ohio began to occur. At the same time in the seventh year 2017 the heroin drug identification in Virginia reached the threshold level.

3. Based on our research on socioeconomic data and the number of drug cases we have obtained 13 factors that may be closely related to the number of drug cases (see table below). If you want to reduce the rate of drug abuse you can try to consider the development of relevant policies and laws and regulations from some aspects.

References

[1] Gabriel Svennerberg. *Beginning Google Maps API 3.* Apress, 2010.

Appendices

Appendix A Percent Margin Error Data and Variable Graph

After the analysis of 2010, we have done the same analysis to the data from 2011 to 2016, and draw the following pictures.

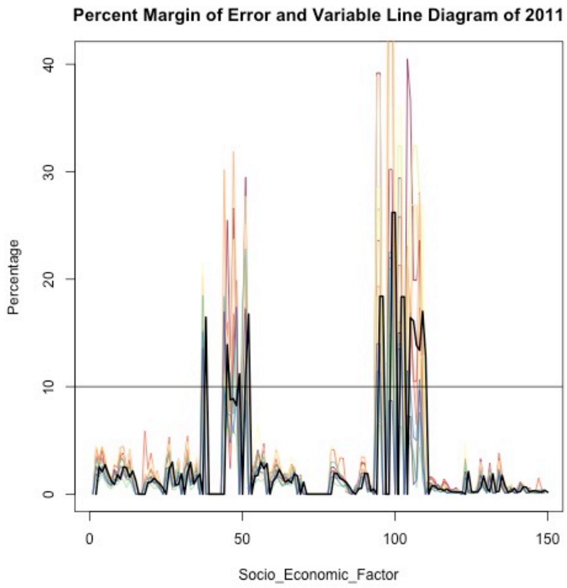


Figure 18: Percent Margin Error Data and Variable Graph 2011

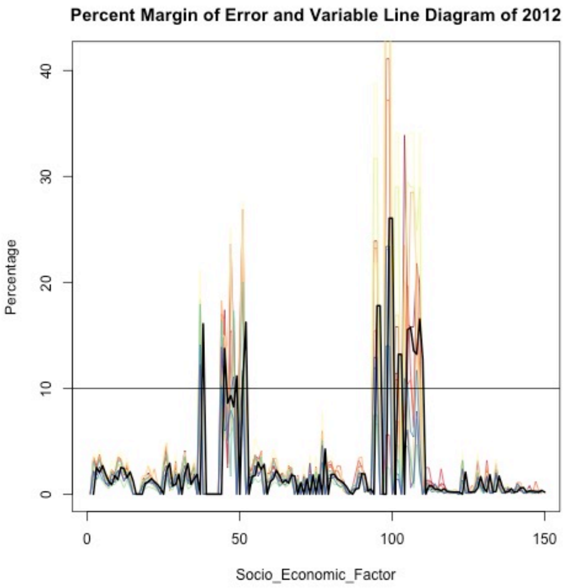
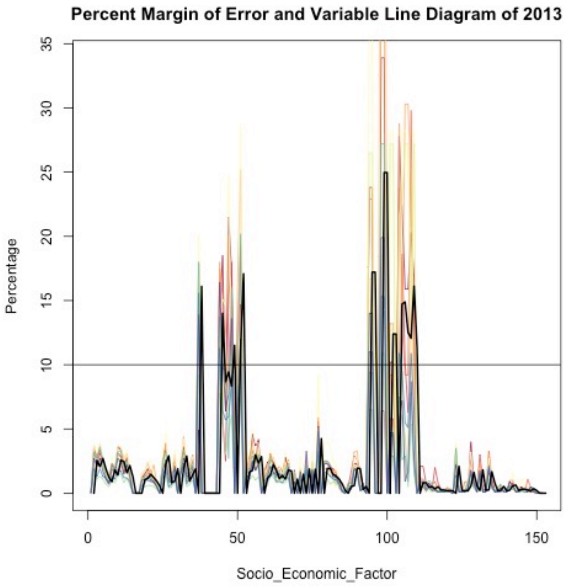


