| For office use only | Team Control Number | For office use only |
|---|---|---|
| T1 _____ | **1920491** | F1 _____ |
| T2 _____ | | F2 _____ |
| T3 _____ | Problem Chosen | F3 _____ |
| T4 _____ | **C** | F4 _____ |

**2019**

**MCM/ICM**

**Summary Sheet**

# The Opioid Crisis

The United States is currently facing serious opioid abuse problems. In response to this situation, we used data science to establish a drug diffusion equation model to describe the current state of drug diffusion and spread and proposed relevant policy recommendations.

As to part one, the opioid spreading process is reflected in the number of cases. Through data preprocessing and visualization, we initially obtained the competitive relationship between heroin and synthetic opioids and established a competitive opioid diffusion equation model based on the reaction-diffusion equation and the Lotka-Volterra competition model. According to the finite difference method and the genetic algorithm, we can get the parameter items in the model. Then we traced back the possible starting position of opioid diffusion and predict the places where the number of cases in the next five years exceeds the threshold. In the end, based on the hierarchical method AGEMES, we get the threshold k which is be regarded as a classification of the number of cases in each region.

For the second part, we considered socioeconomic factors' effect with the US census data which consists of 600 socio-economic characteristics. We used the XGBoost algorithm to rank the correlations between these characteristics and opioid abuse cases, and obtained a top ten rankings of choice correlations to describe the effect. For the model improvement, We integrated the data provided by NFLIS into the characteristics of XGBOOST. The final corrected XGBoost model R2 is 99% on the training set and 78% high-precision model on the test set. For the model in part one, we added the top ten socio-economic factors that we have found. Then, based on the relevant economic and social factors, combined with the national conditions and reality of the United States, we found out the reasons for the widespread use of opioids.

As for the third part, we combined the existing US special management drug testing system and the implementation intervention policy. According to the partial revised model, we proposed the intervention target characteristic testing policy to ensure that the present status is improved as much as possible in the drug abuse population. The characteristic parameters after the intervention are adjusted and substituted into the test model, reflecting the effectiveness of the policy.

**Key Words:** reaction-diffusion equation, Lotka-Volterra competition model, GA， FDM, XGBoost

# Contant

# I. Introduction

## 1.1 Background

Many people in the United States are taking synthetic and non-synthetic opioids, either for medical or recreational purposes. People depending on opioids might have difficulty filling in positions with specific requirements. What's more, the increase of opioid addition cases among the elderly results in more cost of health care. So, it is necessary to control the spread of the opioids and identify a reasonable strategy to counter the national crisis.

## 1.2 Our works

Task1:

Describe the spread and characteristics of the reported synthetic opioid and heroin incidents in and between the five states and their counties over time
Identify any possible locations where specific opioid use might have started in each of the five states
Predict when and where synthetic opioid and heroin abuse will occur

Task2:

Find the trend in the numbers of synthetic and heroin cases
Compare the above numbers
Identify drug threshold levels for synthetic and non-synthetic opioids of each county

Task3:

Determine whether socioeconomic factors have an impact on opioid use
Determine which socioeconomic factors have a significant impact on opioids use

# II. Assumptions

We make the following assumptions to help us perform modeling. These assumptions are the premise for our subsequent analysis.

(1) We ignore state-state boundaries when predicting the spread of opioid abuse.
(2) We assume that the number of reported cases indicates whether there is abuse of opioids.
(3) For heroin and synthetic opioids, they all have geographical origins. We can not only predict their spreading status through the model but also predict these origins.

# III. Data collection and preprocessing

## 3.1 Sum

Due to the different nature of heroin and synthetic opioids, we calculated the sum of the numbers of synthetic opioid cases per county per year. We selected the number of each county each year as the density of the diffusion equation to study the diffusion process of their quantity in time and space.

## 3.2 Interpolation and extraction

However, the diffusion equation needs to know the geographical coordinate information of each county in order to carry out spatial research. We extracted the latitude and longitude information of each county from the United States Board on Geographic Names [1] according to the regional FIPS codes. Then we used Arcgis to display the points on the map. Their colors represent the number of cases of heroin or opioids. The redder the color, the larger the number. Fig1 shows synthetic opioid cases in 2017.

Finally, we used Kriging interpolation [2] for its lower error, the interpolation function is spherical, and the interpolation range is $89.5^{\circ}W-74.5^{\circ}W, 36.5^{\circ}N-42^{\circ}N$. Fig.2 shows the results of Kriging interpolation. In the interpolation region, we extracted the data at intervals of 0.1 degrees to prepare for the analysis of the diffusion equation.

