For office use only	Team Control Number	For office use only
T1 _____	**1902249**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**C**	F4 _____

2019 Mathematical Contest in Modeling (MCM) Summary Sheet

(Attach a copy of this page to each copy of your solution paper.)

Evaluation System for the Opioid Crisis

Summary

In this paper, we analyze and predict the amount and percentage of opioid used in the five states of Ohio, Kentucky, West Virginia, Virginia and Tennessee in 2010-2017, we also put forward suggestions on reducing the number of drug addicts.

Firstly, we use the **K-means clustering algorithm** to analyze the percentage of opioids used by each state and county in the past two years. Then we find out that **46 counties** such as **Virginia** and 21015, 39047, 42003, 51027, 54003 may have started using specific opioids. At the same time, we also determine the drug recognition threshold level through the clustering center.

Secondly, we use the AR(2) model to predict the use of opioid in each state and county in 2019 and 2020 . Then combining with difference equation-linear regression model, we successfully predict that **Pennsylvania** and **12 counties** may exceed the drug identification threshold level in the next two years. More specifically, in 2019, there are **9 counties** such as 39007 and 39023, and there are **3 counties** such as 42019 in 2020.

After that, we screen out the common parts of the annual census social-economic data table, perform polynomial interpolation fitting on the missing data, and then use the **Pearson correlation coefficient** to test the correlation to identify **72 closely related** characteristics of census socio-economic data , including positive correlation features such as "households with one or more people 65 years and over" and negative correlation features like "Females 15 years and over – now married, except separated".

Finally, we select 72 characteristics of census socio-economic data as the new variable and modify **the difference equation - linear regression model**. We also take factors such as proliferation, detoxification and crime of drug users into consideration, and introduce relevant probability parameters to fit a more reasonable model. Using this model, we identify possible strategies to combat the opioid crisis.

In addition, we perform stability and sensitivity analysis on the modified difference equation-linear regression model to identify any important parameter boundary on which success (or failure) depends.

MEMORANDUM

To: the Chief Administer

From: MCM Team

Subject: Our Key Work to Alleviate the Opioid Crisis

Date: Monday, January 28th., 2019

With the boundary between drug's medical use and recreational purposes becoming vaguer, the United States is suffering from great pain brought by opioid. Simply enforcing laws is by no means ideal to alleviate the crisis. To prevent the current situation from worsening, we come up with some strategies to handle this problem. In addition, making sufficient analysis on five states enables us to simplify the process and concentrating on smaller units like 'county' reduces the possibility to miss out potential dangers.

Firstly, finding out possible locations where specific opioid use might have started is so decisive when it comes to pouring more attention to smaller regions. What's more, these places have overdosed opioid so that immediate measures should be taken regarding these regions. But How do we find these places? We employ the K-means clustering algorithm which helps us to classify the locations according to their opioid use. Supported by sufficient data and through careful analysis, we find that there are 46 counties which have started using opioid. At the same time, we incorporate the idea of logarithm to determine the drug threshold level.

Our Second key finding is to predict the tendency of opioid use of each counties. For those places whose drug use is appropriate in recent years, we still cannot ignore their potential risks. Special attention should be taken to those counties which show sharp rise in our prediction. To make our prediction more precise and feasible, we adopt the AR(2) model which can perform interpolation fitting on the present data .Then we come to the conclusion that about 12 counties will exceed the drug threshold level in the next two years with nine counties in 2019 and three counties in 2020. Under this circumstance, government may take measures in advance such as enhancing laws and supervision, strengthening education and so forth.

Next, to give constructive suggestions to the U.S government on the basis of patterns and characteristics, we make full use of census socio-economic data. By using the Pear-

son correlation coefficient, we identify 72 closely related characteristics of census socio-economic data. It is noticeable that households with one or more people 65 years or over shows obvious positive correlation.

Finally, we make some modification to the difference equation-linear regression model. Taking many factors into account like detoxification, proliferation and so on, we tend to identify possible strategies.