Figure 19: Percent Margin Error Data and Variable Graph 2012



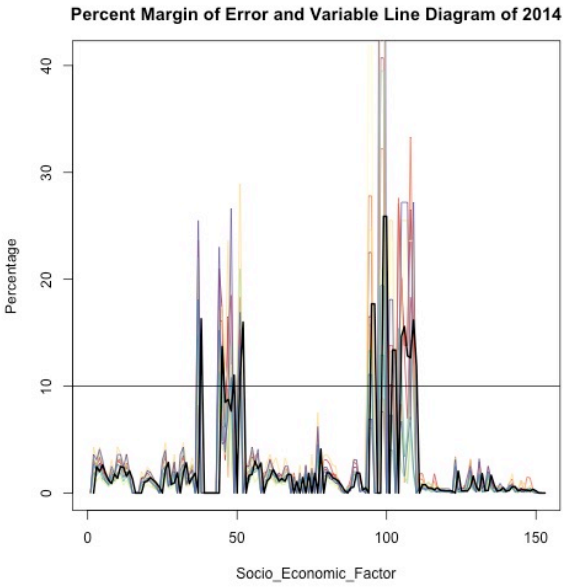
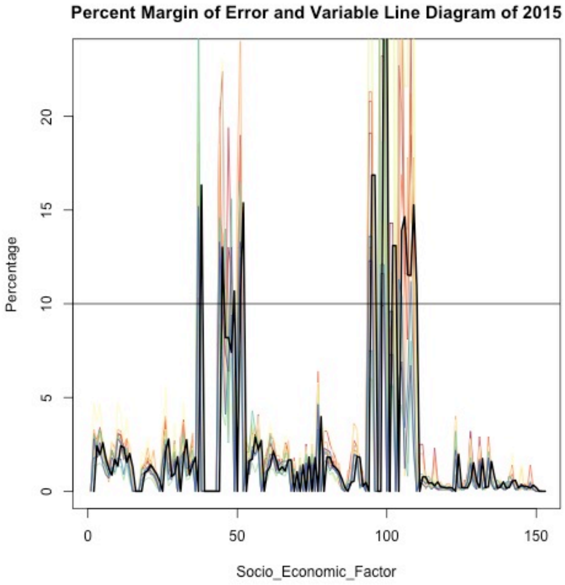


Figure 20: Percent Margin Error Data and Variable Graph 2013

Figure 21: Percent Margin Error Data and Variable Graph 2014



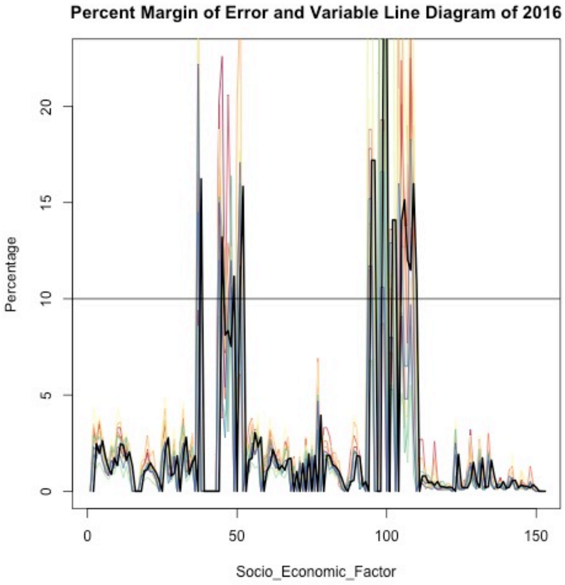


Figure 22: Percent Margin Error Data and Variable Graph 2015

Figure 23: Percent Margin Error Data and Variable Graph 2016

Appendix B Changes of Interpretation Variable and Coefficients Graph in LASSO Regression

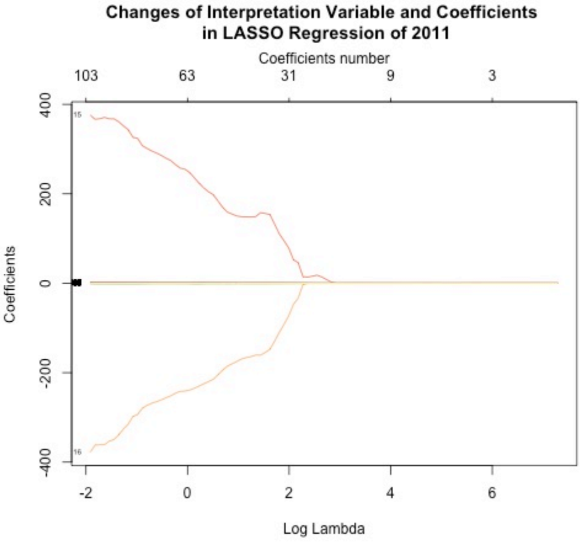


Figure 24: Changes of Interpretation Variable and Coefficients Graph in LASSO Regression of 2011

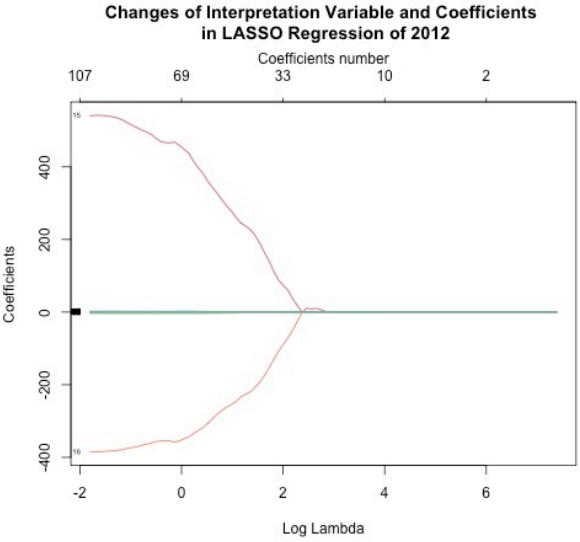


Figure 25: Changes of Interpretation Variable and Coefficients Graph in LASSO Regression of 2012

