For office use only	Team Control Number 1920129	For office use only
T1 ________________		F1 ________________
T2 ________________		F2 ________________
T3 ________________	Problem Chosen	F3 ________________
T4 ________________	C	F4 ________________

2019

MCM/ICM

Summary Sheet

Summary

Currently, USA experience a explosive crisis of synthetic and non-synthetic opioids. How to control the opioids crisis has become a nation concern ever since. Based on the data, we has made a comprehensive study on this problem.

Firstly, we preprocess the data provided, which includes data visualization, Correlation analysis, default value processing, abnormal value process.

Next, based on SIS epidemic model, considering the transmission between the relevant counties and the transmission within the counties, we create an improved model : Transmission-SIS model. That is, we weight the basic transmission rate to obtain the correlation transmission rate. We use the least squares estimation with constraints to estimate parameters, the model testing shows that the result is good. According to our model, we obtain the origins of each drug and predict the drug use in each county of 5 states from 2017 to 2035. We find that drug use counts in each county will eventually converge to the equilibrium value , which can be used as drug identification threshold level. We also describe other characteristics of drug spread, and show several specific concerns that the U.S. government should have.

Next, based on the socio-economic data, we use GBDT to generate a feature combination after weighting the top 4 important factors. Combining with the Transmission-SIS model and the feature combination, we Establish Social-Economic-Relevant transmission-SIS model. Accordingly, we bring up 2 strategies for countering the opioid crisis, which is helping addicts and increase financial investment in local education. The testing of parameter shows the efficiency of strategies.

Finally, we analysis sensitivity of the model, our system highly relied on transmission between regions and the county's drug use also auto related with itself significantly.

Contents

1 Introduction

1.1 Background and statement of the problem

Opioids play a significant role in the treatment and management of pain for prescription use. However, overuse of opioids causes negative effects - It not only hurts peoples' health, but also erodes social economy in many ways, therefore, the USA is experiencing a huge national crisis. Some institutions released an annual report to estimate state and local drug cases. Based on these accessible data, we analyze counties located in five U.S. states: Ohio, Kentucky, West Virginia, Virginia and Pennsylvania, we will use mathematical modeling methods to determine possible strategies to combat the opioid crisis, including its effectiveness and influencing factors.

2 Assumptions and Nomenclature

2.1 Assumptions

- Drug identifications' number always in proportion to the numbers of users.

- The population of each county remains unchanged with time.

- Every individual is likely to use drugs in the future

- Spreads of one drug are unrelated with each other.

- It is impossible for everyone in a region to use drugs, that is, at least one person does not use drugs.

- Drugs can spread between regions, but the speed can be different due to realistic constraints.

2.2 Nomenclature

Table2-1 Nomenclature in our paper

Notations	Definitions
$N^{(j)}$	the total number of individuals in the county numbered j
$N_1^{(j)}(t)$	the number of individuals using synthetic opioids or heroin
$N_0^{(j)}(t)$	the total number of individuals not injecting synthetic opioids or heroin
$\mathcal{A}=\{\alpha\|\alpha=FIPS\_COMBINED\}$	the collection of county FIPS codes within each state
$N_1=\left(N_1^{(j)}\right)$	the column vector of n∗1
$\hat{N}_1=\left(R-I_n\right)N$	the weighted number of drug individuals in other counties
η	the coefficient of δ
$\sigma_i\left(\mathrm{t}\right)$	social-economic features combinations obtained through GBDT.

$R=(r_{ij}), i、 j \in \mathcal{A}$	Pearson correlation coefficient matrix between the total number of individuals taking drugs
Λ	the rate that the susceptible S will become infected by injecting synthetic opioids and heroin through abnormal means
M	the rate that the infected person will change back to be susceptible under certain measures or means.
$\theta_{\alpha j} = r_{\alpha j}\beta$	the correlation transmission rate
B	the basic infection rate between counties.
η	the coefficient of δ
$\sigma_i(t)$	social-economic features combinations obtained through GBDT.
$\varphi(t)$	important feature
w	weight of important feature

3 Date Processing and visualization

3.1 the drug data Processing

The document (MCM_NFLIS_Data.xlsx) contains drug identification data in yea rs 2010-2017 for different substances in each of the counties from Ohio, Kentucky, West Virginia, Virginia and Pennsylvania , that is , the number of individuals that use narcotic analgesics (synthetic opioids) and heroin in each of the counties from the five states in years 2010-2017. The data has three dimensions: geography, time and the drugs categories.

The geographic latitude level is county and state, and the time dimension level is in years, that is, from 2010 to 2017. The minimum unit of this data is the number of in dividuals using a certain drug in a certain county in a certain year. We use python pan das library to manipulate the data in different dimensions.

3.1.1 Data visualization of the drug identification counts

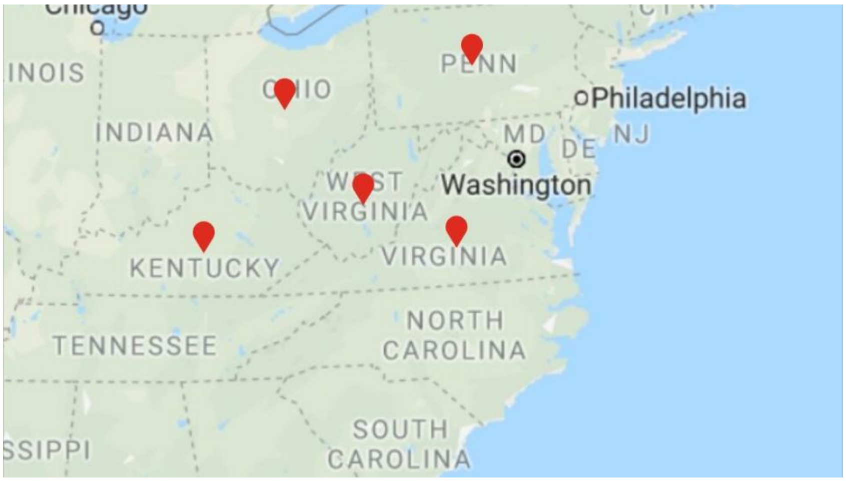


Figure 1 Distribution map of counties in five continents of the US

It can be seen from the figure 1 that the five states where the drug cases occur red are adjacent to each other, we can guess from the map that there must be certai n possibility of transmission and related relations among the states.

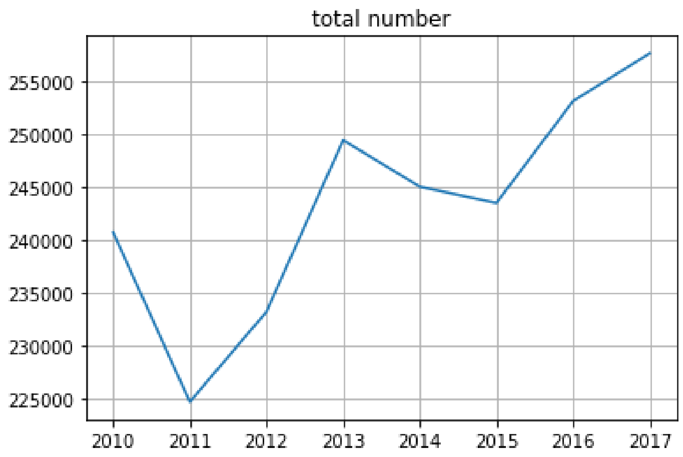


Figure 2 The Trend of Drug Cases Aggregation in Each State

As can be seen from the figure 2, from 2010 to 2011, the total number of drug outbreaks in various states dropped. From 2011 to 2017, the total number of drug outbreaks increased stepwise. Among them, the number declined slightly from 2013 to 2015 and then rose immediately.

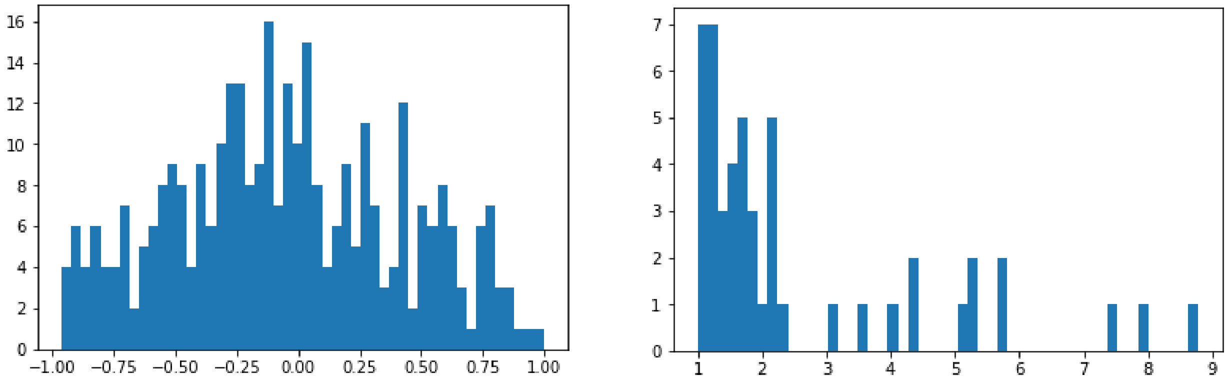


Figure 3 Histogram of Frequency Distribution of Total Growth Rate

We draw the histogram for the total growth rate from 2010 to 2017 in two figures a bove , the left one is for those growth below 1 and it distributes equally, there are still s ome states and counties whose growth rate exceeds 1 (shown on the right). Among the m, the growth rate in a few regions even lies between 7 and 9, indicating that the region is flooded with drugs and lax in supervision. The purpose of dividing the left and right p arts of the frequency distribution histogram range is to make the chart clearer and more

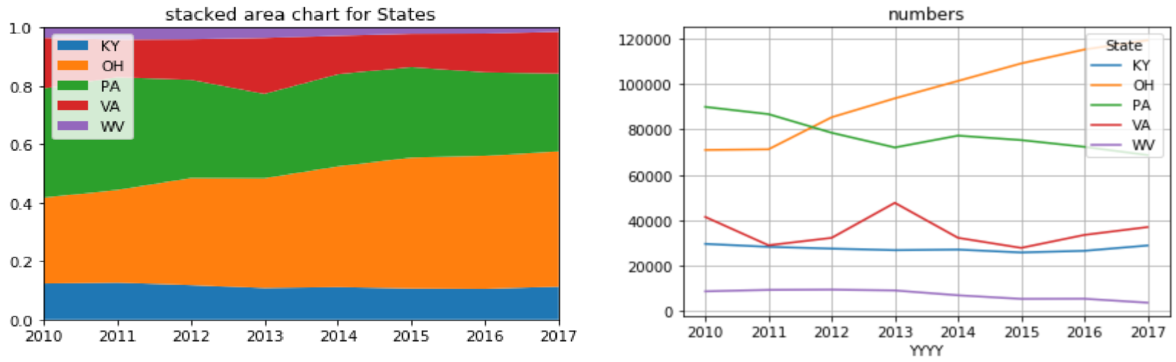


Figure 4 Stacked Area Chart and Line Chart of Frequency Distribution of Drug Cases in Each State

These two figures are the stacked frequency chart and the line chart of each states. OH was growing quickly

3.1.2 Correlation analysis

According to the data and the trend chart of the total number of drugs used, we find that the time series of the total number of drugs used between different counties are correlated, and at the same time, considering that synthetic opioids and heroin will not only spread in each state and county, but also spread between states under certain conditions, we analyze Pearson correlation[1] between counties. From Pearson correlation coefficient:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \qquad 3.1$$

According to the data in the table, the correlation coefficient matrix between counties, we assume that· $R = \left(r_{ij}\right), i、 j \in \mathcal{A}$, which indicates the Pearson correlation coefficient matrix between the total number of individuals taking drugs in a county, which reflects the degree of drug transmission between counties.