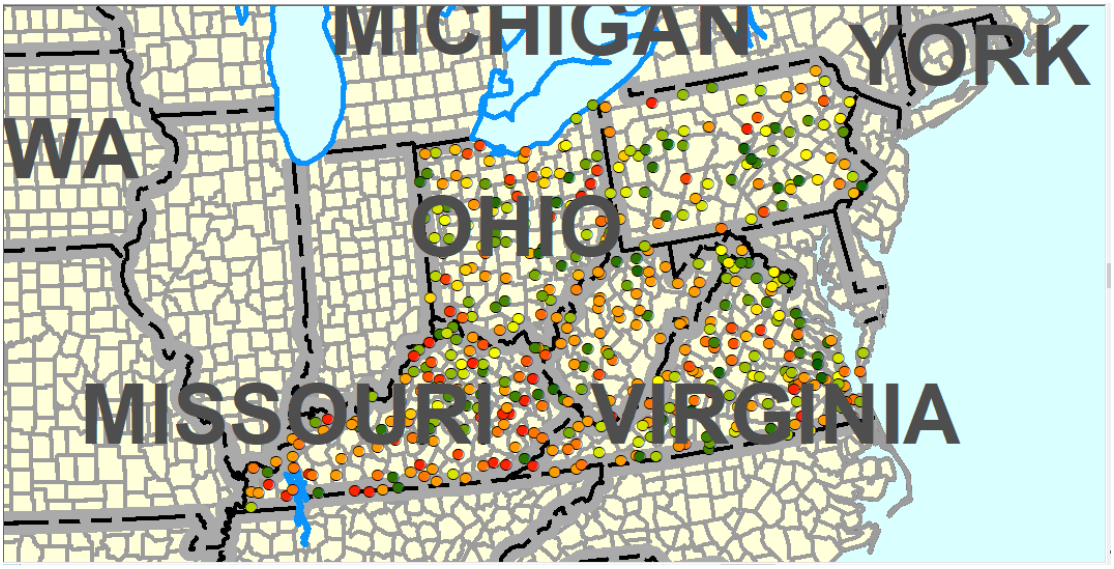



Fig1：Display points

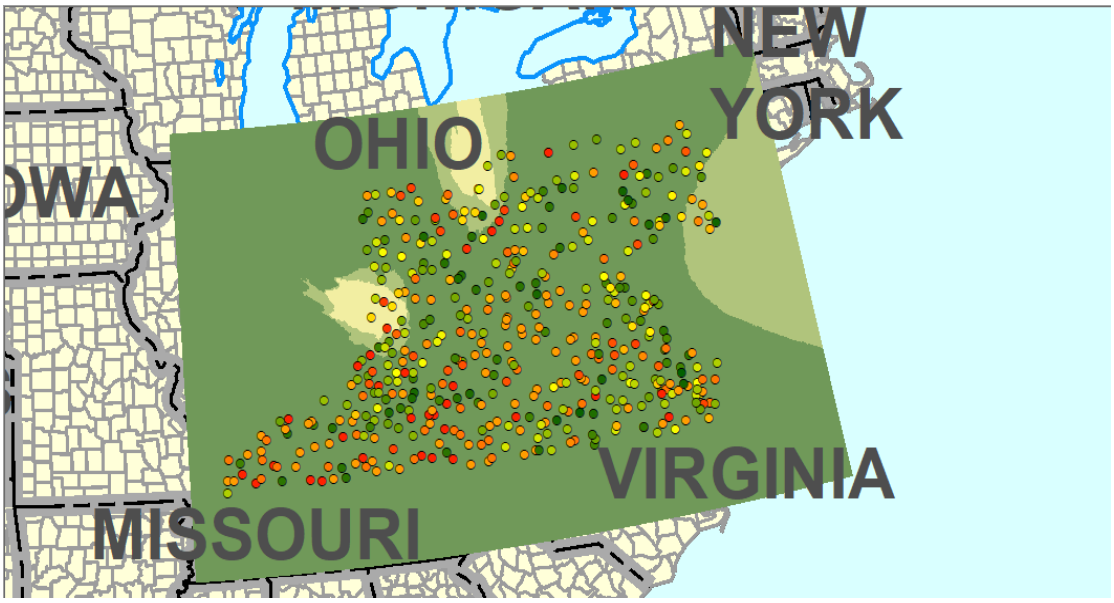



Fig2: Interpolation

# IV. Task1: Problem Analysis

## 4.1 Part one

To describe the spread characteristics of synthetic opioid and heroin cases in five states, it is necessary to identify the statistics that can describe the changes and characteristics of cases in time and space. When identifying the statistics, we should not consider the influence of socio-economic factors.

According to the data given, we divided the opioids into two types roughly: synthetic opioids and heroin. We carried out data preprocessing and visualization to observe and analyze the laws and characteristics of the number of opioids cases. We can see that the spread of opioid cases is similar to the process of infectious disease spread. In addition, it is necessary to consider the interaction between the two processes, and the interaction between two states and the boundaries.

The location where a particular opioid may start should have the following characteristics: the number of opioid cases in the area before the start time is 0; the number of drug cases in the adjacent area before the start time is 0; after the start time, the number of cases increased, and as time grew, the number of opioid cases in adjacent areas begun to increase. After the model is established, we can perform a time traversal to find a region that satisfies the three characteristics, that is, a position where opioid use may begin.

To find out the threshold that the government pays attention to, we should first analyze the process of spread. The threshold setting reflects the change of the spread. We can observe the number of cases in a particular state. We can divide the number of states into two states: safety and danger. The dangerous number needs to be vigilant to the government and has a strong ability to spread to the surrounding; while the security status indicates that the number of cases is within the normal range and the diffusion state tends to be stable and is not easy to spread around. According to the characteristics of the two states, we can use clustering analysis to divide the number of cases into two categories, and the inter-class step value is the threshold.

## 4.2 Part two

Firstly, based on the analysis of part one, we can establish a model describing the spread of synthetic opioid and heroin cases. With the socio-economic data provided by the census, we can use this model to consider the socio-economic factors. The data gives eighteen categories of population classification criteria, describing population characteristics from different perspectives, such as educational attainment, language spoken at home, and place of birth. In each major category, we classified and counted the characteristics of the population.

Then, we need to analyze whether these characteristics are related to opioid abuse cases, what kind of correlations, and how relevant are they, to describe the characteristics of people who abuse opioids, that is, who is using and abusing opioids. At the same time, the correlation coefficients determine how socio-economic factors affect the spread of drugs.

Finally, according to the results obtained, we need to add new related items to the model established in part one, and the model is modified so that the spreading process can be further described through related items. Based on the relevant socio-economic factors, combined with the national conditions and reality of the United States, we can find out the reasons for the widespread use of opioids.

## 4.3 Part three

As for the third part, we combined the existing US special management drug testing system and the implementation intervention policy. According to the partial revised model, we proposed the intervention target characteristic testing policy to ensure that the present status is improved as much as possible in the drug abuse population. The characteristic parameters after the intervention are adjusted and substituted into the test model, reflecting the effectiveness of the policy

# V. Task1: Description, Identification and Prediction

## 5.1 Model 1: Diffusion Equation

The spread of synthetic opioids and heroin exhibits many features of epidemics, such as infectiousness, rapid diffusion, and clear geographic boundaries. Thus, mathematical modeling methods can be applied to describe such problems. Based on the theory of heroin epidemic models, considering the influence of geographical location, the spread of opioids can be regarded as a diffusion process, and a diffusion equation model can be established.

The process of opioid dissemination is reflected in the number of cases. The high number of opioid cases identified indicates that opioid abuse is serious. Conversely, a small number of cases indicates that opioid abuse is relatively light.

The general form of the diffusion equation:

$$\frac{\partial u}{\partial t} = a^2 \Delta u$$

Where:

$u(x, y, t)$ : The concentration of matter

$a^2$ : The diffusion coefficient

Comparing with the physical diffusion equation, we establish the opioid diffusion equation:

$$\frac{\partial u}{\partial t} = d_1 \Delta u + u(a_1 - b_1 u)$$

Where:

$u(x, y, t)$ : The case density

$d_1$ : The diffusion rate of opioids

$a_1$ : Natural growth rate

$b_1$ : The competition coefficient within its type

## 5.2 Model 2: The Logistic Equation and Lotka-Volterra Competition Model

According to Fig1, heroin and synthetic opioids show a competitive relationship between themselves, similar to the competition between the two species. The above drug diffusion equation needs to be modified to fit Lotka-Volterra competition model.

### 5.2.1 The Logistic Equation

Ecologically, subject to limited resources, population growth is often limited and has a maximum carrying capacity. Ecologists make two assumptions: 1) An environmental capacity exists 2) The growth rate of the population decreases linearly with increasing density. Therefore, the logistic equation was put forward:

$$\frac{dx}{dt} = rx\left(1-\frac{x}{K}\right)$$

Where:

$x$ : the size of the population at a given time

$r$ : inherent per-capita growth rate

$K$ : the carrying capacity

### 5.2.2 Lotka-Volterra Competition Model

In nature, a species often does not exist in isolation, and it often has one or the other relationship with the surrounding species. The competitive Lotka–Volterra equations are a simple model of the population dynamics of species competing for some common resource.