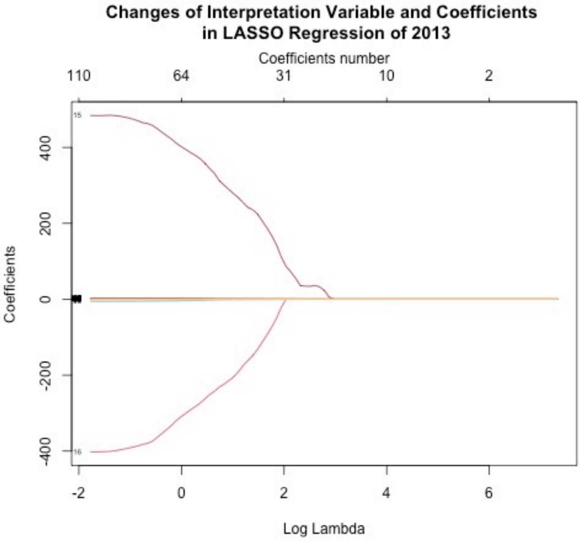


Figure 26: Changes of Interpretation Variable and Coefficients Graph in LASSO Regression of 2013

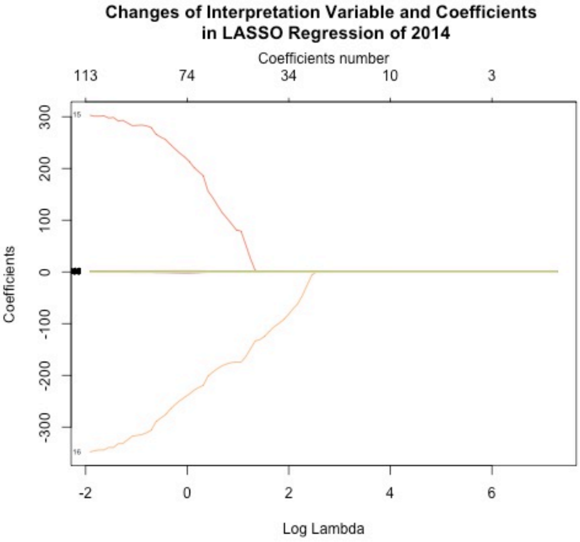


Figure 27: Changes of Interpretation Variable and Coefficients Graph in LASSO Regression of 2014

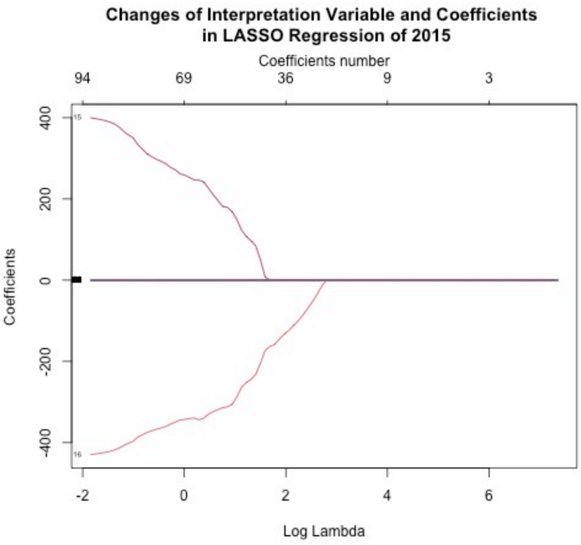


Figure 28: Changes of Interpretation Variable and Coefficients Graph in LASSO Regression of 2015

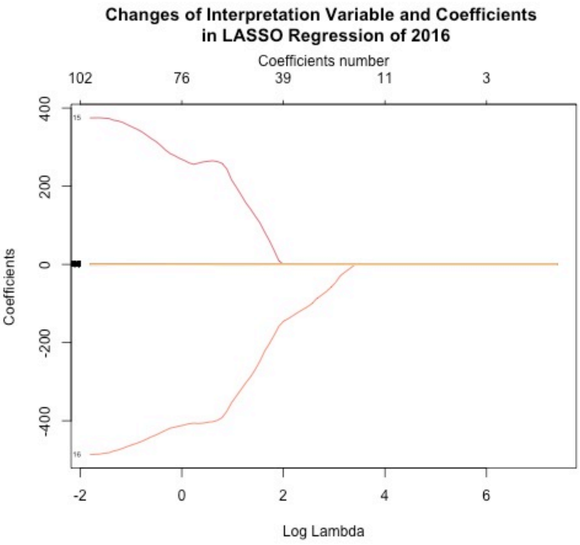
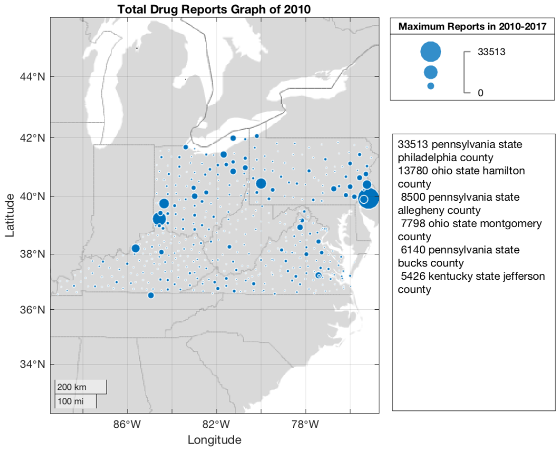