Through previous analysis, you can see that our model has a previous generalization and can be applied to many files. So we sincerely hope that our advice could be taken to deal with the crisis.

Contents

1 Introduction

1.1 Problem Background

As is often pointed out, drug sometimes is a two-waged weapon which can be used equally for good and evil. There is no exception for opioid, the pain relieving drug which is highly addictive, has two uses either for the management of pain(prescription use) or for recreational and illegal use. Especially in recent years, some people have misused prescription opioid transition to heroin. Whats more, opioid overdoses have increased about 30 percent over the last two years. In that case, the United States is struggling with a national crisis concerning the use of opioid that affects public health as well as social and economic welfare. Take the health care sector for instance, if the percentage of people with opioid addiction mounts to a higher level among the elderly, health care costs will be affected.

The DEA publishes a data-heavy annual report addressing drug identification results and related information collected by authorities all around the country. There is amount of data from crime laboratory that handles up to 88% of the state drug cases. More specifically, considerable opioid uses are related to drug crime. It is a grimly disastrous spectacle if the opioid crisis spreads to all cross-sections of US with its ever-increasing power. So proper measures should be taken to prevent the situation from worsening.

1.2 Restatement of the Problem

Supplied with sufficient data, we try to help solve some problems to alleviate the severity of the crisis. Before that, we must be more explicit about the problems which are need to be figured out. Thus, we boil down the problem into the following three questions:

- Identify the possible locations where specific opioid use might have started.
- Find out the threshold level of the overdose of opioid. Predict the time and location that the specific opioid use might occur.
- Determine whether the opioid trend-in-use is associated with the provided socio-economic data. If related, try to modify the model.
- Figure out the specific concern that American government should have.

As for the problems mentioned above, we will give our explanations in different sections. More specifically, the first problem will be solved in section 3.1.1, the second in part 3.1.2 and 3.2, the third in 3.3 and 3.4, and the last in 3.5 and 4.1.

2 Preparation of the Models

2.1 Assumptions

- We only consider how much opioid and heroin account for overall controlled substance rather than the absolute amount of the two prescriptions.
- All else being equal,each state or county that exceeds drug identification threshold levels will continue to be above drug identification threshold levels.
- We only tend to consider the the time variable t in model 2.
- Owing to the fact that five states are close in location, we tend to ignore their geographical differences.
- When calculating the correlation coefficient, we will eliminate the items with large errors and take the average value finally.

2.2 Notations

Symbol	Definition
ωi	the i cluster
m_i	the i clustering center
a_i	counties'(states') percentage of opioid use
μ_i	the characteristic of census socio-economic data
$\overrightarrow{U}_t$	$(u_1(t), u_2(t), \cdots, u_k(t))$
$y_t, y(t, \overrightarrow{U}_t)$	the percentage of opioid use
$x(t)$	the percentage of opioid use which only relates to time
λ	the possibility that one drug user successfully induce others to take drugs
s_1	the possibility of withdrawal from drug addiction
s_2	the possibility of committing crime or suicide because of taking drugs
γ	lagging index

Table 1: Notations

The primary notations used in this paper are listed in the table above.

3 Model Solution

3.1 Model 1:K-means Test Model

3.1.1 K-means Algorithm

The overdose of opioid is a relatively vague concept and is hard to define. In order to get a clear picture of the problem, we tend to classify the data whose rate value is similar to one certain category, which is based on the theory of K-means. Hence, we establish the K-means test model which enables us to pick out the locations whose opioid use rate is relatively high. In fact, we do **not** tend to use the year as the samples variable. However, Taking the migration, political events and the break out of sudden crisis into account, We always think that two years are **independent** of each other.To identify the possible locations where specific opioid use might have started in the five states, we employ year 2016 and 2017 as two variables. Firstly, we employ the two-dimensional vector $(a, b)^T$ to denote each county's the percentage of opioid use. Then we use logarithm to express it as $(2 - lg(80 - a), 2 - lg(80 - b))^T$.