3.2 socio-economic data processing

The socio-economic data is a typical big data, which is from the U.S. Census Bureau represents a common set of socio-economic factors collected for the counties of these five states during each of the years 2010-2016. The data contains information about households, relationship , marital status, fertility, grandparents, educational attainment, school enrollment, veteran status and so on.

First, we deal with the missing data:1. For a variable that loses a large amount of data, we only delete it for that small data is not be able to provide valuable and enough information for our modeling. 2. For variables that lack a small amount of data, we use variable mean value in region to substitute. Then we standardized the remaining data.

4 Part I Analysis, evaluation and prediction of SIS model

4.1Preliminary preparation for model establishment

4.1.1 Propose SIS model

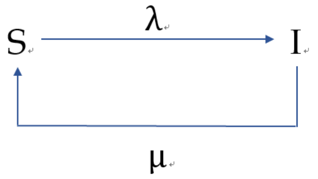
Based on the NFLIS data that provide a mathematical model to describe the spread and characteristics of synthetic opioids and heroin incidents (cases) reported between the five states and their counties over time, we have established a mathematical model to describe the spread and characteristics of synthetic opioids and heroin incidents (cases) reported Within each state and county in the United States over time.

Model Establishment: SIS Epidemic Model

Background: The infectious disease model has a long history. It is generally believed that it began in 1760 when Daniel Bernoulli[2] made a research on smallpox vaccination in one of his papers. After that, the process of establishing the infectious disease model went through a long period. Among them, the two most important models are SIS model and SIR model. In SIR model, susceptible population becomes infected population after being infected, but still has certain recovery ability, thus entering immune state (R). The people in this state are cured and have immunity, and will not infect others or be infected again. In SIS model, the susceptible population becomes the susceptible population after being cured, and does not have immune symptoms R, In this topic, people who inject synthetic opioids and heroin in non-medicinal occasions still have a certain probability of re-injection without immunity after being cured. Therefore, we build SIS mod el[3] to solve this problem.

In our model, we use S to represent the general population of the five states and counties in the United States that are easy to contact and inject synthetic opioids and heroin in non-medical occasions, that is, easy to be infected. At the same time, I indicates those groups that use synthetic opioids and heroin for entertainment （over-the-counter）,that is infected persons. And they have the opportunity to infect vulnerable people. λ represents the probability of being infected by an easily infected person, that is, the probability of being affected by an infected person to use synthetic opioids and heroin in non-medical situations. μ represents the probability that an individual will change from an infected person to a vulnerable person, i.e. stopping injecting synthetic opioids, i.e. heroin, for entertainment and other purposes. Arrows represent the direction and mechanism of conduction between individuals.

As can be seen, the model represents the probability λ that the susceptible S will become infected by injecting synthetic opioids and heroin through abnormal means, and μ that the infected person will change back to be susceptible under certain measures or means. The model is as follows.



Once the model and its corresponding dynamics are determined, an ordinary differential equation is derived as follows

$$\begin{cases} Ndi = \lambda Nsidt - \mu Nidt \\ s + i = N \end{cases} \tag{1}$$

Assume that T is the individual state space, T={0,1}, where individuals belonging to s are noted as being in state 0, that is, individuals using synthetic opioids or heroin. And the individuals belonging to I are recorded as being in state 1, that is, not injecting synthetic opioids or heroin.

Therefore, we use SIS model to build the differential equations that spread in each county.

$$\begin{cases} N_0^{(j)}(t) + N_1^{(j)}(t) = N^{(j)} \\ dN_1^{(j)}(t) = \dfrac{\lambda N_1^{(j)}(t)\, N_0^{(j)}(t)}{N^{(j)}} dt - \mu N_1^{(j)}(t)\, dt \end{cases} \tag{2}$$

According to the data in the table, the total number $N^{(j)}$ of individuals in each county was obtained through regression . For any j$\in \mathcal{A}$, we set up SIS model of correlated infectious diseases, where $N^{(j)}$ denotes the total number of individuals in the county numbered j, $N_1^{(j)}(t)$ denotes the number of individuals using synthetic opioids or heroin, and $N_0^{(j)}(t)$ denotes the total number of individuals not injecting synthetic opioids or heroin.

4.2 Improved model

4.2.1 Analysis of Drug Transmission between Counties

Consider that the correlation revealed according to the data and synthetic opioids and heroin will not only spread in each state and county but also spread between states under certain conditions, we determine the correlation between the use and development of synthetic opioids and heroin between states according to Pearson correlation coefficient so as to make further analysis. The correlation is expressed by the basic transmission rates of synthetic opioids and heroin between neighboring states. If the correlation is strong, it can be judged that the two states have a higher infection rate, that is, the two states are more likely to carry out mutual transmission of infection. At the same time, we weighted the basic infection rate to get the actual infection rate of synthetic opioids and heroin between counties and draw further conclusions.

Assume that $\mathcal{A}=\{\alpha|\alpha=FIPS\_COMBINED\}$ is the collection of county numbers within each state. $R=(r_{ij}), i、j \in \mathcal{A}$, which indicates the Pearson correlation coefficient matrix between the total number of individuals taking drugs in a county, which reflects the degree of drug transmission between counties. β represents the basic infection rate between counties.

According to the correlation, we weighted the basic transmission rate to obtain the correlation transmission rate

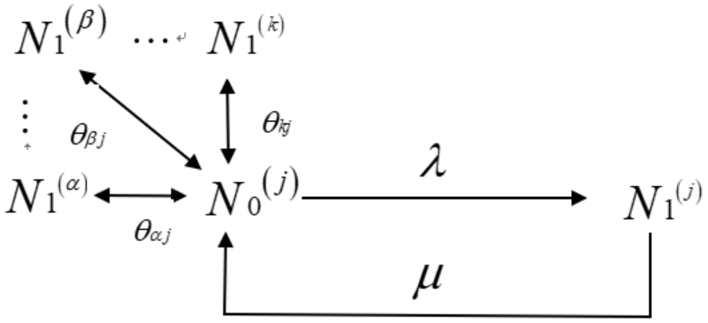
$$\theta_{\alpha j} = r_{\alpha j}\beta$$

Then the number of drug-using individuals spread by the total individuals in other areas in county I is

$$\left(\sum_{\alpha\in\mathcal{A}}\theta_{\alpha j}N_1^{(\alpha)}(t)\right)N_0^{(j)}(t)$$

4.2.2 Relevant Transmission -SIS model

Based on SIS model, considering the transmission between the relevant counties and the transmission within the counties. The propagation process is as follows.



we have established the following Relevant Transmission -SIS model.

$$\begin{cases} N_0^{(j)}(t)+N_1^{(j)}(t)=N^{(j)} \\ dN_1^{(j)}(t)=\left(\lambda N_1^{(j)}(t)+\sum\limits_{\alpha\in\mathcal{A},\ \alpha\neq j}r_{\alpha j}\beta N_1^{(\alpha)}(t)\right)\dfrac{N_0^{(j)}(t)}{N^{(j)}}dt-\mu N_1^{(j)}(t)dt \end{cases} \quad (3)$$

Considering the characteristics of the data, we discretize the differential equations and construct the differential equations model.

$$\begin{cases} N_0^{(j)}(t)+N_1^{(j)}(t)=N^{(j)} \\ N_1^{(j)}(t+\Delta t)-N_1^{(j)}(t)=\left(\lambda N_1^{(j)}(t)+\sum\limits_{\alpha\in\mathcal{A},\ \alpha\neq j}r_{\alpha j}\beta N_1^{(\alpha)}(t)\right)\dfrac{N_0^{(j)}(t)}{N^{(j)}}\Delta t-\mu N_1^{(j)}(t)\Delta t \end{cases} \quad (4)$$

4.3 results of the model

4.3.1 Origin of drugs

First, we screen the data of each drug and use the geographical distribution of individuals using each drug to screen. We believe that if individual users of a certain drug only appear in one state, then the drug begins in that state. According to this method, we screened and determined the birthplace of 19 drugs.

Secondly, for the remaining xx drugs, we found the starting time when it appeared in the report. If the individual users of this drug only appeared in one state in that year, then this drug began in that state. On the other hand, if the individual drug user appears in multiple states in that year, the state with the largest total drug user is considered as the birthplace.

According to our model, this method is reasonable.

Kentucky	Mitragynine','Pethidine','Methorphan','Pentazocine','Thebaine','Hydrocodone', fentanyl', fentanyl', 'Cyclopentyl fentanyl', 'Desmethylprodine ','Cyclopropyl fentanyl', 'Cyclopentyl 'Methadone','Desmethylprodine ','Cyclopropyl, 'Methadone'
Ohio	Acetyl fentanyl', 'Cyclopropyl/Crotonyl, 'Methoxyacetyl fentanyl', Fentanyl', 'Dextropropoxyphene', 'Crotonyl fentanyl' 'Fluorofentanyl','p-Fluorobutyryl fentanyl','4-Methylfentanyl', 'o-Fluorofentanyl','Carfentanil','U-51754','p-Fluorofentanyl','U-49900','Furanyl fentanyl', 'Phenyl fentanyl', 'Tetrahydrofuran fentanyl', 'Acryl fentanyl','Fluoroisobutyryl,fentanyl','U-48800','3Methylfentanyl','U47700','Morphine','Oxycodone','Oxymorphone','Propoxyphene','Opiates','Benzylfentanyl'
Pennsylvania	'Dihydrocodeine','Metazocine','MT-45','Buprenorphine','Valeryl fentanyl','Dihydromorphone','Heroin','Hydrocodeinone','cis-3-methylfentanyl','p-methoxybutyryl fentanyl', 'Butyryl fentanyl', '4-Fluoroisobutyryl fentanyl','Alphaprodine','3,4-Methylenedioxy U-47700','Codeine','trans-3-Methylfentanyl','Fentanyl', 'Remifentanil''Acetyldihydrocodeine','Levorphanol', 'ANPP'
Virginia	'Opium','Hydromorphone','Meperidine','3Fluorofentanyl','Nalbuphine',' Butorphanol',Tramadol',
West Virginia	Furanyl/3-Furanyl fentanyl','Isobutyryl fentanyl','Fluorobutyryl fentanyl ','Acetylcodeine'

4.3.2 parameters estimation

According to equation (4), considering the practical significance and correlation of the parameters, we consider using the least squares estimation with constraints, that is, limiting the parameter space to a certain range to prevent over - fitting. Because the transmission rate is less than 1 and the total number of individuals is greater than the number of individuals using drugs, we obtain the following constraints:

$$\begin{cases} 0 < \beta < 1 \\ 0 < \lambda < 1 \\ 0 < \mu < 1 \\ N^{(j)} > \max\left(N_1^{(j)}(t)\right) \end{cases} \tag{5}$$

In actual fitting, the parameters with practical significance can be expressed as the spread score of the county and all other counties. As shown in the following table （this table is the score corresponding to explanatory variables and cross items）.

variable	**score**	**Parameter expression**
$N_1^{(j)}$	a	$\lambda - \mu$
$\hat{N}_1$	b	β

$\left(N_1^{(j)}\right)^2$	c	$-\dfrac{\beta}{N^{(j)}}$
$N_1^{(j)}\hat{N}_1$	d	$-\dfrac{\lambda}{N^{(j)}}$

where $N_1=\left(N_1^{(j)}\right)$ is the column vector of n×1; $\hat{N}_1=\left(R-I_n\right)N$ indicates the weighted number of drug individuals in other counties.