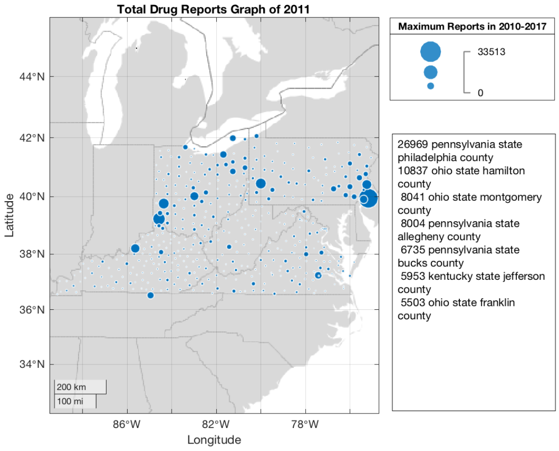
Figure 29: Changes of Interpretation Variable and Coefficients Graph in LASSO Regression of 2016

Appendix C Total Drug Reports County



33513 pennsylvania state philadelphia county
13780 ohio state hamilton county
8500 pennsylvania state allegheny county
7798 ohio state montgomery county
6140 pennsylvania state bucks county
5426 kentucky state jefferson county

Figure 30: Total Drug Reports County in 2010



26969 pennsylvania state philadelphia county
10837 ohio state hamilton county
8041 ohio state montgomery county
8004 pennsylvania state allegheny county
6735 pennsylvania state bucks county
5953 kentucky state jefferson county
5503 ohio state franklin county

Figure 31: Total Drug Reports County in 2011

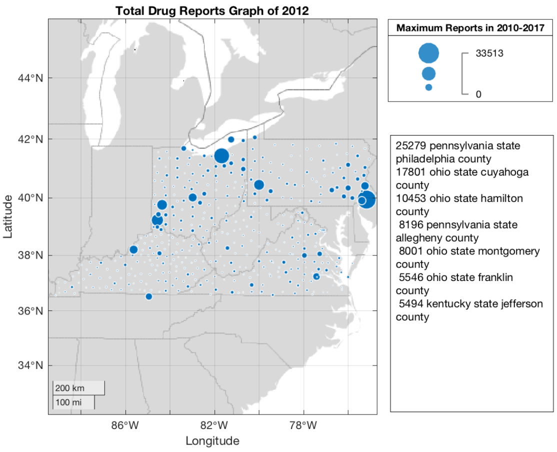

Total Drug Reports Graph of 2012
Maximum Reports in 2010-2017
33513
0
25279 pennsylvania state philadelphia county
17801 ohio state cuyahoga county
10453 ohio state hamilton county
8196 pennsylvania state allegheny county
8001 ohio state montgomery county
5546 ohio state franklin county
5494 kentucky state jefferson county
44°N
42°N
40°N
38°N
36°N
34°N
Latitude
200 km
100 mi
86°W
82°W
78°W
Longitude

Figure 32: Total Drug Reports County in 2012

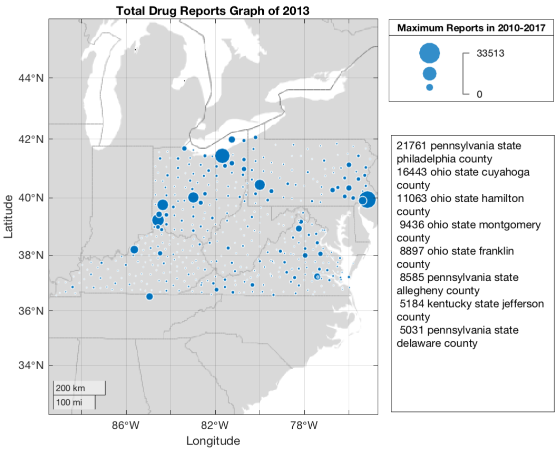

Total Drug Reports Graph of 2013
Maximum Reports in 2010-2017
33513
0
21761 pennsylvania state philadelphia county
16443 ohio state cuyahoga county
11063 ohio state hamilton county
9436 ohio state montgomery county
8897 ohio state franklin county
8585 pennsylvania state allegheny county
5184 kentucky state jefferson county
5031 pennsylvania state delaware county
44°N
42°N
40°N
38°N
36°N
34°N
Latitude
200 km
100 mi
86°W
82°W
78°W
Longitude

Figure 33: Total Drug Reports County in 2013

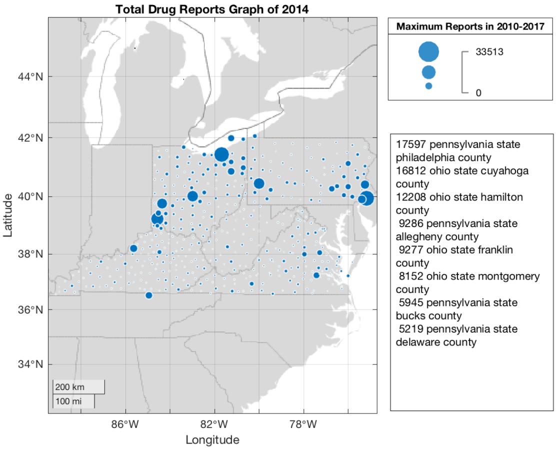

Total Drug Reports Graph of 2014
Maximum Reports in 2010-2017
33513
0
17597 pennsylvania state philadelphia county
16812 ohio state cuyahoga county
12208 ohio state hamilton county
9286 pennsylvania state allegheny county
9277 ohio state franklin county
8152 ohio state montgomery county
5945 pennsylvania state bucks county
5219 pennsylvania state delaware county
44°N
42°N
40°N
38°N
36°N
34°N
Latitude
200 km
100 mi
86°W
82°W
78°W
Longitude