Next, We discuss the basic processes of the K-means test model:

- Sort out k objects whatever we want from overall data as the clustering center.
- According to the average value, also known as the center target, we calculate the distance between every single target and the center target.
- Recalculate the average value of each cluster until the center of the category does not change any more.
- Repeat step 2 and step 3 until there is no change in the classification which enables the following value attains its minimum.

$$E = \sum_{j=1}^{k} \sum_{x \in \omega_j} ||x_i - m_j||^2 \tag{1}$$

After implementing Model 1 to MATLAB, we get the K-means cluster analysis graph as follows:

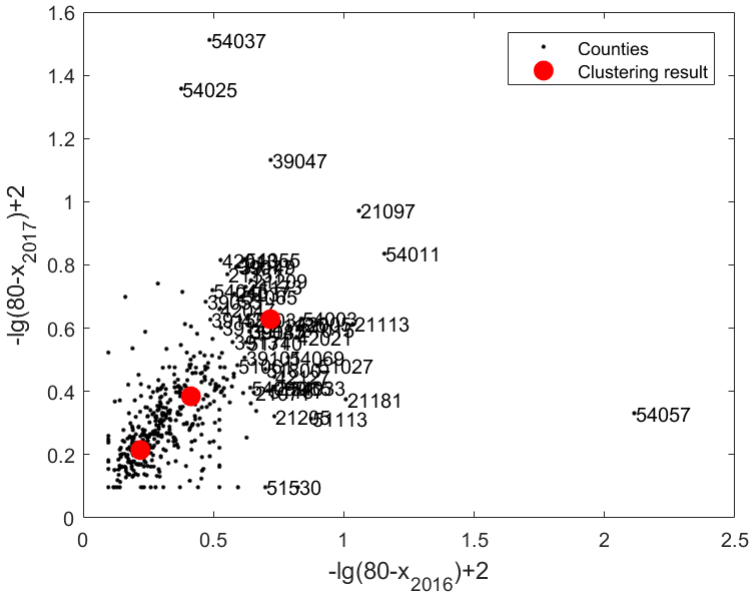

Figure 1: K-means Cluster Analysis Graph

As is shown in the graph above, three red dots are clustering centers. Rest points with numbers are the possible locations where specific opioid use might have started of the five states. To get a clear picture of the problem, we list the corresponding Table 2 as follows:

Kentucky(21)	**Ohio(39)**	**Pennsylvania(42)**	**Virginia(51)**	**West Virginia(54)**
21015	39047	42003	51027	54003
21037	39053	42005	51061	54011
21065	39077	42013	51113	54025
21071	39087	42021	51187	54033
21097	39101	42047	51530	54037
21113	39131	42127	51740	54045
21135	39145		51800	54055
21151	39147			54057
21173	39149			54065
21181	39155			54069
21205				54099
21209				

Table 2: Possible Locations Where Specific Opioid Use might Have Started

3.1.2 Threshold Level Identification

Then we set k to 3 and implement our model in Matlab . To facilitate understanding and further analysis, we give schematic diagram of the model, shown in Figure 2. The