Using this model and constrained least square method， we calculate parameters and curves of heroin and synthetic opioid:

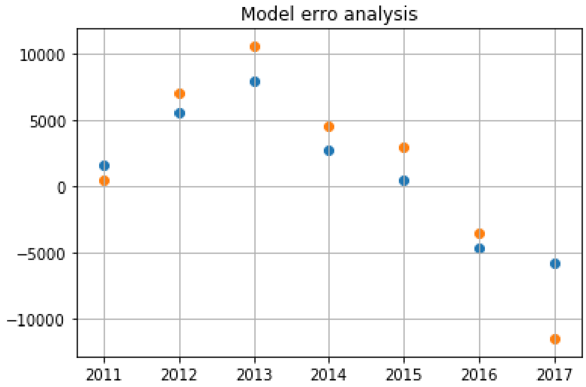


Figure 5 Comparison of Heroin Fitting Data and Original Data

The RMSD[4] of predicted values for times t of a regression's dependent variable is computed for n different predictions as the square root of the mean of the squares of the deviations:

$$\mathrm{RMSD}=\sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t-y_t)^2}{n}}.$$

The total square root of the mean of the squares of the deviations of heroin models in all counties is 2734.38118, which means the average RMSD is 5.931412. According to the comparison chart of fitting and original data， we see the heroin curve fits well to real data， but synthetic opioid curve fits not that well to real data.However, we have generated the curve of synthetic opioid for the next few years and found that the curve trend is very consistent with the original data. We think that our model has a certain lag, but the lag is not serious, and it can still reflect the changing trend and characteristics of synthetic opioid.

（图）

4.3.4 Forecast result

According to the previous model based on data, we predicted the total drug use in the next xx years, and the results are as follows:

Unfortunately, according to this model, the data after more than x years show that ... the prediction effect is not good. We consider adjusting the parameters, … , the results are as follows:

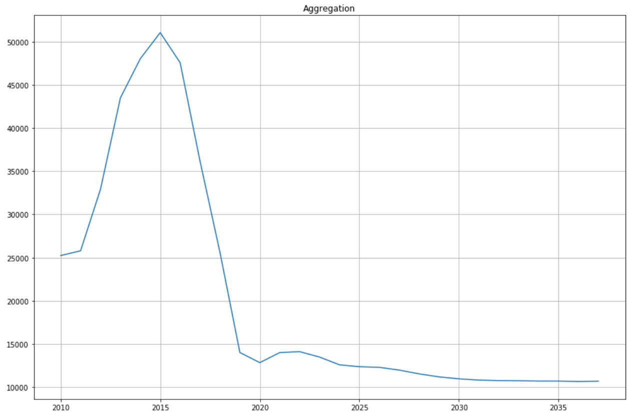


Figure 5 Total number of heroin injection cases by state

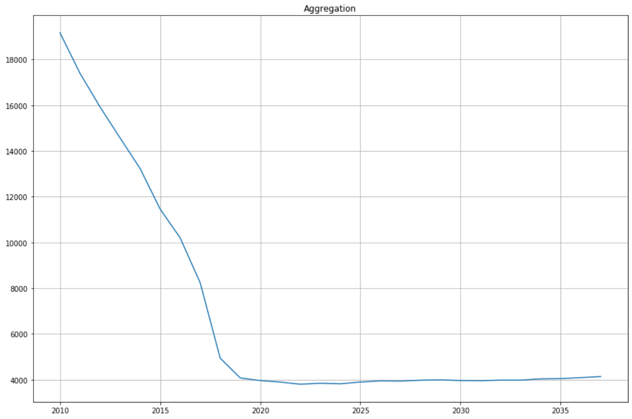


Figure 6 Total number of Oxycodone injection cases by state

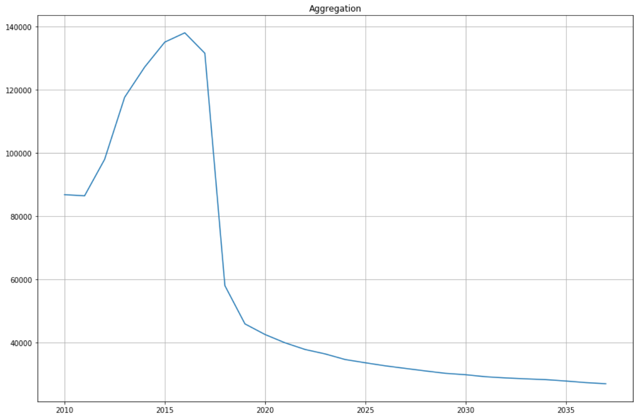


Figure 7 Total number of injection cases of other drugs by state

As can be seen from figs. 5, 6 and 7, the total amount of some drug users will tend to be stable in the future after a fluid.

As heroin and oxycodone injection events account for a large proportion of the total number of incidents, we will analyze heroin and oxycodone separately. According to fig.5, the total amount of heroin users goes up till it reach the peak and then will tend to be stable in the future while going down . Heroin injection cases reached the peak in 2015. According to fig.6, The number of oxycodone injection cases go down and tends to be stable in the future.It means that the number of oxycodone injection cases has reached the peak before 2010, thus showing a downward trend in the figure. The number of injecting cases of other drugs reached the explosion point near 2016, then decreased and showed a stable trend in the future.

The following three charts describe the distribution of cases of heroin, oxycodone and other drugs in each county from 2010 to 2035 and their forecast trend. The trend of different counties is shown by curves of different colors.

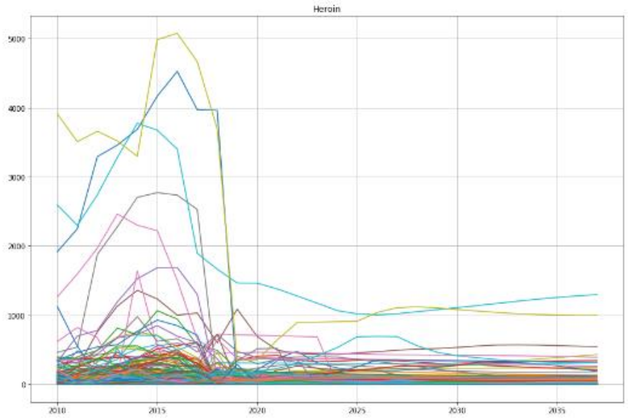


Figure 8 Forecast of Heroin Injection Cases in Counties

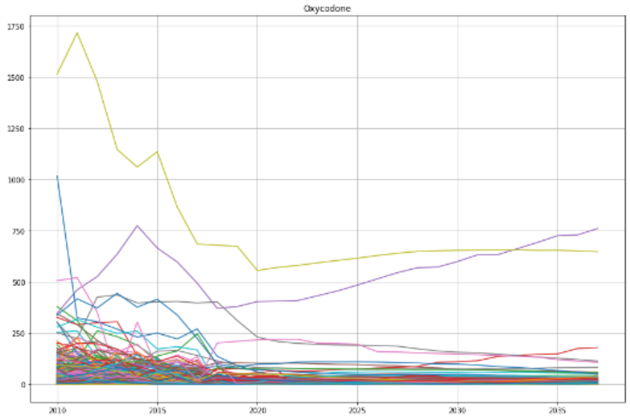


Figure 9 Forecast of Oxycodone Injection Cases in Counties

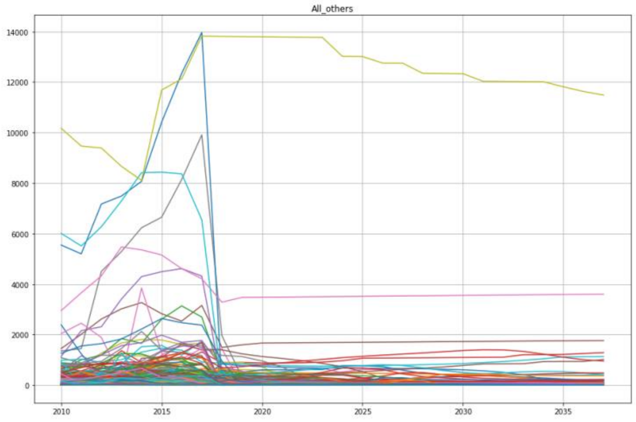


Figure 10 Forecast of Injection Cases of Other Drug in Counties

We found that different drugs and different counties have different equilibrium values and equilibrium times. As can be seen from the above three figures, as for the use of heroin and Oxycodone and other drugs, most counties will stabilize near 2021, which indicates that before that, there has been a sharp increase in the number curve caused by a large amount of injected drugs in these counties。

We found that the trend of total drug use in each county depends on the initial value. For counties where the total amount of initial drug use is greater than the equilibrium value, the total amount of initial drug use will decrease and eventually reach the equilibrium point (some drugs will first increase to reach the peak value and finally decrease and reach the equilibrium point); For counties where the total amount of starting drugs used is greater than the equilibrium value, the total amount of starting drugs used will decrease and increase and eventually reach the equilibrium point. This means that the total amount of new drugs with low initial drug is on the rise and will not peak. This is consistent with the model.

So we generated the county balance values and balance schedules (see appendix) The following figure shows the balance time distribution and balance value distribution of heroin total usage in each county.