Figure 34: Total Drug Reports County in 2014

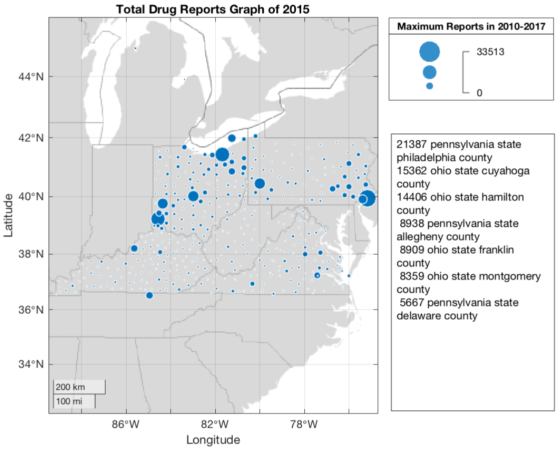

Total Drug Reports Graph of 2015
Maximum Reports in 2010-2017
33513
0
21387 pennsylvania state philadelphia county
15362 ohio state cuyahoga county
14406 ohio state hamilton county
8938 pennsylvania state allegheny county
8909 ohio state franklin county
8359 ohio state montgomery county
5667 pennsylvania state delaware county
44°N
42°N
40°N
38°N
36°N
34°N
Latitude
200 km
100 mi
86°W
82°W
78°W
Longitude

Figure 35: Total Drug Reports County in 2015

Appendix D Matlab Code

Here are simulation programmes we used in our model as follow.

Input matlab source:

```
%      state=convertCharsToStrings(MCMNFLISDataS1.State);
% import data
[~, ~, MCMNFLISDataS1] = xlsread('/Users/rexzhang/Downloads/2019_MCM-ICM_Problems/2018_MCMProblemC_DATA/MCM_NFLIS_Data.xlsx','Data');
MCMNFLISDataS1 = MCMNFLISDataS1(2:end,:);

MCMNFLISDataS1 = string(MCMNFLISDataS1);
MCMNFLISDataS1(ismissing(MCMNFLISDataS1)) = '';
% search for location
FIPS_Combined=unique(MCMNFLISDataS1(:,6));

```

```
county=MCMNFLISDataS1(:,3);
county=strcat(county, ” County”);

state=MCMNFLISDataS1(:,2);
state=strrep(state,'OH','Ohio State');
state=strrep(state,'KY','Kentucky State');
state=strrep(state,'WV','WestVirginia State');
state=strrep(state,'PA','Pennsylvania State');
state=strrep(state,'VA','Virginia State');


StateCountyName=unique(strcat(state,” ”, county));
%连接Google Map API获取地点的经纬度数据
% for i=1:length(StateCountyName)
%%          for i=[209:274]
%%     i,StateCountyName(i)
%      [lng(i),lat(i),value(i)] = geoCode(StateCountyName(i),'AIzaSyDn06hjcJEFz2HRfPI463e7oVIY5ZKGPss');
% end
%CombinedGeoData = [StateCountyName FIPS_Combined lng' lat'] ;

%已经获取数据，直接导入
% Import the data
[~, ~, raw] = xlsread('/Users/rexzhang/Desktop/CombinedGeoData.xlsx','Sheet1');
stringVectors = string(raw(:,1));
stringVectors(ismissing(stringVectors)) = '';
raw = raw(:,[2,3,4]);
data = reshape([raw{:}],size(raw));
CombinedGeoData = table;
CombinedGeoData.StateCountyName = stringVectors(:,1);
CombinedGeoData.FIPS_Combined = data(:,1);
CombinedGeoData.lng = data(:,2);
CombinedGeoData.lat = data(:,3);
clearvars data raw stringVectors;


% array2table
MCMNFLISDataS1 =array2table(MCMNFLISDataS1,'VariableNames',{'YYYY','State','    COUNTY','    FIPS_State','    r','     FIPS_Combined','SubstanceName','    DrugReports  ','TotalDrugReportsCounty','TotalDrugReportsState'});
CombinedFinalData=outerjoin(MCMNFLISDataS1,CombinedGeoData);
% plot and record

% plot(str2double(CombinedFinalData.lng),str2double(CombinedFinalData.lat),'.r', 'MarkerSize', 20);
% h = scatter3(str2double(CombinedFinalData.lng),str2double(CombinedFinalData.lat),str2double(CombinedFinalData.TotalDrugReportsCounty));
% plot_google_map('MapScale', 1,'MapType','roadmap')

vedio = VideoWriter('test.avi'); %初始化一个avi文件
vedio.FrameRate = 1;
open(vedio);
maxTotalDrugReportsCounty=max(str2double(CombinedFinalData.TotalDrugReportsCounty));
BadCountyData=cell(8,1);
for j=2010:2017
    a=CombinedFinalData(CombinedFinalData.YYYY==num2str(j),{'YYYY','StateCountyName','TotalDrugReportsCounty','TotalDrugReportsState','lat','lng'});
    a.TotalDrugReportsCounty=str2double(a.TotalDrugReportsCounty);
    a.lat=a.lat;
    a.lng=a.lng;
    BadCountyData{j-2009}=sortrows(unique(a(a.TotalDrugReportsCounty>5000,:)),3,'descend') ;