Given two populations, x1 and x2, with logistic dynamics, the Lotka–Volterra formulation adds an additional term to account for the species' interactions. Thus the competitive Lotka–Volterra equations are:

$$\frac{dx_1}{dt} = r_1x_1\left(1-\left(\frac{x_1+\alpha_{12}x_2}{K_1}\right)\right)$$

$$\frac{dx_2}{dt} = r_2x_2\left(1-\left(\frac{x_2+\alpha_{21}x_1}{K_2}\right)\right)$$

Where:

$\alpha_{12}$ : The effect species 2 has on the population of species 1

$\alpha_{21}$ : The effect species 1 has on the population of species 2

In this case, it can be further developed into:

$$u_t = d_1\Delta u + u\left(a_1 - b_1u - c_1v\right) \quad x\in\Omega,\ t>0$$

$$v_t = d_2\Delta v + v\left(a_2 - b_2u - c_2v\right) \quad x\in\Omega,\ t>0$$

Where:

$\Delta$ : $\Delta=\sum_{i=1}^{N}\frac{\partial^2}{\partial x_i^2}$ Laplace operator

$\Omega$ : Smooth bounded regions on a two-dimensional plane

$u(x,y,t), v(x,y,t)$ : Case density of heroin and synthetic opioids

$d_1, d_2$ : Diffusion rates of $u$ and $v$

$a_1, a_2$ : Natural growth rates of $u$ and $v$

$b_1, b_2$ : Competition coefficients within the type of $u$ and $v$

$c_1, c_2$ : Competition coefficient between types of $u$ and $v$

$$a_1>0, a_2>0, b_1>0, b_2>0, c_1>0, c_2>0, d_1>0, d_2>0$$

It can be seen from the map that the five states are geographically adjacent. To avoid the influence of state-state interdiffusion, we consider the five states as a holistic model. Since we consider the occurrence of cases in these five states, the speed and state of opioid diffusion are not affected by geographical state boundaries, so these partial differential equations are the Cauchy problem (unbounded problem). The initial conditions are the number of heroin and synthetic opioid cases occurring in five states in 2010.

In the above model, competition can be divided into weak competition and strong competition.

Weak competition:

$$\frac{b_1}{b_2}>\frac{a_1}{a_2}>\frac{c_1}{c_2}$$

Srong competition:

$$\frac{b_1}{b_2}<\frac{a_1}{a_2}<\frac{c_1}{c_2}$$

In the case of weak competition, there is no eviction in the competition between the two, that is, two types can coexist.

As to strong competition, there is an expulsion phenomenon in the competition between the two, that is, one type disappears completely and the other type survives.

### 5.2.3 Results & analysis

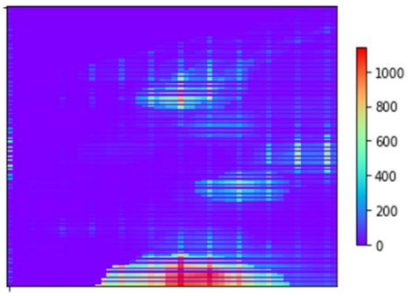
We calculated the parameters by using the genetic algorithm(GA) and finite difference method it is showed in table1:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| error | 16218.66 | | | | | | |
| d1 | 675.9507 | a1 | 803.0308 | b1 | 864.2577 | c1 | 686.8703 |
| d2 | 973.1102 | a2 | 570.5362 | b2 | 964.1142 | c2 | 627.0014 |

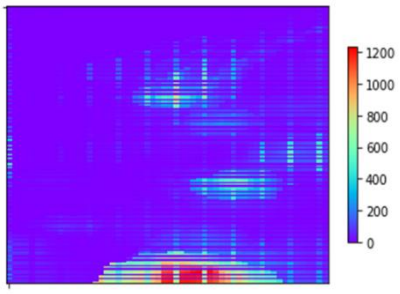
Table1.

Here is the thermal maps of heroin and synthetic opioids.