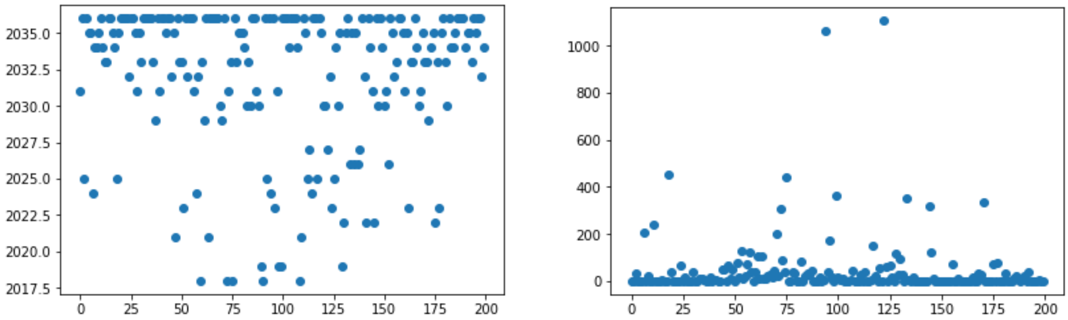


Figure 11 the balance time and value distribution of heroin total usage in each county.

It is worth noting that the problem of heroin and oxycodone drug use is serious in Kent province. among its counties, Allen's heroin stationary value is 48 times of the average stationary value and Adair's oxycodone stationary value is 72 times of the average stationary value. For the remaining synthetic opioid, Philadelphia county in Pennsylvania state has a serious drug problem, with a stationary value 216 times the average stationary value.

What can be also seen in fig.8,9,10 is that the total number of individuals using some drugs increase explosively in some years. Different drugs and different counties have different peaks and outbreak times. We reasonably suspect that there will be an explosion of drug use at this time.

For example, for heroin injection, some counties had reached the peak around 2015, with different degrees of explosion. As for the injection of Oxycodone, some counties had reached the peaks around 2012 and 2014, that is, a large number of drugs were taken in those counties. For injection of other drugs, there will be an explosion point in 2016, and then it will drop rapidly and stabilize. However, the number of drug injections in individual counties will remain high.

Therefore, we have generated the peak value and outbreak schedule of each county (appendices 1, 2 and 3). the following figure shows the outbreak time distribution and peak value distribution of the total amount of heroin used in each county.

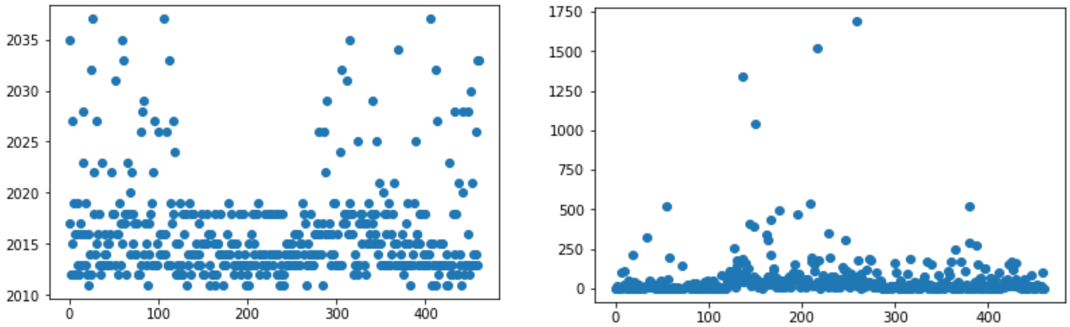


Figure 12 the outbreak time and value distribution of the total amount of heroin used in each county

If the patterns and characteristics of our model continue, the government's focus should be on peak value, outbreak time, equilibrium value and equilibrium time. To say it more specifically:1. The total amount of heroin and other synthetic opioids injected in most counties has reached a peak in the past and will show a stable state in the future, but its stable value is still at a high level. Therefore, the government should still focus on using certain effective means and measures to vigorously control drugs. 2. New drugs may still be in a stable state with a high value, that is, it may develop into epidemic drugs or trends. 3. Drug use is characterized by geographical clustering. For any drug control, the province should be taken as the unit, the county with the highest stable value should be taken as the center, and the circulation between counties should be controlled. 4, individual counties have very high stable values, so the government should pay more attention to these counties and take effective measures to focus on governance. For example, in Kentucky state, the heroin quantity in Allen county and the Oxycodone quantity in Adair county are both at high values. Therefore, the government should strengthen the governance of such counties.

5 Part two analysis

5.1 Background

As a substance harmful to human health, drugs are closely related to the economic development of society. Drugs will reduce the individual's ability to work and enthusiasm for work, and reduce the available labor force in society, thus inhibiting economic growth and producing negative effects on social and economic development. Based on this contradictory background, we will further analyze the problem and the model.

5.2 GBDT generates feature combinations

The socio-economic data has many variables and is a typical big data. So we consider using machine learning to screen variable. GBDT[6] (Gradation Boosting Decision Tree) is an iterative decision tree algorithm, which consists of multiple decision trees, and the conclusions of all trees are accumulated to make the final answer. Because GBDT can flexibly process various types of data, including continuous values and discrete values, and can effectively filter features, it is very suitable for processing socio-economic data. The flowchart of GBDT algorithm is as follows:

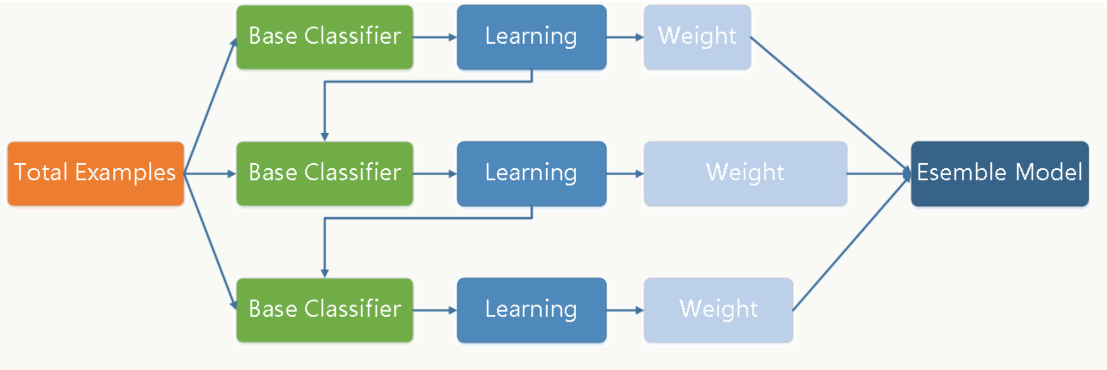


Figure 13 The flowchart of GBDT algorithm

We try to create an index constructed by all the important socio-economic features that can help explain the difference in drug usage, that is, the change in drug usage.

Firstly, we use GBDT combined with drug identification data and socio-economic data to do feature processing。The figure below shows that importance of each feature

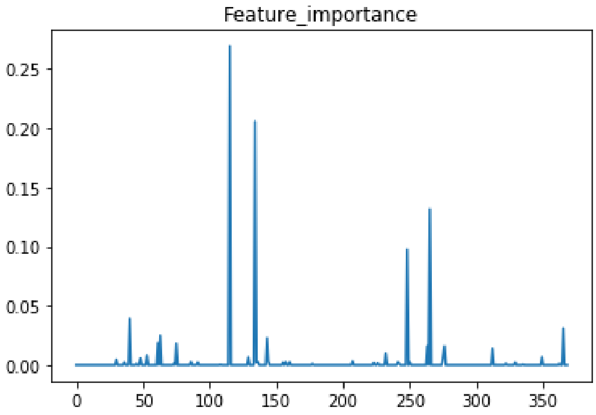
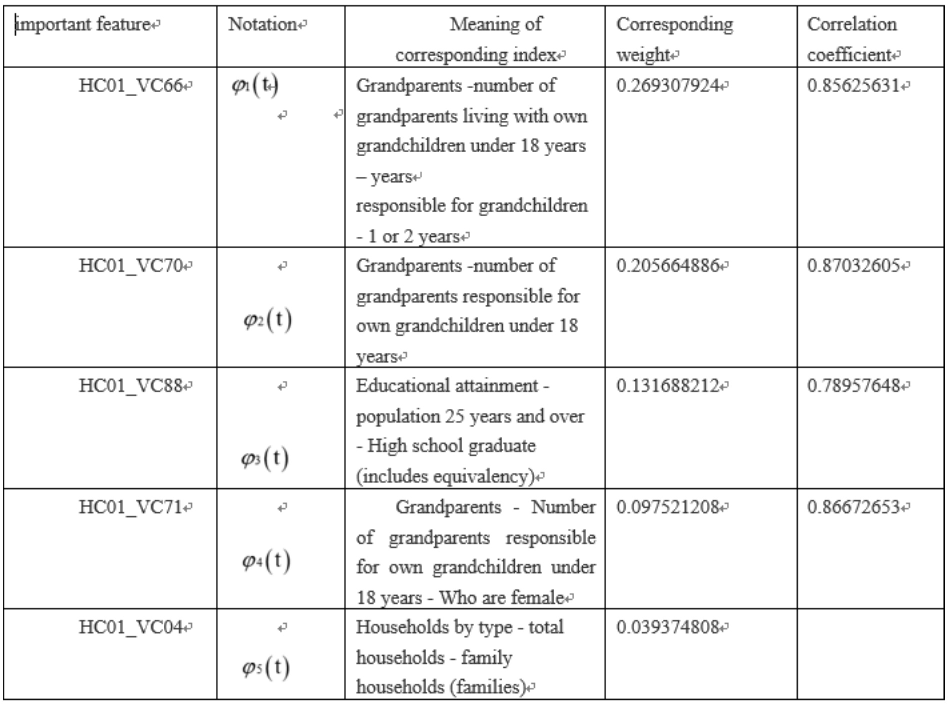


Figure 14 the importance of each feature

We checked the error of GBDT model. Then, we select the top 4 features of importance and get the corresponding weights. Then analyze the correlation between features and drug usage. The following table shows the important features, corresponding index significance, corresponding weight and correlation.



important feature	Notation	Meaning of corresponding index	Corresponding weight	Correlation coefficient
HC01_VC66	$\varphi_1(t)$	Grandparents -number of grandparents living with own grandchildren under 18 years – years responsible for grandchildren - 1 or 2 years	0.269307924	0.85625631
HC01_VC70	$\varphi_2(t)$	Grandparents -number of grandparents responsible for own grandchildren under 18 years	0.205664886	0.87032605
HC01_VC88	$\varphi_3(t)$	Educational attainment - population 25 years and over - High school graduate (includes equivalency)	0.131688212	0.78957648
HC01_VC71	$\varphi_4(t)$	Grandparents - Number of grandparents responsible for own grandchildren under 18 years - Who are female	0.097521208	0.86672653
HC01_VC04	$\varphi_5(t)$	Households by type - total households - family households (families)	0.039374808	

Results show that the number of grandparents responsible for or living with own grandchildren under 18 years is positively related to drug usage, and the number of people and 25 and above with only high school qualifications is positively related to drug usage.