```

```
%         figure
%         geobubble(figure,a.lat,a.lng,a.TotalDrugReportsCounty,'BubbleWidthRange',[1 20],'MapLayout','normal',...
%             'SizeLimits',[0 maxTotalDrugReportsCounty],'Basemap','grayland',...
%             'SizeLegendTitle','Maximum Reports in 2010-2017','Title',sprintf('%s%s','Total Drug Reports Graph of ',num2str(j))...
%             );
%    %     text(figure,BadCountyData{1, 1}.lat,BadCountyData{1, 1}.lng,BadCountyData{1, 1}.StateCountyName)
%         annotation(figure,'textbox',[0.798118847539015 0.598993288590604 0.177871548619448 0.216442953020134],'String',{'geobubble'},'FitBoxToText','on');
%    set(0,'DefaultFigureVisible', 'off')
    figure=createfigure(a.lat, a.lng, a.TotalDrugReportsCounty, BadCountyData{j-2009},j);

    saveas(figure,sprintf('%s%s%s','TotalDrugReportsCounty',num2str(j),'.png'));


    fname=strcat('TotalDrugReportsCounty',num2str(j),'.png');
    frame = imread(fname);
    writeVideo(vedio,frame);
end
close(vedio);

%统计2010-2017大于5000次毒品案的县的频数
l=1;

for i=1:8
lagthan5000name(l:l+length(BadCountyData{i,1}.StateCountyName)-1,1)=BadCountyData{i,1}.StateCountyName ;
l=length(lagthan5000name);
end
tabulate(lagthan5000name);
```

```
function [lng,lat,value] = geoCode(address, key)

address=char(address);
if isempty(address) || ~ischar(address) || ~isvector(address)
    error('Invalid address provided, must be a string');
end

address = regexprep(address, ' ', '+');

SERVER_URL = 'https://maps.googleapis.com/maps/api/geocode/json?address=';
queryUrl = sprintf('%s%s',SERVER_URL, address);
if ~isempty(key)
    queryUrl = sprintf('%s&key=%s', queryUrl, key);
end
queryUrlContent=urlread(queryUrl);
value = jsondecode(queryUrlContent);
if length(value.results)<2 && ~isempty(value.results)
    lat=value.results.geometry.location.lat;
    lng=value.results.geometry.location.lng;
elseif isempty(value.results)
    lat=0;
    lng=0;
    warning('value.results is empty');
else
    lat=value.results{1, 1}.geometry.location.lat;
    lng=value.results{1, 1}.geometry.location.lng;
end
%         parseFcn = @parseGoogleMapsXML;


```

```
end
```

```
function [value] = geoDist(startaddress,endaddress)

startaddress=char(startaddress);
endaddress=char(endaddress);

startaddress = regexprep(startaddress, ' ', '+');
endaddress = regexprep(endaddress, ' ', '+');
SERVER_URL = 'https://maps.googleapis.com/maps/api/distancematrix/json?origins=';
queryUrl = sprintf('%s%s&destinations=%s',SERVER_URL, startaddress, endaddress);
queryUrl = sprintf('%s&key=%s', queryUrl, 'AIzaSyDn06hjcJEFz2HRfPI463e7oVIY5ZKGPss');

queryUrlContent=urlread(queryUrl);
value = jsondecode(queryUrlContent);
end
```

```
function [figure1]=createfigure(lat1, lon1, sizedata1,badcountydata,j)

% Create figure
figure1 = figure;

% Create geobubble
geobubble(figure1,lat1,lon1,sizedata1,...
    'Title',sprintf('%s%s','Total Drug Reports Graph of ',num2str(j)),...
    'Basemap','grayland',...
    'SizeLimits',[0 33513],...
    'BubbleWidthRange',[1 20],...
    'SizeLegendTitle','Maximum Reports in 2010-2017');
geolimits('auto');

% Create textbox
annotation(figure1,'textbox',...
    [.7 .73 .265 .38],...
    'String',{'Top Reports County'},...
    'FitBoxToText','on');

annotation(figure1,'textbox',...
    [.7 .1 .265 .6],...
    'String',{strjoin(strcat(num2str(badcountydata.TotalDrugReportsCounty),...
    {' '},lower(badcountydata.StateCountyName)),'\n')},...
    'FitBoxToText','off');
```

Appendix E R Code

Here are simulation programmes we used in our model as follow.