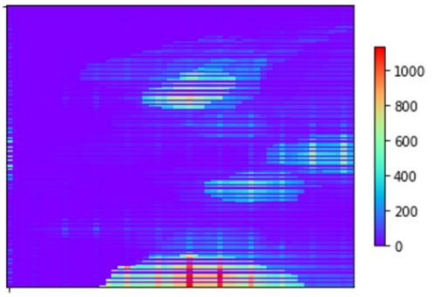
Heroin:2010



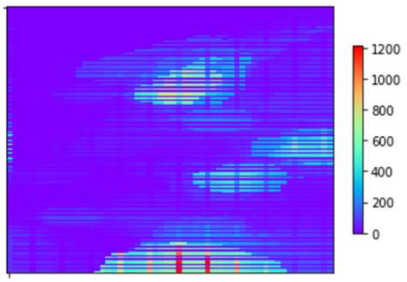

2011



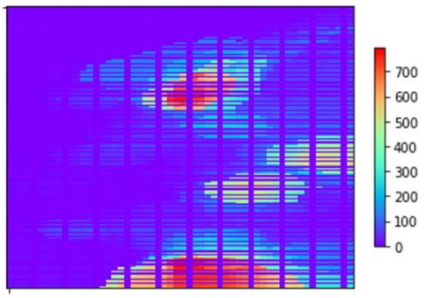

2012



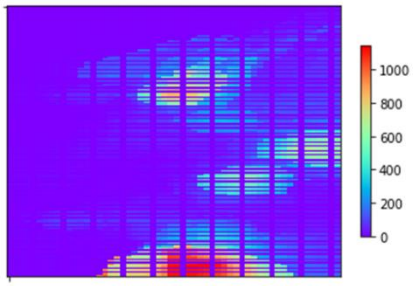

2013



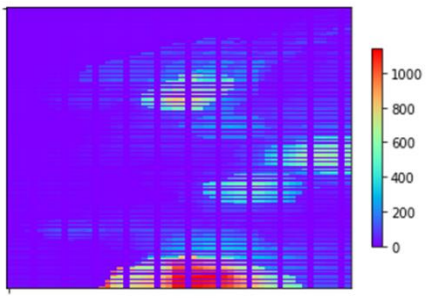

2014



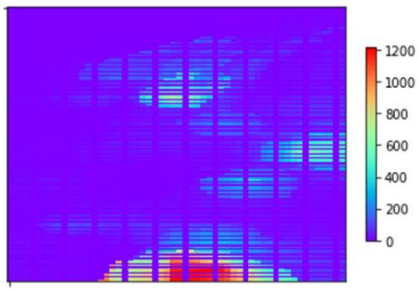

2015




2016

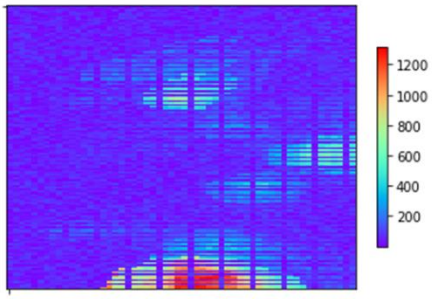



2017

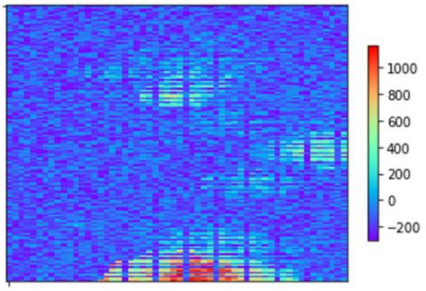



2018

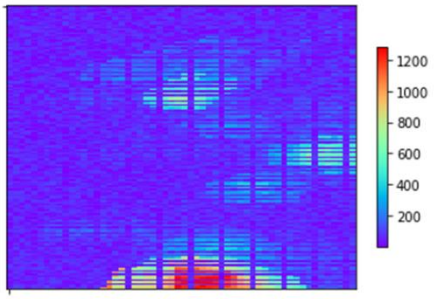
2019



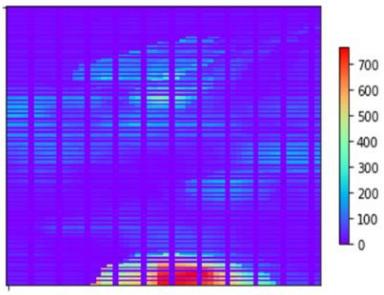

2020



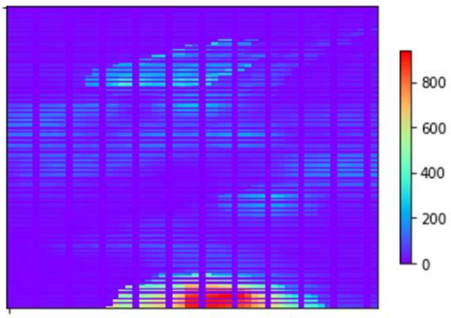

Synthetic opioids:2010



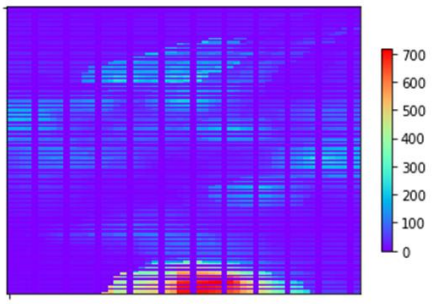

2011



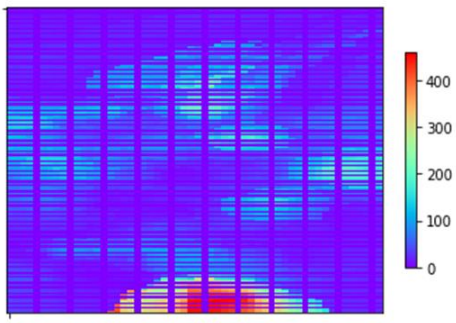

2012



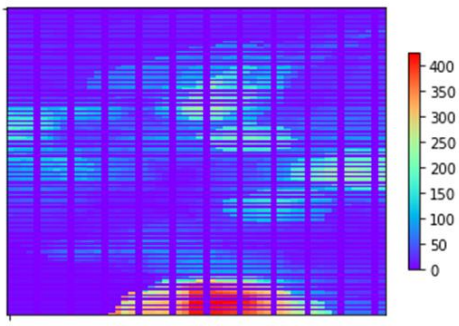

2013




2014




2015




2016

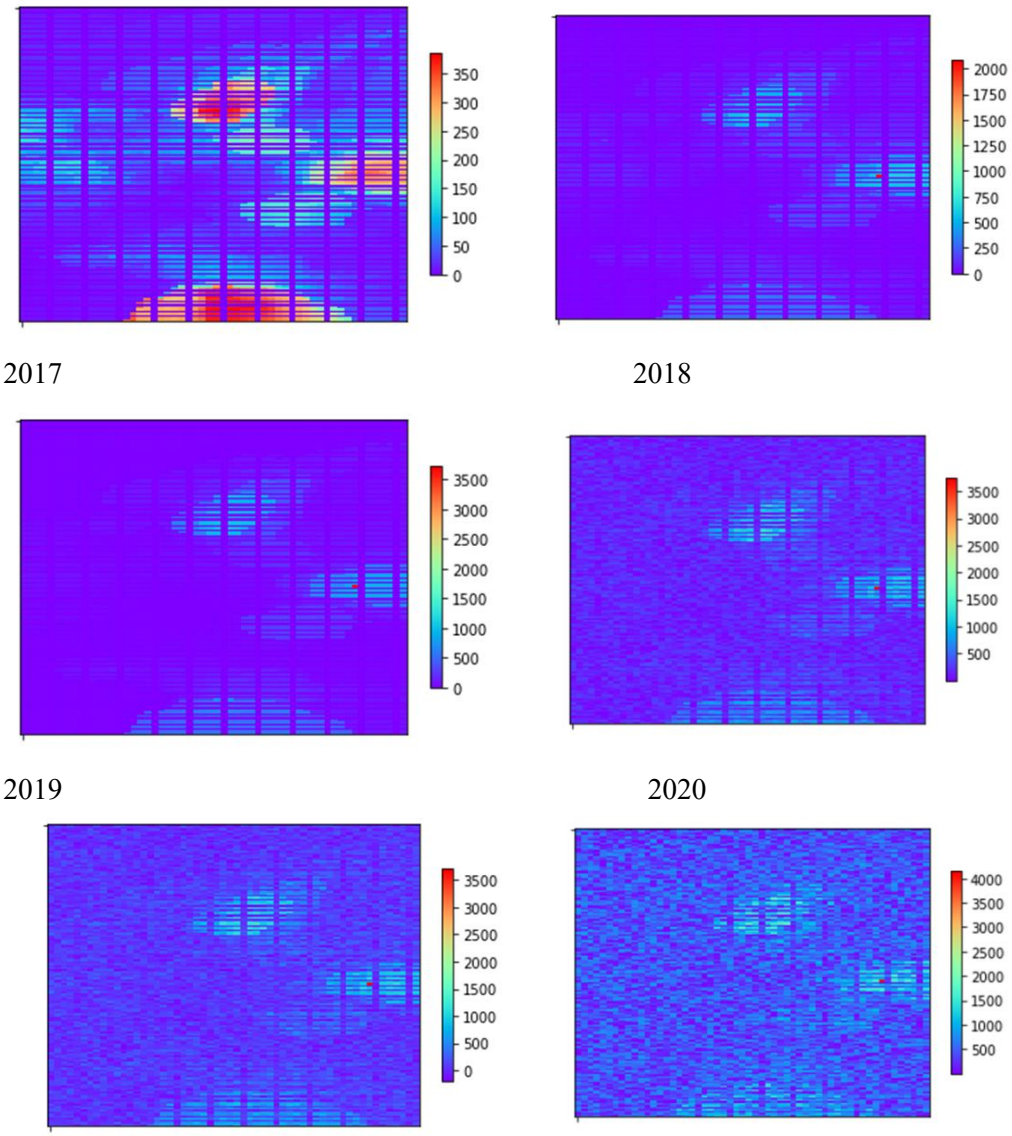



Fig3. Thermal maps of opioid use

From the thermal maps, we can conclude that at the beginning, the spread of heroin and synthetic opioids use was very similar., they have the similar original beginner As time went by, it can be predicted that the heroin use gradually stabilized into some specific regions whereas synthetic opioid spread all over the place.

# VI. Task2: Finding the thresholds

## 6.1 Comparison

From Fig3,Fig4, we can clearly see a competitive relationship between heroin and synthetic opioids in the process of dissemination and diffusion. They fluctuated between 2010 and 2017,

but the sum of the two shows a linear growth trend. This means that the problem of abuse is generally deteriorating over time.