Then we use the weighted data to obtain an index. In more detail, all the important socio-economic features construct this index to explain the difference in drug usage, that is, the change in drug usage.

$$\mathrm{w}_1\varphi_1(\mathrm{t})+\mathrm{w}_2\varphi_2(\mathrm{t})+\mathrm{w}_3\varphi_3(\mathrm{t})+\mathrm{w}_4\varphi_4(\mathrm{t})+\mathrm{w}_5\varphi_5(\mathrm{t})$$

Where $\varphi(\mathrm{t})$ denotes the important feature and w denotes the weight of the important feature.

It reflects the socio-economic characteristics of a certain region in a certain year.

5.3 Establishment of Social-Economic-Relevant transmission-SIS model

According to the index we obtained in the previous step, we modified the Relevant Transmission -SIS model to obtain the social-economic-relevant transmission-sis model. This model is reasonable because (指标)directly explains the changes in drug usage.

$$\begin{cases} N_0{}^{(j)}(t)+N_1{}^{(j)}(t)=N^{(j)} \\ dN_1{}^{(j)}(t)=\left(\lambda N_1{}^{(j)}(t)+\sum\limits_{\alpha\in\mathcal{A},\ \alpha\neq \mathrm{j}} r_{\alpha j}\beta N_1{}^{(\alpha)}(t)\right)\dfrac{N_0{}^{(j)}(t)}{N^{(j)}}dt-\mu N_1{}^{(j)}(t)dt+\eta\sigma_i(t)dt \end{cases} \tag{5}$$

Where η denotes the coefficient of δ ; $\sigma_\mathrm{i}(\mathrm{t})$ denotes social-economic features combinations obtained through GBDT.

6 Strategy for Countering the Opioid Crisis

According to the social-economic-relevant transmission-sis model we have established and the GBDT feature selection methods:

(1)the grandparent-child household and the educational level has great feature importance and big correlation coefficient.

(2)the transmission between different counties and regions can be significant, it makes the whole states as a system.

(3)each counties and each drugs has their own property and auto relation.

6.1 Help addicts

According to the social-economic-relevant transmission-sis model, μ represents the rate that the infected person will change back to be susceptible under certain measures or means. We believe that providing help to addicts is an effective strategy for countering the opioid crisis, and because providing help to addicts can increase and reduce future drug use.

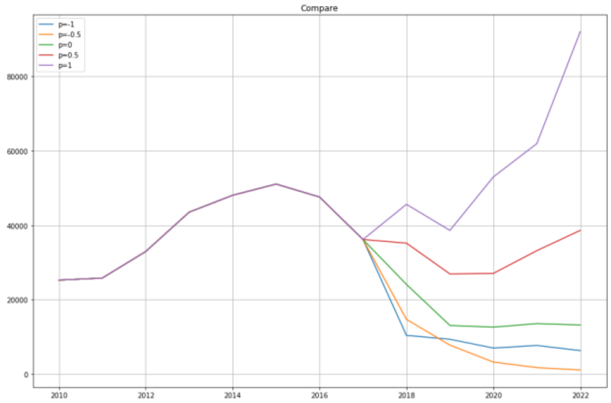
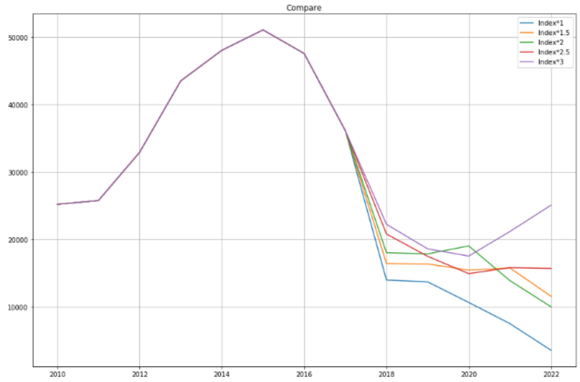


Figure 15

6.2 Increase Financial Investment in Local Education

As the model we have established the number of people aged 25 and above with only high school qualifications are positively related to drug usage. Increasing the financial investment in local education will directly encourage more people to go to university, thus reducing the number of people aged 25 and above who only have a high school diploma.



7 Sensitivity analysis

We change two key parameters a and b of the explanatory variables and also do the simulation for the index we get from part two (because we don’t have the data for future).The sensitive analysis proved that our model is meaningful and successful:

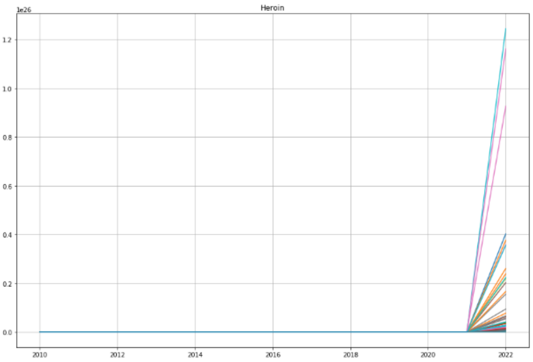


Figure 17

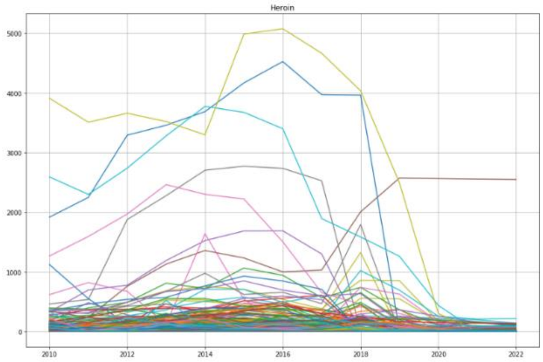


Figure 18

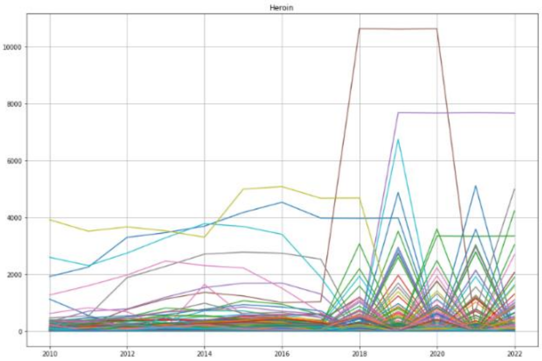


Figure 19

For the parameter b of Regional impact, we set it within [0.001,0,0.001], and we find it is highly sensitive, but this is simply because the impact is very large in scale. What interest us most is when b for all counties is all negative, the whole system turns to recycle (f19) and when b for all counties is all 0, they grow totally by themselves(f18) and when b for all counties is all positive, they accumulate the growth then bomb into infinity. (f17).

For other parameters and explanatory variables, we have don’t it in the previous part.

8.Conclusion

9. Strengths and Weaknesses

9.1 Strengths

Based on the mechanism of drug transmission, we have constructed a reasonable and realistic（social - economic - ）relevant transmission - sis model. The model has practical significance and provides theoretical basis for formulating prevention and control strategies and measures.

1. After a systematic consideration， we improved the traditional SIS model, weighted the mutual transmission rate by the correlation coefficient between regions, and innovatively dealt with the drug transmission problem between different regions.

2.GBDT's method of dealing with socio-economic data is obviously superior to other non-machine learning algorithms.

9.2 weakness

1.We have neglected the interaction between drugs, such as substitution. That is, if an infected person abandons injecting a drug, he may choose to inject another drug instead.

Memorandum

To: the Chief Administrator, DEA/NFLIS Database

From: Team #72969

Date: Jan 28th, 2019

Subject: Energy Profiles Characterization, Prediction and Future Goals

Honorable Chief Administrator of DEA/NFLIS Database

Currently, the United States is experiencing a national opioid crisis. Based on the drug data from 2010 to 2017 obtained from your database, our team has made a comprehensive study on the spread of the reported synthetic opioid and heroin incidents in and between the five states and their counties over time, including modeling the spread of specific drugs, identifying Origin of drugs, finding out characteristics of the drug spread, predicting drug use to 2035, identifying and testing possible strategies for countering the opioid crisis.

spread of drugs and predictions: We create Relevant Transmission -SIS model to...

We make predictions of drug use from 2018 to 2035 and found out that drug use counts in each county will eventually tend to reach the equilibrium point in the future. Most drug use will not converge till 2026.Furthermore, the trend of total drug use in each county depends on the initial value. Drug with initial value higher than the equilibrium value will increase explosively before going down and converge to equilibrium value. Besides, the total amount of heroin and other synthetic opioids injected in most counties has reached a peak in the past and will show a stable state in the future, but its equilibrium value is still at a high level.

Characteristics of drugs spread: equilibrium value and time, peak value and time, geographical clustering...

Concerns:

1.the government should focus on peak value, outbreak time, equilibrium value and equilibrium time.

2.equilibrium value can be used as drug identification threshold

3. Drug use is characterized by geographical clustering. For any drug control, the province should be taken as the unit. Especially, after all the drugs use in 5 states converge, the drugs use in Pennsylvania will predominate by 54%.

4. Several individual counties have very high stable values, for example， Allen in Kentucky state.

5.According to the social-economic-relevant transmission-sis model we have established, The number of grandparents responsible for or living with own grandchildren under 18 years old and the number of people aged 25 and above with only high school qualifications are positively related to drug usage.

Strategies:helping addicts and increasing financial investment in local education is efficient strategies.

The above is the summary of our study. We sincerely hope that it will provide you with useful information.

Thanks!

Reference

[1]Adler J, Parmryd I. Quantifying colocalization by correlation: The Pearson correlation coefficient is superior to the Mander's overlap coefficient[J]. Cytometry Part A, 2010, 77a(8):733-742.

[2] Bacaër N. Daniel Bernoulli, d’Alembert and the inoculation of smallpox (1760)[M]// A Short History of Mathematical Population Dynamics. 2011.

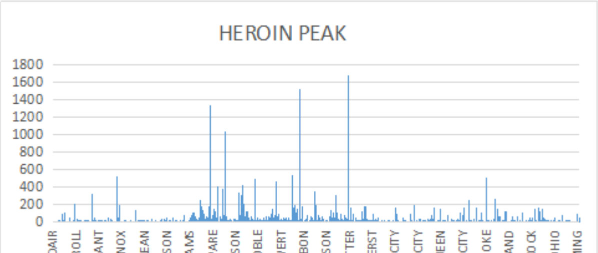
[3] 田蓓蓓，李青，周美莲．复杂网络上病毒传播的元胞自动机模拟[J]．计算机工程, 2008, 34(23):278-279.

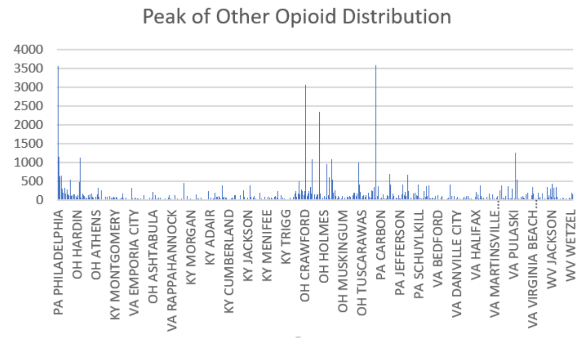
[4] Kirchmair J, Markt P, Distinto S, et al. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection--what can we learn from earlier mistakes?[J]. J Comput Aided Mol Des, 2008, 22(3-4):213-228.