Input R source:

```
library(RColorBrewer) ##载入RColorBrewer程序包
library(readxl)
library(readr)
library(ggplot2)
library(dplyr)
library(glmnet)
library (plyr)

```

```
rm(list = ls())

display.brewer.all() ;##显示所有调色板以供选择
cols <- brewer.pal(n=11,name="Spectral") ;##设定颜色为set3调色板，n根据调色 板相应的改变
cols_x<-cols[1:445]; ##设定颜色矩阵


dev.new


allmainfactor=c();

for (z in 10:16)
{
MCM_NFLIS_Data <- read_excel("/Users/rexzhang/Downloads/2019_MCM-ICM_Problems/2018_MCMProblemC_DATA/MCM_NFLIS_Data.xlsx",
                                   sheet = "Data")
Book <- read_excel("/Users/rexzhang/Downloads/Book2.xlsx")
ACS_10_5YR_DP02_with_ann <- read_csv(paste("/Users/rexzhang/Downloads/2019_MCM-ICM_Problems/2018_MCMProblemC_DATA/ACS_",z,"_5YR_DP02/ACS_",z,"_5YR_DP02_with_ann.csv",sep=""))
ACS_10_5YR_DP02_with_ann <- ACS_10_5YR_DP02_with_ann[-1,]
CS_10_5YR_DP02_metadata <- read_csv(paste("/Users/rexzhang/Downloads/2019_MCM-ICM_Problems/2018_MCMProblemC_DATA/ACS_",z,"_5YR_DP02/ACS_",z,"_5YR_DP02_metadata.csv",sep=""))


#连接某一年数据
Book=Book[which(Book$YYYY == paste("20",z,sep="")),]

#除去了多余的县， all=F
CombinedData=merge(x=Book,y=ACS_10_5YR_DP02_with_ann,by.x="FIPS_Combined",by.y = "GEO.id2",all=F);

#处理空值
CombinedData[CombinedData=="(X)"]<-0
CombinedData[CombinedData=="*****"]<-NA
CombinedData[CombinedData=="**"]<-100


#提取误差率HC04<10%的变量名HC01goodname
HC04columIndices=seq(from=14,to=length(CombinedData),by=4)
HC04=CombinedData[,HC04columIndices];

Socio_Economic_Factor=1:dim(HC04)[2];
Percentage=HC04[1,]
HC04=as.matrix(sapply(HC04,as.numeric))

setwd("/Users/rexzhang/Downloads/mcmthesis-V6/figures")
jpeg(file=paste("PercentMarginErrorDataGraphVariable20",z,".jpeg",sep=""),width=500,height=500,quality = 75)

par(mar=c(5,5,3,2)+0.1, oma=c(0, 0, 0, 0))
plot(Socio_Economic_Factor,Percentage,'l',col=cols_x[1],main=paste("Percent Margin of Error and Variable Line Diagram of 20",z,sep=""),cex=0.5)
for (i in 2:dim(CombinedData)[1])
{Percentage=HC04[i,]
j=i
lines(Socio_Economic_Factor,Percentage,type='l',col=cols_x[j])
}
#误差率平均线
lines(Socio_Economic_Factor+1,colMeans(HC04),type='l','black',lwd=2)
#10%误差率截断线
abline(h =10, col ="black",lwd=1)

dev.off()