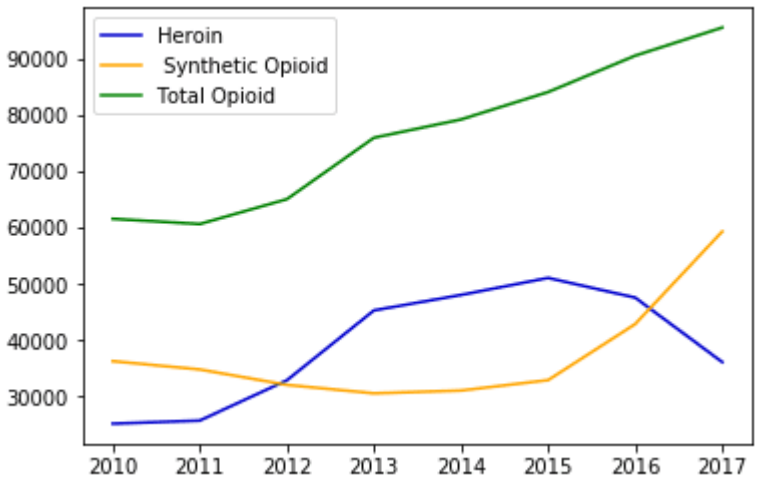



Fig4: Total cases of heroin, synthetic opioids, and their sum

## 6.2 Model: K-means Clustering Analysis

Now we observe the situation of various opioid abuse cases in five states, which are also divided into heroin and synthetic opioids. We calculate the total number of opioid abuse cases at the same time.

The goal is to find out the threshold value $k$. It can be regarded as the classification of the number of cases in each region. If the number of cases is less than $k$, then the use of opioids in the region is normal. Otherwise, governments should pay attention to opioid use in their own regions. Therefore, we can cluster counties in the five states according to the number of cases.

Take heroin as an example, we carried out the normal distribution test of the number of heroin cases in each region in each year. We found that the number of cases in each region did not obey the normal distribution, and the data was concentrated when they are lower than 200 and discrete higher than 200. The degree of dispersion has a great influence on the clustering result, so we considered clustering the data into multiple classes through the k-means algorithm.

### 6.2.1 Modeling step

**Step 1:Initialization**

Given the number of classification and set $j=0$. Pick $m$ vectors from sample vectors as $k_1^j, k_2^j, ..., k_m^j$ .then consider them as the clustering center $k_i^j = \left[k_{i1}^j, k_{i2}^j, ..., k_{im}^j\right], (i=1,2,...,k)$.

**Step 2:Classification of samples**

Include each sample vector $x_l = \left[x_{l1}, x_{l2}, ..., x_{\ln}\right]^T$ in the class which owns $\boldsymbol{k}_i^j$ as the center by this formula:

$$\left\|x_l - k_i^j\right\| = \min_{1\le h\le m}\left\|x_l - k_m^j\right\|$$

**Step 3:Center adjustment**

Adjust the clustering center by:

$$K_{ih}^{j+1} = \frac{\sum_{x_{lih} \in k_i^j} x_{lih}}{N_i}$$

Where $N_i$ represents the number of vectors of the clustering block $K_i^j$ .

**Step 4:Condition judgment**

Construct the objective function iteration as:

$$J = \sum_{m=1}^{n} \sum_{x_m \in K_i} |x_k - k_i|$$

Put the data which got from Step 1 into the formula and judge the result. The iteration ends when the result doesn't have obvious change. Or $j = j+1$ and turn to Step 1.

### 6.2.2 Results & analysis

Considering the eight years of 2010-2017 as the dimensions of data, it describes the characteristics of the number of opioid abuse cases at the time level. We use the L2-norm as the distance between the data and set the number of classes to 4. Data of the former two dimensions is used as the horizontal and vertical coordinates to observe the clustering results.

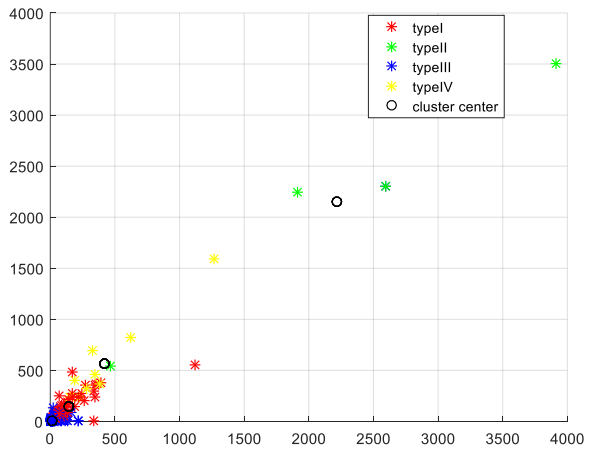



Fig5: Cluster results

Based on the hierarchical method AGEMES, we condensed and merged multiple categories until they are divided into two categories. Then, we detected the cluster boundary. Because the data is discrete, the average of the data at the boundaries of two types can be obtained. Since the data dimension is 8, the average of the eight boundaries is averaged again as the inter-class threshold. In this way, we calculated thresholds $k_1$ and $k_2$ for heroin and synthetic opioids respectively.

After calculating the threshold $k_3$ for the total number of heroin and synthetic opioids, we compare $k_3$ with $k_1 + k_2$. For the sake of reality and avoidance of misjudgment, the threshold of the total number is $k = \min\{k_3, k_1 + k_2\}$. With the threshold, governments can monitor possible opioid abuse as much as possible.

When they are detecting the number of opioid abuse cases in a certain area, if any of the data( $k_1, k_2, k$ ) reaches or exceeds the threshold, it indicates that opioid abuse occurs at this time, and the government should pay attention to it.

| year | h | s | sum | h+s | min |
|---|---|---|---|---|---|
| 2010 | 75 | 109 | 169 | 184 | 169 |
| 2011 | 71 | 98 | 216 | 169 | 169 |
| 2012 | 81 | 76 | 177 | 157 | 157 |
| 2013 | 105 | 58 | 169 | 163 | 163 |
| 2014 | 144 | 71 | 212 | 215 | 212 |
| 2015 | 184 | 77 | 193 | 261 | 193 |
| 2016 | 205 | 95 | 305 | 300 | 300 |
| 2017 | 221 | 160 | 446 | 381 | 381 |
| average | 135.75 | 93 | 235.875 | 228.75 | 218 |
| final | 135 | 93 | 218 | | |

Table 2. K-means

After K-means clustering analysis, we get the results of Table1. From the Table we can see that the final thresholds are(135, 93, 218). If any of the data reaches or exceeds the threshold, the government should pay more attention to them.

# VII. Task3: Determining the Factors

## 7.1 Model: XGBoost

Tree boosting is a highly effective and widely used machine learning method. In this paper, we describe a scalable end-to-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges. There are many factors in the social economics data given. We use XGBoost to determine the weights of the factors. This can help the government pre-monitor specific populations, and prevent the abuse of opioids at the lowest possible cost.