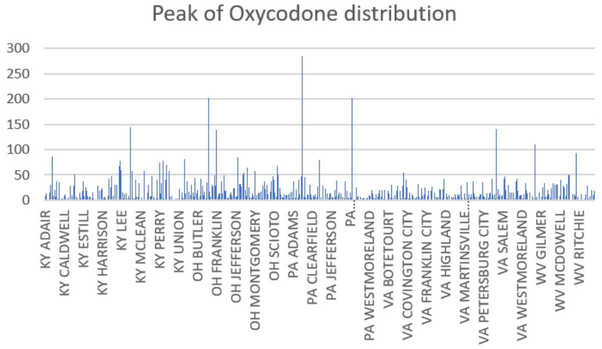
[5] https://www.drugabuse.gov/drugs-abuse/opioids

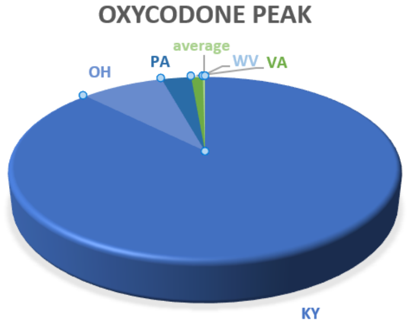
[6] 郑凯文，杨超．基于迭代决策树(GBDT)短期负荷预测研究[J]．贵州电力技术, 2017, 20(2):82-84.

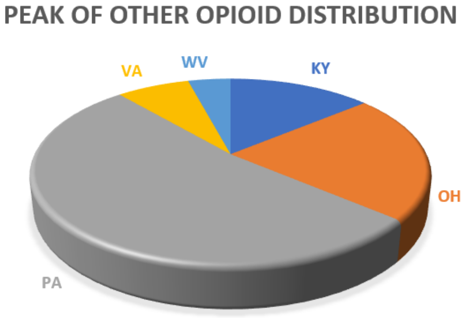
Appendix A Figures of peak and balanced distribution

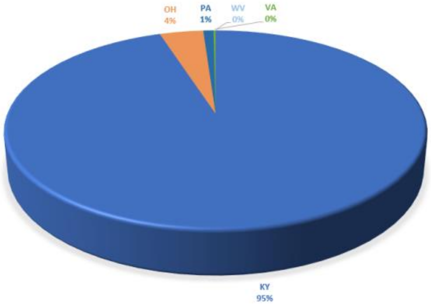


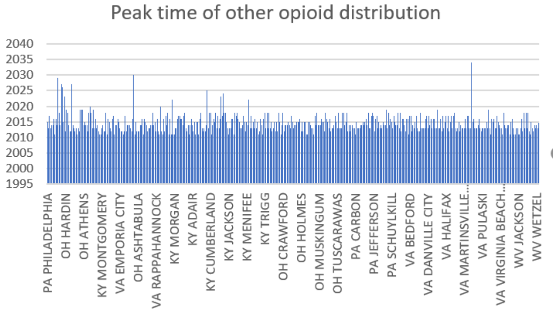

Peak of Other Opioid Distribution
4000
3500
3000
2500
2000
1500
1000
500
0
PA PHILADELPHIA
OH HARDIN
OH ATHENS
KY MONTGOMERY
VA EMPORIA CITY
OH ASHTABULA
VA RAPPAHANNOCK
KY MORGAN
KY ADAIR
KY CUMBERLAND
KY JACKSON
KY MENIFEE
KY TRIGG
OH CRAWFORD
OH HOLMES
OH MUSKINGUM
OH TUSCARAWAS
PA CARBON
PA JEFFERSON
PA SCHUYLKILL
VA BEDFORD
VA DANVILLE CITY
VA HALIFAX
VA MARTINSVILLE...
VA PULASKI
VA VIRGINIA BEACH...
WV JACKSON
WV WETZEL

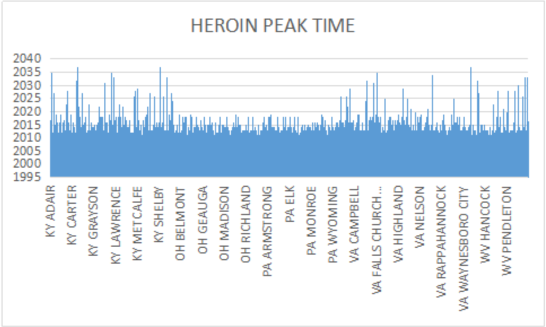

Peak of Oxycodone distribution
300
250
200
150
100
50
0
KY ADAIR
KY CALDWELL
KY ESTILL
KY HARRISON
KY LEE
KY MCLEAN
KY PERRY
KY UNION
OH BUTLER
OH FRANKLIN
OH JEFFERSON
OH MONTGOMERY
OH SCIOTO
PA ADAMS
PA CLEARFIELD
PA JEFFERSON
PA...
PA WESTMORELAND
VA BOTETOURT
VA COVINGTON CITY
VA FRANKLIN CITY
VA HIGHLAND
VA MARTINSVILLE...
VA PETERSBURG CITY
VA SALEM
VA WESTMORELAND
WV GILMER
WV MCDOWELL
WV RITCHIE

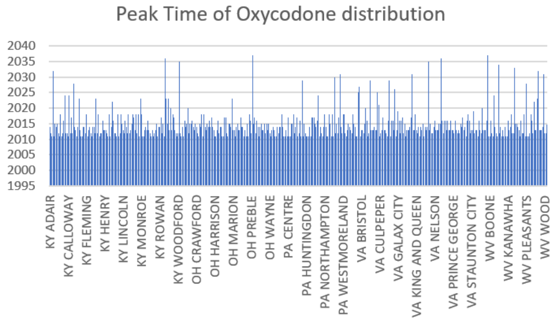

OXYCODONE PEAK
average
OH
PA
WV
VA
KY


PEAK OF OTHER OPIOID DISTRIBUTION
VA
WV
KY
OH
PA


OH
4%
PA
1%
WV
0%
VA
0%
KY
95%


Peak time of other opioid distribution
2040
2035
2030
2025
2020
2015
2010
2005
2000
1995
PA PHILADELPHIA
OH HARDIN
OH ATHENS
KY MONTGOMERY
VA EMPORIA CITY
OH ASHTABULA
VA RAPPAHANNOCK
KY MORGAN
KY ADAIR
KY CUMBERLAND
KY JACKSON
KY MENIFEE
KY TRIGG
OH CRAWFORD
OH HOLMES
OH MUSKINGUM
OH TUSCARAWAS
PA CARBON
PA JEFFERSON
PA SCHUYLKILL
VA BEDFORD
VA DANVILLE CITY
VA HALIFAX
VA MARTINSVILLE...
VA PULASKI
VA VIRGINIA BEACH...
WV JACKSON
WV WETZEL


HEROIN PEAK TIME
2040
2035
2030
2025
2020
2015
2010
2005
2000
1995
KY ADAIR
KY CARTER
KY GRAYSON
KY LAWRENCE
KY METCALFE
KY SHELBY
OH BELMONT
OH GEAUGA
OH MADISON
OH RICHLAND
PA ARMSTRONG
PA ELK
PA MONROE
PA WYOMING
VA CAMPBELL
VA FALLS CHURCH...
VA HIGHLAND
VA NELSON
VA RAPPAHANNOCK
VA WAYNESBORO CITY
WV HANCOCK
WV PENDLETON