```

```
HC04goodname=names(subset(colMeans(HC04),colMeans(HC04)<10))
HC01goodname=gsub("HC04", "HC01", HC04goodname)





#采用总数而非比率HC_01
#HC03_VC93  Percent; EDUCATIONAL ATTAINMENT - Percent high school graduate or higher 仅有比率
columIndices=seq(from=11,to=length(CombinedData),by=4)


x=CombinedData[,columIndices];
#除去x中的误差率HC04>10%的指标
x=subset(x,select=HC01goodname)
x=as.matrix(sapply(x,as.numeric))
#x=matrix(x,nrow=437,ncol=152,byrow=FALSE)
y=as.numeric(CombinedData[,7]);



#lasso
#,lambda.min.ratio=0.001 , nlambda=100
fit=glmnet(x, y, family="gaussian", alpha=1)
#print(fit)

setwd("/Users/rexzhang/Downloads/mcmthesis-V6/figures")
jpeg(file=paste("ChangesofInterpretationVariableandCoefficientsinLASSORegressionof20",z,".jpeg",sep=""),width=500,height=500,quality = 75)

par(mar=c(5,4,9,2)+0.1, oma=c(0, 0, 0, 0))
plot(fit, xvar="lambda",label=TRUE,col=cols,main=paste("Changes of Interpretation Variable and Coefficients \n in LASSO Regression of 20",z,sep=""));
mtext("Coefficients number", side=3,line=2);
###lambda变化趋势
n=1:length(fit$lambda)
#qplot(x=n,fit$dev.ratio,ylab =expression(R^2))

#par(mar=c(5,4,2,2)+0.1, oma=c(0, 0, 0, 0))
#plot(x=n,y=fit$lambda,ylab ='Log Lambda',cex=0.7,pch=20)

dev.off()


coef=coef(fit, s=c(fit$lambda[length(fit$lambda)*0.2]))
coef@Dimnames[[1]][coef@i+1]

#交互检验 lasso ,lambda.min.ratio=0.00005
cvfit = cv.glmnet(x, y, family = "gaussian",alpha=1)


setwd("/Users/rexzhang/Downloads/mcmthesis-V6/figures")
jpeg(file=paste("CVLASSOLambdaCorrespondingMSEGraphof20",z,".jpeg",sep=""),width=500,height=500,quality = 75)

par(mar=c(5,5,7,2)+0.1, oma=c(0, 0, 0, 0))
plot(cvfit,main=paste("CV LASSO Lambda Corresponding MSE Graph of 20",z,sep=""))
mtext("Coefficients number", side=3,line=2)
dev.off()

###lambda变化趋势

jpeg(file=paste("SelectedLambdaCorrespondingRegressR2Graphof20",z,".jpeg",sep=""),width=500,height=500,quality = 75)
par(mar=c(5,4,2,2)+0.1, oma=c(0, 0, 0, 0))
```

```
m=1:length(cvfit$lambda)
plot(cvfit[["glmnet.fit"]][["lambda"]],cvfit[["glmnet.fit"]][["dev.ratio"]],xlab ='Log Lambda',ylab =expression(R^2),cex=0.7,pch=20,main=paste("Scatter Plot of Lambda's Value of 20",z,sep=""))
abline(v =c(cvfit$lambda.min, cvfit$lambda.1se), col ="black",lwd=3)
dev.off()

c(cvfit$lambda.min, cvfit$lambda.1se)
#coef_cvfit=coef(fit,s=c(cvfit$lambda.min, cvfit$lambda.1se))
coef_cvfit=coef(fit,s=c(cvfit$lambda.min))
abbr=data.frame(coef_cvfit@Dimnames[[1]][coef_cvfit@i+1],coef_cvfit@x)
names(abbr)[1]="abbr"
mainfactor=merge(x=abbr,y=CS_10_5YR_DP02_metadata,by.y= "GEO.id",by.x = names(abbr)[1],all=F)
mainfactor$year=2000+z
mainfactor$CorrespondingR2=cvfit[["glmnet.fit"]][["dev.ratio"]][which(cvfit[["glmnet.fit"]][["lambda"]] == cvfit$lambda.min)]

#setwd("/Users/rexzhang/Downloads/mcmthesis-V6/figures")
#yourfilenamecsv=paste("mainfactor20",z,".csv",sep="")#循环命名
#write.csv(mainfactor, file =yourfilenamecsv)



allmainfactor$abbr=c(allmainfactor$abbr,as.character(mainfactor$abbr));
allmainfactor$Id=c(allmainfactor$Id,mainfactor$Id);
allmainfactor$year=c(allmainfactor$year,mainfactor$year);
allmainfactor$coef_cvfit.x=c(allmainfactor$coef_cvfit.x,mainfactor$coef_cvfit.x);
allmainfactor$CorrespondingR2=c(allmainfactor$CorrespondingR2,mainfactor$CorrespondingR2);


#标准化x
#处理空值
CombinedData[CombinedData=="(X)"]<-0
nx<- matrix(nrow=dim(x)[1],ncol=dim(x)[2])
colnames(nx)=colnames(x)
for (i in 1:dim(x)[2]) {
  nx[,i]=scale(x[,i],center=T,scale=T);
}



#绘制标准化数据
Variable=1:dim(x)[2];
Value=nx[1,]

setwd("/Users/rexzhang/Downloads/mcmthesis-V6/figures")
jpeg(file=paste("StandardDataLASSOVariableof20",z,".jpeg",sep=""),width=500,height=500,quality = 75)


par(mar=c(5,5,3,2)+0.1, oma=c(0, 0, 0, 0))
plot(Variable,Value,'l',col=cols_x[1],main=paste("Standardized Data Graph and LASSO Selected Variable of 20",z,sep=""),ylim=c(-2,2),cex=0.5)
for (i in 2:dim(CombinedData)[1])
{Value=nx[i,]
j=i
lines(Variable,Value,type='l',col=cols_x[j])
}
#绘制选择的变量位置
abline(v =coef_cvfit@i+1, col ="black",lwd=1)

dev.off()
}
dataframeallmainfactor=data.frame(t(rbind(allmainfactor$Id,allmainfactor$abbr,allmainfactor$year,allmainfactor$coef_cvfit.x,allmainfactor$CorrespondingR2)))
```

```
freqx1=data.frame(table(dataframeallmainfactor))
#publicmainfactor=merge(x=freqx1 ,y=CS_10_5YR_DP02_metadata ,by.x=names(freqx1)[1],by.y=names(CS_10_5YR_DP02_metadata)[2], all.x=TRUE)
write.csv(dataframeallmainfactor, file =paste(”Publicmainfactor.csv”,sep=””))

avd=merge(x=data.frame(table(dataframeallmainfactor$X2)),y=data.frame(CS_10_5YR_DP02_metadata),by.y= names(data.frame(CS_10_5YR_DP02_metadata))[1],by.x = names(data.frame(table(dataframeallmainfactor$X2)))[1], all.x=T)
write.csv(avd, file =paste(”PublicmainfactorCount.csv”,sep=””))

#publicmainfactor=data.frame(table(dataframeallmainfactor$X1, dataframeallmainfactor$X2))
```