### 7.1.1 Modeling

**Step1: Output**

For a given data set with $n$ examples and $m$ features, a tree ensemble model uses $K$

additive functions to predict the output.

Data set:

$$D=\{(x_i,y_i)\}\left(|D|=n,x_i\in\mathbb{R}^m,y_i\in\mathbb{R}\right)$$

Output:

$$\hat{y}_i=\phi(x_i)=\sum_{k=1}^{K}f_k(x_i),\ f_k\in F$$

Where:

$F=\left\{f(x)=\omega_{q(x)}\right\}\left(q:\mathbb{R}^m\to T,\omega\in\mathbb{R}^T\right)$: The space of regression trees

$q$: The structure of each tree that maps an example to the corresponding leaf index

$T$: The number of leaves in the tree

$\omega$: Leaf weights

$f_k$: Correspond to $q$ and $\omega$

**Step2: Regularized Learning Objective**

$$L(\phi)=\sum_i l\left(\hat{y}_i,y_i\right)+\sum_k\Omega(f_k)$$

Where $\Omega(f)=\gamma T+\frac{1}{2}\lambda\|\omega\|^2$

$l$: Dfferentiable convex loss function

$\Omega$: Complexity penalty

$\lambda$, $\gamma$: Coefficients

Our task is to minimize the above-mentioned regularized objective, which is divided into two parts. The former is a loss function, such as residual sum, and the latter is a complexity penalty, which includes the number of leaves and each weight.

**Step3: Gradient Tree Boosting**

Let $\hat{y}_i^{\,t}$ be the prediction of the $i$-th instance at the $t$-th iteration, we will need to add $f_t$ to minimize the following objective.

$$L^{(t)}=\sum_{i=1}^{n}l\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)+\Omega(f_t)$$

We greedily add the $f_t$ that most improves our model. Second-order approximation can be used to quickly optimize the objective in the general setting.

$$L^{(t)}\simeq\sum_{i=1}^{n}\left[l\left(y_i,\hat{y}^{(t-1)}\right)+g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t)$$

Where:

$$g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$$

They are first and second order gradient statistics on the loss function. Next, we remove the constant terms to simplify the objective.

$$\tilde{L}^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t)$$

It can be rewritten into:

$$\begin{aligned}\tilde{L}^{(t)} &= \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)\omega_j^2\right] + \gamma T\end{aligned}$$

Where:

$I_j = \{i \mid q(x_i) = j\}$: The instance set of leaf j

By deriving, we can find the optimal weight:

$$\omega_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}$$

Then calculate the optimal value:

$$\tilde{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T$$

We can use this equation to describe the structural quality of the tree. The more scores, the worse the quality.

We can also use this equation to decide if we should split at a certain node. Let $I_L$ and $I_R$ be the instance sets of left and right nodes after the split. Letting $I = I_L \cup I_R$, then the loss reduction after the split is given by

$$L_S = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma$$

### 7.1.2 Results & Analysis

Through XGBoost model, we calculated that $r^2$ is 0.68 and identified the top ten influencing

factors of F-Score as follows.

| | |
|---|---|
| HC01_VC03 | 0.0650724 |
| HC02_VC03 | 0.0216142 |
| HC01_VC15 | 0.0213842 |
| HC01_VC40 | 0.0197747 |
| HC03_VC120 | 0.011267 |
| HC02_VC193 | 0.0108071 |
| HC02_VC04 | 0.0103472 |
| HC02_VC107 | 0.0103472 |
| HC01_VC67 | 0.0101173 |
| HC02_VC147 | 0.0101173 |

Table 3

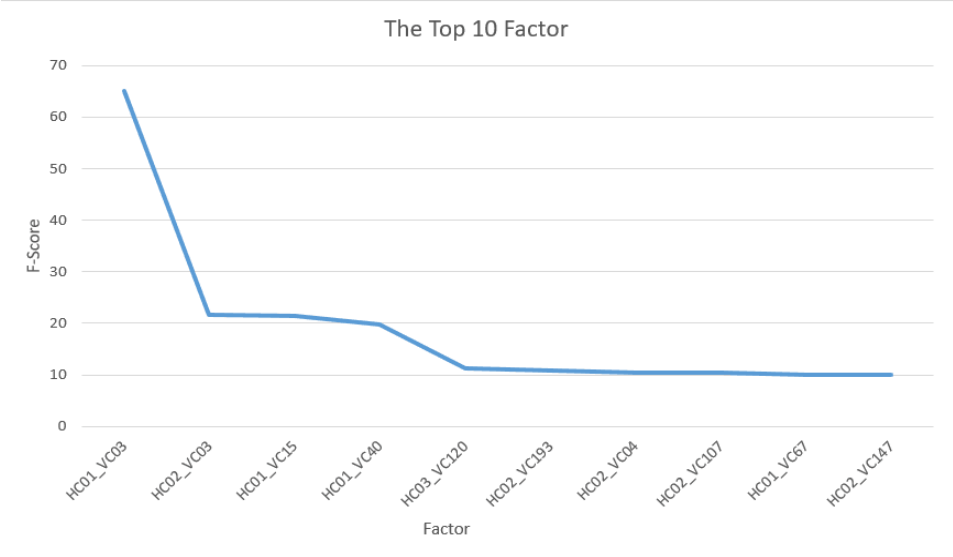



Fig 6. Top 10 factors in XGBoost

By doing this what we can find is that the most important factors in the social-economic factors. Here we will give a short discussion about some factors we find to understand the real mechanism between the factors and spread of heroin or synthetic opioids

HC02_VC107: Estimate Margin of Error; DISABILITY STATUS OF THE CIVILIAN NONINSTITUTIONALIZED POPULATION. That is an quite simple relationship between the HC02_VC107 and the abuse of heroin and synthetic opioids. People who is disabled is more likely to enjoy the pleasant which is brought by some special synthetic opioids.

HC01_VC40: Estimate; MARITAL STATUS – Divorced. Felling quite sorry the people who is unfortunately like people in HC02_VC107 and HC01_VC40. Divorced makes people fell blue, they may try to escape the truth by addicting in the heroin and synthetic opioids.

So negative factor as the HC02_VC107 and HC01_VC40, there are also positive factor between the 10 factors we found. HC01_VC67: Estimate; GRANDPARENTS - Responsible for grandchildren - Years responsible for grandchildren - 5 or more years. Just image you have an happy family with lovely grandson, will you try to addict with the heroin and synthetic opioids? Haha the answer must be NOT I guess.

In conclusions, in this part we have 10 factors which is the most important factors in social-economics conditions which may be related to the spread of heroin or synthetic opioids. And by analysis we find that all the factors is meaningful in the mechanism.

In next part, we will try to improve the model we build including the XGBoost model and the multi diffusion- Lotka-Volterra Competition Model.

# VIII. Task4: Model Improvement

## 8.1 Improvement for the Diffusion Equation Model

### 8.1.1 the ideal for improvement

For the model in part one, we added the top ten socio-economic factors that we have found using the XGBoost algorithm to the model, that is, we added the influence of socio-economic factors on the diffusion process to make corrections to the model.