Peak Time of Oxycodone distribution
2040
2035
2030
2025
2020
2015
2010
2005
2000
1995
KY ADAIR
KY CALLOWAY
KY FLEMING
KY HENRY
KY LINCOLN
KY MONROE
KY ROWAN
KY WOODFORD
OH CRAWFORD
OH HARRISON
OH MARION
OH PREBLE
OH WAYNE
PA CENTRE
PA HUNTINGDON
PA NORTHAMPTON
PA WESTMORELAND
VA BRISTOL
VA CULPEPER
VA GALAX CITY
VA KING AND QUEEN
VA NELSON
VA PRINCE GEORGE
VA STAUNTON CITY
WV BOONE
WV KANAWHA
WV PLEASANTS
WV WOOD

```
# This script is used to do the Data visualiation and Data Preprocessing

# Data visualiztion and manipulation
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# 获取文件路径和表格名
S =
'C://Users/26529/Desktop/Data/2018_MCMProblemC_DATA/MCM_NFLIS_Data.xls
x'

s = 'Data'

def Get_Data(file, sheetname):

    # 获取数据

    Data_itseIf = pd.read_excel(file, sheet_name=sheetname)
    return Data_itself

Data = Get_Data(S,s)

Data['Total_name'] = Data['State']+' '+Data['COUNTY'] #通过这个标签来唯一标识




# Draw the frequency accumulation curve for the states

# 画出频率累积曲线对 States 来划分而言

X = Data[['YYYY','TotalDrugReportsState','State']]

X=X.drop_duplicates()

S = X.groupby('YYYY')['TotalDrugReportsState'].cumsum()

X=X.pivot(index='YYYY',columns='State',values='TotalDrugReportsState')

# 计算总的各州的数据用于计算后面的比例

X.plot(title='numbers',grid=True)
```

```
X['T']=X['KY']+X['OH']+X['PA']+X['VA']+X['WV']

for i in X.columns:
    if i!='T':
        X[i] = np.array(X[i])/np.array(X['T'])


# Make the plot
plt.figure()
plt.stackplot(X.index,X['KY'],X['OH'],X['PA'],X['VA'],X['WV'],labels=['KY','OH','PA','VA','
WV'])
plt.legend(loc='upper left')
plt.margins(0,0)
plt.title('stacked area chart for States')
plt.show()


#  在这里我们计算一个总的趋势

S = S[4::5]    #这里我们取出来的是五个州的案例数的加和

S.index= list(range(2010,2018))

S.plot(title='total number',grid=True)

# Draw the histogram for the annual increase rate for each counties

#   计算 county 增长频率分布
X = Data[['YYYY','Total_name','TotalDrugReportsCounty']]

X = X.drop_duplicates()


n_day=7

def calcu_rate(A,n):
        A['return']
=A['TotalDrugReportsCounty']/A['TotalDrugReportsCounty'].shift(n) -1
        return A
    # Using the apply function to calculate the historical return for all the indexes
respectly
X=X.groupby('Total_name').apply(lambda x:calcu_rate(x,n_day))
```

```
X = X.dropna()

A = X[X['return']<1]

B = X[X['return']>1]

B = B[B['return']<60]

plt.figure()
plt.hist(A['return'].values,bins=50)
plt.show()

plt.figure()
plt.hist(B['return'].values,bins=50)
plt.show()
```

```
# Codes for Part1

# 1) package imported

import pandas as pd
import numpy as np
#import matplotlib.pyplot as plt
import statsmodels.regression as streg




# 获取文件路径和表格名
# 2) Get data source path and sheet name

S = 'C://Users/26529/Desktop/Data/2018_MCMProblemC_DATA/MCM_NFLIS_Data.xlsx'
s = 'Data'

def Get_Data(file, sheetname):

    #   Get Data

    Data_itself = pd.read_excel(file, sheet_name=sheetname)

    return Data_itself

Data = Get_Data(S,s)

Data['Total_name'] = Data['State']+' '+Data['COUNTY'] #通过这个标签 Total_name 来唯一标识


# 获取地区数据并计算他们之间的相关关系
# 3) Divide and aggregate data by county then calculate the relationship

X = Data[['YYYY','Total_name','TotalDrugReportsCounty']]

X = X.drop_duplicates()

X = X.pivot(index='YYYY',values='TotalDrugReportsCounty',columns='Total_name')

X = X.fillna(value=0)
```

```
T = X.corr()

# 计算每个药品的时间序列数据

D = Data[['YYYY','SubstanceName','Total_name','DrugReports']]

D=D.sort_index(by=['SubstanceName','YYYY'])

Substance_name = list(set(D.SubstanceName))    # 这里存放所有药物名

Total__name = list(set(D.Total_name)) # 这里存放所有 county 名

Raw_Data = dict()    # 这里存放所有药品数据

for i in Substance_name:

    d = D[D['SubstanceName']==i]

    #name_temp = list(set(d.Total_name.values))

    d=d.pivot(columns='Total_name',values='DrugReports',index='YYYY')

    dd = pd.DataFrame(columns=Total__name,index=d.index)
    dd[d.columns] = d[d.columns]
    dd = dd.sort_index(level=T.columns,axis=1)
    dd = dd.fillna(0)
    Raw_Data[i] = dd

# 这是我们第一个核心的线性模型;
'''
我们在这里对一阶滞后项
'''

def Model1(Corr, r):

    Future_data = dict() #为预测准备好起始的解释变量数据

    # 因子的构造

    n = r.shape[0]-1
    print(n)

    c = np.array(Corr)
    r_dff = np.diff(r,axis=0)    #成功构造被解释变量

    c=c-np.diag(np.diag(c))

    Impact = np.dot(r,c)[0:n]    #解释变量 1：外部影响因子  n*461
    Future_data['1'] = np.dot(r,c)[n].reshape([461,1]) #未来因子 1

    Impact_plus = Impact*r[0:n,:] #解释变量 2：   外部感染因子
    Future_data['2'] = Future_data['1']*(r[n,:].reshape([461,1])) #未来因子 2

    r1 = r[0:n,]
    Future_data['3'] = r[n,].reshape([461,1])

    r1_square = np.power(r[0:n,],2)
    Future_data['4'] = np.power(r[n,].reshape([461,1]),2)

    Beta0 = dict()
    Beta1 = dict()
    Beta2 = dict()
    Beta3 = dict()

    Result = dict()

    for i in range(461):

        Explain = np.array([Impact[:,i],Impact_plus[:,i],r1[:,i],r1_square[:,i]]).T

        Md2 = streg.linear_model.OLS(r_dff[:,i].reshape([n,1]),Explain)
        result = Md2.fit()
        Beta0[i] = float(result.params[0])
        Beta1[i] = float(result.params[1])
        Beta2[i] = float(result.params[2])
        Beta3[i] = float(result.params[3])
        Result[i] = result
    return Beta0, Beta1, Beta2, Beta3, Result, Future_data

#[a1,a2,a3,a4,R,F] = Model1(T,Raw_Data['Heroin'])


def Predict(Future_data, Result,r,a1,a2,a3,a4):

    Ran = 3
    Mean1 = np.mean(list(a1.values()))
    Mean2 = np.mean(list(a2.values()))
    Mean3 = np.mean(list(a3.values()))
    Mean4 = np.mean(list(a4.values()))
    Std1 = np.std(list(a1.values()))
    Std2 = np.std(list(a2.values()))
    Std3 = np.std(list(a3.values()))
    Std4 = np.std(list(a4.values()))
    Max1 = Mean1+Ran*Std1
    Max2 = Mean2+Ran*Std2
    #Max3 = Mean3+Ran*Std3
    #Max4 = Mean4+Ran*Std4
    Min1 = Mean1-Ran*Std1
    Min2 = Mean2-Ran*Std2
    #Min3 = Mean3-Ran*Std3
    Min4 = Mean4-Ran*Std4
    Max_3 = np.max([abs(np.min(list(a3.values()))),abs(np.max(list(a3.values())))])
    explain =
np.array([Future_data['1'],Future_data['2'],Future_data['3'],Future_data['4']])[:,:,0]

    Temp =np.zeros(461)


    '''
    进行严格的系数预处理和限制，
    a1 和 a2 必须符号相反
    且 a2 必须很小
    a3 在-1-1 之间
    a4 必须也很小并且是负数
    '''


    for i in range(461):

        if(Result[i].params[3]>0):

            Result[i].params[3]=0

        if(Result[i].params[3]<Min4):

            Result[i].params[3]=Min4

        #Result[i].params[2]= Result[i].params[2]/Max_3
        if(Result[i].params[2]>0):
            Result[i].params[2]=0


        if(Result[i].params[0]<Min1):

            Result[i].params[0]=Min1

        if(Result[i].params[0]>Max1):

            Result[i].params[0]=Max1

        if(Result[i].params[1]<Min2):

            Result[i].params[1]=Min2

        if(Result[i].params[1]>Max2):

            Result[i].params[1]=Max2

        #Result[i].params[2]=1
        Result[i].params[1]=0.001
        Temp[i] = Result[i].predict(explain[:,i])



    Temp[np.where(Temp==Temp.max())]=np.log(abs(Temp.max()))

    Temp[np.where(Temp==Temp.min())]=-np.log(abs(Temp.min()))

    k=1

    Temp = Temp+k*np.random.standard_normal(Temp.shape)

    r = np.insert(r,r.shape[0],values=Temp+r[r.shape[0]-1,:],axis=0)

    r[r<0]=0

    return r

def Find_the_first_State(Data):

    # Try to find where the drug derived from, the state
    # 找毒品的起源地
    # N  将是我们算法操作的唯一数据集；
    N = Data[['YYYY','SubstanceName','State','DrugReports']]

    #  我们通过一系列操作将 N 取出每一个 substance 最新的一年,并合理排序

    def Find_first(A):
        A['first_year'] =np.min(list(set(A.YYYY.values)))
        return A

    N=N.groupby('SubstanceName').apply(lambda x:Find_first(x))
    N=N.sort_index(by=['SubstanceName','YYYY'])

    #  药品的总名字和地区的总名字
    S_name = list(set(N.SubstanceName))
    ST_name = list(set(N.State))


    #  这里提前我们的模型的可能会在的州
    df_result = dict()


    for i in range(len(set(N.SubstanceName))):

        #  将每一个数据的药品都读取过来
        D = N[N['SubstanceName']==S_name[i]]

        #  如果一直在一个洲里面活动，就是那个洲
        if len(set(D.State.values))==1:
            df_result[S_name[i]]=list(set(D.State.values))[0]
        else:
            D = D[D['YYYY']==D['first_year']]
            D = D.groupby(by=['State'])['DrugReports'].sum()

            #  我们其实已经很轻松的将最先的一年数据找出来了
            #  接下来将最大的作为起始点就可以了
            df_result[S_name[i]]=D[D.values==D.max()].index[0]


    DF_RESULT = dict()
    #  将键值对调转

    for j in ST_name:
```

```
            DF_RESULT[j]=list()


        for k in list(df_result.keys()):

            DF_RESULT[df_result[k]].append(k)

        return DF_RESULT

# s=Find_the_first_State(Data)

def Get_New_Explain(Corr,r):


        Future_data = dict() #为预测准备好起始的解释变量数据



        n = r.shape[0]-1

        c = np.array(Corr)

        c=c-np.diag(np.diag(c))

        Future_data['1'] = np.dot(r,c)[n].reshape([461,1]) #未来因子 1

        Future_data['2'] = Future_data['1']*(r[n,:].reshape([461,1])) #未来因子 2

        Future_data['3'] = r[n,].reshape([461,1])

        Future_data['4'] = np.power(r[n,].reshape([461,1]),2)

        return Future_data

def Run_predict(Corr, rawdata,k_steps):

        r = np.array(rawdata)

        [a1,a2,a3,a4,R,F] = Model1(Corr, r)
        r=Predict(F,R,r,a1,a2,a3,a4)
        S = dict()
        for i in range(k_steps-1):

            [a1,a2,a3,a4,R,F] = Model1(Corr, r)
            S[i]=R

            r=Predict(F,R,r,a1,a2,a3,a4)


        return r,S


def Combine_Data(Raw_Data, Name):


S=pd.DataFrame(index=Raw_Data['Heroin'].index,columns=Raw_Data['Heroin'].column
s)

        S = S.fillna(0)

        n = len(Name)

        tem1 =
pd.DataFrame(data=Raw_Data[Name[0]],index=Raw_Data['Heroin'].index,columns=Ra