### 8.2.2 the ideal for improvement

So here we have the new model as follow:

$$u_t = d_1\Delta u + u\left(a_1 - b_1 u - c_1 v\right) + \sum_{i=1}^{10} e_i X_i \quad x \in \Omega,\ t > 0$$

$$v_t = d_2\Delta v + v\left(a_2 - b_2 u - c_2 v\right) + \sum_{i=1}^{10} h_i X_i \quad x \in \Omega,\ t > 0$$

Where:

$\Delta$ : $\Delta = \sum_{i=1}^{N} \frac{\partial^2}{\partial x_i^{\,2}}$ Laplace operator

$\Omega$ : Smooth bounded regions on a two-dimensional plane

$u(x, y, t), v(x, y, t)$ : Case density of heroin and synthetic opioids

$d_1, d_2$ : Diffusion rates of $u$ and $v$

$a_1, a_2$ : Natural growth rates of $u$ and $v$

$b_1, b_2$ : Competition coefficients within the type of $u$ and $v$

$c_1, c_2$ : Competition coefficient between types of $u$ and $v$

$e_i$ : The weights of the $X_i$ for u

$h_i$ : The weights of the $X_i$ for v

$X_i$ : The top ten factors we found

## 8.2 Improvement for the XGBoost Model

### 8.2.1 the ideal for improvement

By the restriction of problem, when we construct the XGBoost Model in last chapter, we only use the U.S. Census socio-economic data. But the data from NFLIS is ignored. I think the data NFLIS provided is also important in for the spread of heroin or synthetic opioids.

So here is our ideal for the XGBoost Model improvement. We will construct a new XGBoost model to estimate the relationship between the case number and the data NFLIS provided plus the top 10 factors we find in the last part.

### 8.2.2 the improvement results

Here are the result of our new model:

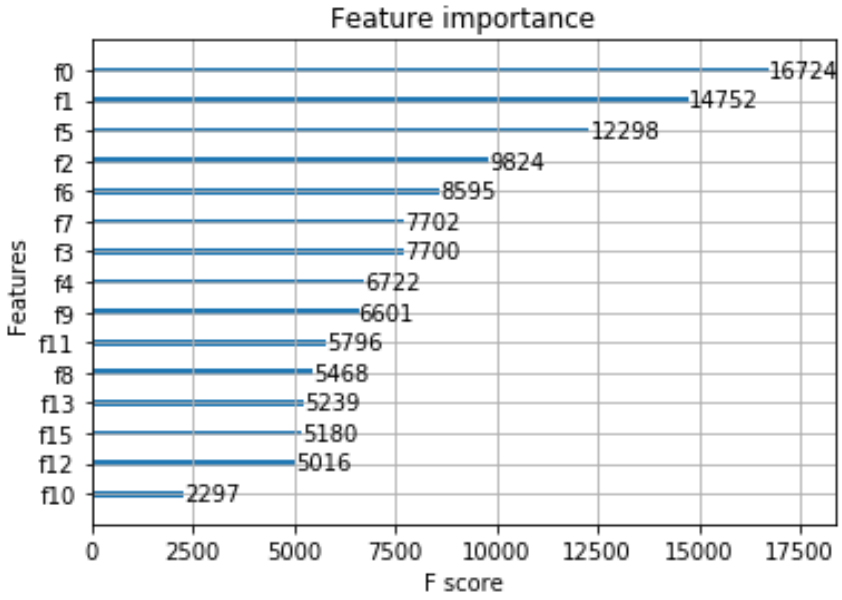



Fig 7. Top new factors for the XGBoost Model we improved

We have the new 16 factors from the NFLIS and the top 10 factors we find in the last chapter to describe the relationship: The Time, the NIPS code, Lat, Lon, TotalDrugReportsCounty, TotalDrugReportsState, and the top 10 factors without HC01_VC67 and HC02_VC015.

The results is also reasonable: the top factors is the time, the second factors is heroin or synthetic opioids, while the third factors is TotalDrugReportsCounty. The similarity result with the conclusion we get from the hot point. And the second factors is also an prove of the competition we get.

And here I want to emphasize another problem of this new XGBoost model is the high accuracy, not just in training data but also in the test data. The $r^2$ in the training data is 0.99 while 0.87 in the test data. Quite impressive!

Then the tree we constructed can be seen in the **appendix**.

# IX. Task5 Policy Control

The existing monitoring system for special management drugs in the United States has three aspects: first is the drug classification system, second is the monitoring system, and the last is the intervention target system. Special management drugs in the United States today are divided into five levels, and each level has corresponding regulatory measures, which are relatively complete. Combined with the regional analysis of the number of opioid abuse cases and the analysis of relevant socio-economic factors, we can propose improvement measures in the intervention and prevention system.

Statistics are carried out on areas where the number of cases exceeds the thresholds and are prone to become a source of spread. The government should increase supervision and intervention to prevent them from becoming the catalyst in the process of opioid spread.

For those who have the relevant characteristics of part two, especially those who have a history of drug abuse, drug purchase registration system should be implemented to reduce the probability of drug abuse. If such methods are implemented, the correlation between the specific population and the number of drug abuse cases can be reduced, that is, the coefficient of this item can be modified, and the modified model can be reapplied to observe whether the diffusion of the predicted case is reduced.

It is necessary to revise the prescription drug detection strategy. Doctors ought to use the drug rationally, understand patient's medication, reduce the risk of drug abuse. In this way, we can control drug abuse from the source.

# X. Task6: Memo

From: Team 1920491, MCM 2019
To: The Chief Administrator
Date: January 28, 2019
Subject: Suggestions to curb the opioids spread

Dear chief administrator, we are honored to inform you our achievement after performing data analysis and modeling.

First, we analyzed the data you provided to us and firstly had an intuitive feel for them. In

terms of total volume, the use of heroin and synthetic opioids is increasing, and there is a competitive relationship between the two. This also indicates that the abuse of opioids in these five states is getting worse.

Then we used a diffusion model to predict the spread of opioids. We predicted that cases of heroin will be geographically aggregated over time. Cases of synthetic opioids, however, will be geographically dispersed.

Therefore, it may be more difficult for the government to control synthetic opioids.

On top of that, we also identified some socioeconomic factors that will worsen the situation.

According to the requirements and the characteristics of opioids, we offer you the following suggestions:

- For areas with a large number of cases and areas that are about to become sources, the government should strengthen supervision and management.
- Pay special attention to the synthetic opioids, invest more financial resources, material resources, manpower, and prevent the spread of synthetic opioids.
- It is necessary to implement a special drug purchase registration system and strictly supervise people with a history of drug abuse.
- Pay attention to the higher ranked people in part two.
- Improve the prescription drug prescribing system, strengthen the management of doctors, and control the drug abuse from the source.

We sincerely hope that you can control opioid abuse as soon as possible!
Please contact us if you have any problems.

# XI. Evaluation of Our Models

## 8.1 Strengths of Our Model

- **Data supplement.** Using the method of interpolation, we can supply the data where is vacant originally.
- **Data preprocessing.** The data processing is very important when we are faced with numerous data. Through this step, we greatly improve the quality of the data. Thus, it is more efficient and convenient for us to solve the problem.

- **Accuracy.** The opioid diffusion model is inspired by the infectious disease model. Combined with the diffusion equation, it accurately describes the competitive relationship between heroin and synthetic opioids. What's more, we joined the influence of socio-economic factors, so the model is closer to reality, and it's versatility and popularization are stronger. What's more, the XGBoost model we improved is really powerful not just in the training data, but also test data.
- **Simplicity.** The model can simultaneously describe the changes in the time and space of the opioid case diffusion process. The forward or backward difference can be used to complete the backtracking and prediction through the finite difference method. It is powerful and harmonious.

## 8.2 Weaknesses of Our Model

- **Complexity.** The algorithm is highly complex and takes a long time.
- **Limitation.** Threshold determination is based on previous year data, not dynamically determined. We should adjust thresholds according to specific circumstances. Therefore, our model exists limitation to some degree.
- **Local error.** There are many modified model parameters. Although we can predict cases, it is inevitable that local extreme conditions will occur when forecasting.

# References

[1] United States Board on Geographic Names. https://geonames.usgs.gov/domestic/download_data.htm.

[2] JIN Guo-dong, LIU Yan-cong, NIU Wen-jie. Comparison between Inverse Distance Weighting Method and Kriging. *Journal of Changchun University of Technology*, 1006-2939(2003)03-0053-05.

[3] Lili Liu, Xianning Liu. Mathematical Analysis for an Age-Structured Heroin Epidemic Model. *Springer Nature B.V. 2019,* 34D23 · 34K20 · 92D30.

[4] https://en.wikipedia.org/wiki/Logistic_equation.

[5] https://en.wikipedia.org/wiki/Lotka%E2%80%93Volterra_equations.

[6] JIANG Wen-na. Solutions of Several Kinds of Competition Equations with Diffusion or Interleaved Diffusion. *Shanghai Normal University.*

[7] https://en.wikipedia.org/wiki/K-means_clustering.

[8] Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. *2016 ACM. ISBN 978-1-4503-4232-2/16/08*

# Appendix

The appendix consists of two main code of our work, which is GA plus PDM to fitting the PDE and XGBoosting. And the tree structure we construct.

----------fitting----------

```
# -*- coding: utf-8 -*-
"""
Created on Fri Jan 25 21:36:33 2019

@author: 30517
"""

from scipy import linalg
import numpy as np
from sopt.SGA import SGA
from math import sin
from sopt.GA.GA import GA
from sopt.util.functions import *
from sopt.util.ga_config import *
from sopt.util.constraints import *
from time import time




def func1(x):
    print(x)
    DH=x[0]

    h1=x[1]
    h2=x[2]
    h3=x[3]

    DS=x[4]

    s1=x[5]
    s2=x[6]
    s3=x[7]


    maxInter = 100

    lenX= 151
    lenY= 56
    lent=8
```

```
delta = 1


HH=0
SS=0

guess = 30


X,Y,t =np.meshgrid(np.arange(0,lenX),np.arange(0,lenY),np.arange(0,lent))

H = np.empty((lenX,lenY,lent))
S = np.empty((lenX,lenY,lent))

H.fill(guess)
S.fill(guess)

H[:,:,0]=realHData[:,:,0]
H[:,0,:]= realHData[:,0,:]
H[0,:,:]= realHData[0,:,:]
H[150,:,:]= realHData[150,:,:]
H[:,55,:]= realHData[:,55,:]
H[:,:,7]= realHData[:,:,7]

S[:,:,0]=realSData[:,:,0]
S[:,0,:]= realSData[:,0,:]
S[0,:,:]= realSData[0,:,:]
S[150,:,:]= realSData[150,:,:]
S[:,55,:]= realSData[:,55,:]
S[:,:,7]= realSData[:,:,7]




print('Solving, Please wait')

for interation in range(0,maxInter):
    print(interation)
    for i in range(1,lenX-1,delta):
        for j in range(1,lenY-1,delta):
            for t in range(1,lent-1,delta):
                A = np.array([[4*DH+h2-1,h3],[s3,4*DS+s2-1]])
                B = np.array([DH*(H[i+1,j,t]+H[i-1,j,t]+H[i,j+1,t]+H[i,j-1,t])+h1-
```

```
H[i,j,t+1],DS*(S[i+1,j,t]+S[i-1,j,t]+S[i,j+1,t]+S[i,j-1,t])+s1-S[i,j,t+1]])
                        idea=linalg.solve(A,B)
                        H[i,j,t]=idea[0]
                        S[i,j,t]=idea[1]




    a=(((H-realHData)**2).sum() + ((S-realSData)**2).sum() )*(1/103888)

    print(a)
    return a




class TestGA:
    def __init__(self):
        self.func = func1
        self.func_type = 'min'
        self.variables_num = 8
        self.lower_bound = 0
        self.upper_bound = 1000
        self.cross_rate = 0.8
        self.mutation_rate = 0.05
        self.generations = 200
        self.population_size = 100
        self.binary_code_length = 20
        self.cross_rate_exp = 1
        self.mutation_rate_exp = 1
        self.code_type = code_type.binary
        self.cross_code = False
        self.select_method = select_method.keep_best
        self.rank_select_probs = None
        self.tournament_num = 2
        self.cross_method = cross_method.two_point
        self.arithmetic_cross_alpha = 0.1
        self.arithmetic_cross_exp = 1
        self.mutation_method = mutation_method.gaussian
        self.none_uniform_mutation_rate = 1
        #self.complex_constraints = [constraints1,constraints2,constraints3]
```

```
            self.complex_constraints = None
            self.complex_constraints_method = complex_constraints_method.penalty
            self.complex_constraints_C = 1e6
            self.M = 1e8
            self.GA = GA(**self.__dict__)

        def test(self):
            start_time = time()
            self.GA.run()
            print("GA costs %.4f seconds!" % (time()-start_time))
            self.GA.save_plot()
            self.GA.show_result()



if __name__ == '__main__':
        TestGA().test()

--------------------------------XGBoosting-------------------------------------
import xgboost as xgb
from xgboost import plot_importance
from matplotlib import pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error

#  读取文件原始数据
X=data[:,7:]
Y=datay









# XGBoost 训练过程
X_train, X_test, y_train, y_test = train_test_split(X,Y, test_size=0.2, random_state=0)

model = xgb.XGBRegressor(max_depth=4,
                                    learning_rate=0.04,
                                    n_estimators=500,
```

```
                    silent=True,
                    objective='reg:linear',
                    nthread=-1,
                    gamma=0,
                    min_child_weight=1,
                    max_delta_step=0,
                    subsample=0.85,
                    colsample_bytree=0.7,
                    colsample_bylevel=1,
                    reg_alpha=0,
                    reg_lambda=1,
                    scale_pos_weight=1,
                    seed=1440,
                    missing=None)



model.fit(X_train, y_train)

# 对测试集进行预测
ans = model.predict(X_test)
print(r2_score(ans,y_test))
print(mean_squared_error(ans,y_test))
# 显示重要特征
plot_importance(model)
plt.show()

------------------------------Tree----------------------------------------
```