w_Data['Heroin'].columns)

        tem1 = tem1.fillna(0)

        # 这里我们输入列名就将想要的毒品数据进行加总计算
        S = S + tem1

        if n>1:
            for i in range(n-1):
                tem1 =
pd.DataFrame(data=Raw_Data[Name[i]],index=Raw_Data['Heroin'].index,columns=Ra
w_Data['Heroin'].columns)
                tem1 = tem1.fillna(0)
                S = S + tem1
        return S


def Get_Balance_point(x,k):

        n1 = x.shape[0]
        r = (x[1:n1]-x[0:n1-1])/x[0:n1-1]
        r[np.isnan(r)]=0
        r=list(r)
        r.reverse()
        for i in range(len(r)):
            if abs(r[i])>k:
                break
        C = len(x)-i
        return C



def Generate_chart(MMM,Raw_Data):


        # 我们先将原始数据转换为 DataFrame 格式
        Source_chart =
pd.DataFrame(data=MMM,index=list(range(2010,MMM.shape[0]+2010)),columns=Raw_
Data['Heroin'].columns)

        # 然后再将他进行查分之后取出差分部分
        S_diff = Source_chart.diff().iloc[1:,:]



        Information =
pd.DataFrame(index=Source_chart.columns,columns=['B_time','B_value','G_time','G_val
ue'])
        for i in S_diff.columns.values:

            Information.loc[i,'G_value']=S_diff[i].max()

Information.loc[i,'G_time']=S_diff.index[np.where(S_diff[i]==S_diff[i].max())].values[0]

            a = Get_Balance_point(Source_chart[i].values,0.05)

            Information.loc[i,'B_time']=a+2010-1
            Information.loc[i,'B_value']=Source_chart.loc[a+2010-1,i]

        return Information



#D1 =
pd.DataFrame(data=D_Heroin,columns=Raw_Data['Heroin'].columns,index=range(2010,
2018))

Substance_name.remove('Oxycodone')
Substance_name.remove('Heroin')
```

```
Data_1 = Combine_Data(Raw_Data,['Heroin'])
D_Heroin,S=Run_predict(T,Data_1,5)
D1 =
pd.DataFrame(data=D_Heroin,columns=Raw_Data['Heroin'].columns,index=range(2010,
2023))

Data_2 = Combine_Data(Raw_Data,['Oxycodone'])
D_Oxy,S=Run_predict(T,Data_2,5)
D2 =
pd.DataFrame(data=D_Oxy,columns=Raw_Data['Heroin'].columns,index=range(2010,20
23))

Data_3 = Combine_Data(Raw_Data,Substance_name)
D_allother,S=Run_predict(T,Data_3,5)
D3 =
pd.DataFrame(data=D_allother,columns=Raw_Data['Heroin'].columns,index=range(201
0,2023))

#D1.sum(axis=1).plot(title='Aggregation',grid=True,figsize=[15,10])
'''
plt.scatter(range(len(I1['G_value'].values)),I1['G_value'].values)
'''
'''
Real_data = Data_3


Predicted_data = pd.DataFrame(columns=Real_data.columns,index=Real_data.index[1:])

for i in range(461):
    r_p = list(R[i].predict())
    for j in range(7):
        Predicted_data.iloc[j,i]=r_p[j]



Real_data1=Real_data1.loc[2011:]

PPP=Predicted_data.sum(axis=1)
RRR=Real_data1.sum(axis=1)

plt.figure()
plt.scatter(np.array(range(2011,2018)).T,np.array(PPP.values).T)
plt.scatter(np.array(range(2011,2018)).T,np.array(RRR.values).T)
plt.title('Model erro analysis')
plt.grid()
plt.show()

np.sqrt(np.square(PPP.values-RRR.values).sum()/7)
'''

plt.figure(figsize=[15,10])
plt.grid()
plt.title('Compare')
plt.plot(d1.sum(axis=1),label='Beta<0')

plt.legend('1')
plt.plot(d2.sum(axis=1),label='Beta=0')


#plt.plot(d3.sum(axis=1),label='Beta>0')


#plt.plot(d4.sum(axis=1),label='Index*2.5')


#plt.plot(d5.sum(axis=1),label='Index*3')

plt.legend()
plt.show()
```

```
# Codes for Part2

# 包的加载
# Package loaded
# 这里有很多包其实在这个方法里是用不着的，比如我们仅仅用了 GBDT
import pandas as pd
import numpy as np
np.random.seed(10)
import matplotlib.pyplot as plt

from sklearn.datasets import make_classification
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.preprocessing import OneHotEncoder
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_curve
from sklearn.pipeline import make_pipeline     #做模型之间的管子链接


# Getdata_ML
# Data Preprocessing




S = 'C://Users/26529/Desktop/Data/2018_MCMProblemC_DATA/ML_Dataset'

s=['ACS_10_5YR_DP02_with_ann.csv','ACS_11_5YR_DP02_with_ann.csv','ACS_12_5YR_DP02_with_ann.csv','ACS_13_5YR_DP02_with_ann.csv','ACS_14_5YR_DP02_with_ann.csv','ACS_15_5YR_DP02_with_ann.csv','ACS_16_5YR_DP02_with_ann.csv']

def Get_Data1(filename,path):


    Whole_Data_set=dict()

    c = list(range(2010,2017))

    j=0
    #   Get Data
    for i in filename:

        Data_itself = pd.read_csv(path+'/'+i)
```

```
            Whole_Data_set[c[j]] = Data_itself

            j = j+1


      return Whole_Data_set


Whole_Data_set = Get_Data1(s,S)

# 下面是数据预处理的过程

# 删除缺失数据和不需要的数据，正则表达式 0.5h
# 将剩下的数据进行标准化处理    0.2h
# 将所在 county 的该年毒品数做连接 0.3h

def Data_drop(Whole_Data_set):

      # 在这里我们丢弃所有的含有(X)的变量

      for i in list(Whole_Data_set.keys()):


            #print(Whole_Data_set[i].columns)
            cols=[x for j,x in enumerate(Whole_Data_set[i].columns) if
Whole_Data_set[i].iat[3,j]=='(X)']
            Whole_Data_set[i]=Whole_Data_set[i].drop(cols,axis=1)
            Whole_Data_set[i]=Whole_Data_set[i].drop(['GEO.id'],axis=1)
            Whole_Data_set[i].drop(axis=0, index=0, inplace=True)    #删除第一行重复数
据


      # 接下来我们将通过交叉处理取得最小的数据共同交集;
      # 利用这个数据交叉集合

      A0=set(Whole_Data_set[2010].columns)
      A1=set(Whole_Data_set[2011].columns)
      A2=set(Whole_Data_set[2012].columns)
      A3=set(Whole_Data_set[2013].columns)
      A4=set(Whole_Data_set[2014].columns)
      A5=set(Whole_Data_set[2015].columns)
      A6=set(Whole_Data_set[2016].columns)
```

```
    A = list(A0.intersection(A1).intersection(A3).intersection(A4).intersection(A5).intersection(A6))

    for i in list(Whole_Data_set.keys()):


        Whole_Data_set[i]=Whole_Data_set[i][A]

    return Whole_Data_set

Whole_Data_set = Data_drop(Whole_Data_set)

# 将得到的 DataFrame 打上标签，我们先获取一下被解释变量数据

S1 = 'C://Users/26529/Desktop/Data/2018_MCMProblemC_DATA/MCM_NFLIS_Data.xlsx'

s1 = 'Data'

def Get_Data(file, sheetname):

    # 获取数据

    Data_itself = pd.read_excel(file, sheet_name=sheetname)
    return Data_itself

Data1 = Get_Data(S1,s1)

def Join_Data(Whole_Data_set,Data1):

    Data = Data1[['YYYY','FIPS_Combined','TotalDrugReportsCounty']]

    Data = Data.drop_duplicates()

    for i in list(Whole_Data_set.keys()):

        data = Data[Data['YYYY']==i]
        #print(data.shape)

        Whole_Data_set[i]['GEO.id2']=Whole_Data_set[i]['GEO.id2'].values.astype(int)
        Whole_Data_set[i] = Whole_Data_set[i].merge(data,left_on='GEO.id2',right_on='FIPS_Combined')
        print(i)

    return Whole_Data_set



Whole_Data_set = Join_Data(Whole_Data_set,Data1)

L = list(Whole_Data_set[2010].columns)
L.remove('YYYY')
L.remove('FIPS_Combined')

# 丢弃所有缺失值，正态化

from sklearn import preprocessing

def Dropspecialstr_columns(L,Whole_Data_set):

    c = 0

    for i in list(Whole_Data_set.keys()):

        if c==0:

            New_Df = Whole_Data_set[i][L]
            c = c+1

        else:
            AAA = Whole_Data_set[i][L]

            New_Df = pd.concat([New_Df,AAA])

    New_Df = New_Df.apply(pd.to_numeric, errors='coerce')

    New_Df = New_Df.dropna(axis=1)

    # 正态化


New_Df=pd.DataFrame(data=preprocessing.scale(New_Df),index=New_Df.index,columns=New_Df.columns)

    return New_Df
```

```
New_Df = Dropspecialstr_columns(L,Whole_Data_set)




# 用 GBDT 模型做特征处理和特征筛选并输出权重 0.5h
# 用该权重加权数据获得一个指标 0.5h
# 用该指标作为原模型的一个变量加入回归之中去，但是这个时候无法预测，因为没有未来的因子数据 1h
# 通过进行参数估计，对这个变量进行分解后得到每一部分的灵敏度，据此进行分析，并给出一些假设数据进行未来模拟 1h
# 写后面的所有回答和两页信纸   你们的工作
# 进行分析   你们的工作
# 修改 剩余时间




    ## 设置机器学习的参数，区分预测集和训练集
    ## Set parameters and load data

X =   New_Df.drop('TotalDrugReportsCounty',axis=1)
y = pd.DataFrame(New_Df.TotalDrugReportsCounty)

    #print(y.columns)
    #print(X.columns)
n_estimator = 5

    # 在这里我们实际上不需要做测试集和训练集的区分，因为本部分本来就是训练的部分
X_train = X
y_train = y

    # 需要将对 LR 和 GBDT 的训练集给区分开来
    # It is important to train the ensemble of trees on a different subset of the training data than the linear regression model to avoid overfitting, in particular if the total number of leaves is similar to the number of training samples
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)

    # Supervised transformation based on gradient boosted trees
    # 这里是训练好的模型，GBDT 模型，编码模型和逻辑回归模型
```

```
grd = GradientBoostingRegressor(n_estimators=10)

grd.fit(X_train, y_train)

Impact = list(list(grd.feature_importances_))

plt.figure()
plt.plot(Impact)
plt.title('Feature_importance')
plt.show()


y_pred_grd = list(grd.predict(X_test))

y_test.reset_index(inplace=True)

y_test=list(y_test.TotalDrugReportsCounty.values)
plt.figure()
plt.scatter(range(len(y_test)),y_test)

plt.scatter(range(len(y_pred_grd)),y_pred_grd)
plt.title('error analysis')
plt.grid()
plt.show()

RMSD = np.sqrt(np.square(np.array(y_pred_grd)-np.array(y_test)).sum()/len(y_test))

Iter =10
s_lo = list()
s_value = list()

a=Impact.copy()

for i in range(Iter):

    s_lo.append(a.index(max(a)))
    s_value.append(a[s_lo[i]])
    a[s_lo[i]] = 0

plt.figure()
plt.plot(range(10),s_value)
plt.grid()
plt.title('Feature selection')
plt.show()
```

```
# 开始着手处理 part2 参数

Feature = ['HC01_VC66', 'HC01_VC70', 'HC01_VC88', 'HC01_VC71']
def Extract_data(Whole_Data_set,Feature):

    a = dict()


    for i in list(Whole_Data_set.keys()):

        a[i] = Whole_Data_set[i][Feature]
        a[i] = a[i].apply(pd.to_numeric, errors='coerce')
        a[i] = a[i].dropna(axis=1)
        a[i] = pd.DataFrame(data=preprocessing.scale(a[i]),index=a[i].index,columns=a[i].columns)
        a[i]['GEO.id2']=Whole_Data_set[i]['GEO.id2']
    return a


a = Extract_data(Whole_Data_set,Feature)

Weight = s_value[0:4]

for i in list(a.keys()):
    a[i]['Index'] =Weight[0]*a[i].iloc[:,0]+Weight[1]*a[i].iloc[:,1]+Weight[2]*a[i].iloc[:,2]+Weight[3]*a[i].iloc[:,3]
    a[i] = a[i][['Index','GEO.id2']]
```