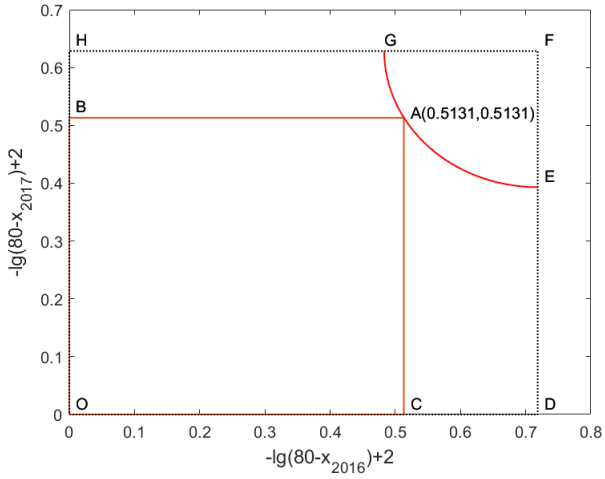

Figure 2: Drug Threshold Level Delimitation Graph

utilization of the Daniel test theory makes our model close to reality and practical. Based on the theory, we implement our model successfully via MATLAB. Then we can safely reach the conclusion that there are 120 counties whose corresponding time series have the tendency to ascend. Suppose the threshold level is a, then we can get back to Model 1 and it is rational for us to assume vector $\overrightarrow{\alpha} = (2 - lg(80 - a), 2 - lg(80 - a))^T$ to be critical point. In addition, $\overrightarrow{\alpha}$ should be located at the boundary of ω_3 that is close to the original point. To facilitate understanding and illustrate our thought more explicitly, we draw the following schematic diagram ,shown in figure 2. As is shown in the diagram, F is the clustering center m_3 and $A(2 - lg(80 - a), 2 - lg(80 - a))^T$ is the boundary point. To find out the threshold level, we consider $x \in \omega_3$ and is located in the lower left quarter of F. Because A is the boundary point, we have:

$$|\overrightarrow{\alpha} - m_3|_2 \geq |x - m_3|_2 \qquad \forall x \in \omega_3$$

which can also be expressed as:

$$|\overrightarrow{\alpha} - m_3|_2 \geq \max_{x \in \omega_3} |x - m_3|_2 \tag{2}$$

By calculation, we have $a \leq 49.3352$. For better and more precise estimation, we assign the value 49.3352 to a.

To guarantee the rationality of the data, we consider the fluctuation beyond 45 from 2019 to two 2020 to be reasonable and pick out those data which does not qualify.

3.2 Model 2: AR(2) Test Model

3.2.1 The Introduction to the Model

By now, we have figured out the possible locations where specific opioid use might have started. In order to pay more attention to those counties and states which suffer greatly from the crisis, we should predict the tendency of the prescription use. To look for a suitable method to solve this problem, we introduce the Daniel test theory.

3.2.2 Spearman Correlation Coefficient

Spearman correlation coefficient is a nonparametric measure of rank correlation. It assesses how well the relationship between two variables can be described using a monotonic function.

Here we introduce the two-dimensional random variable denoted as (X,Y).

Accordingly, we can get a set of data which is denoted by $x_1, x_2, \cdots, x_n$ and $y_1, y_2, \cdots, y_n$.Let $R_1, R_2, \cdots, R_n$ denotes the rank statistic of $x_1, x_2, \cdots, x_n$,

$S_1, S_2, \cdots, S_n$ denotes the rank statistic of $y_1, y_2, \cdots, y_n$, then let q_{xy} be the correlation coefficient of the two groups of rank statistics, which is also known as Spearman coefficient:

$$q_{xy} = \frac{\sum_{i=1}^{n}(R_i - \overline{R})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \overline{S})^2}} \tag{3}$$

where

$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n} R_i;\ \overline{S} = \frac{1}{n}\sum_{i=1}^{n} S_i$$

By careful calculation, we can get the following equation:

$$q_{xy} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2 \tag{4}$$

where

$$d_i = R_i - Si, i = 1, 2, \cdots, n$$

3.2.3 Daniel Test Theory

Let $a_1, a_2, \cdots, a_n$ be the sample we consider concerning time series and we denote the rank of a_t as $R_t = R(a_t)$. Assume q_s is the Spearman correlation coefficient which reflects the dependence of variable with (t, R_t), and it can be expressed as:

$$q_{xy} = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2 \tag{5}$$

Then we can obtain the statistic:

$$T = \frac{q_s \sqrt{n-2}}{\sqrt{1-q_s^2}} \tag{6}$$

Next, we use the hypothesis test theory:

H_0: The series X_t is stable

H_1: The series X_t is unstable, which means the series has shown tendency to ascend or descend. For the sake of determining the stability of series, we adopt the Daniel test theory, that is: As for the notable levelα,we calculate the Spearman correlation coefficient q_s of $(t, R_t), t = 1, 2, \cdots, n$ by making use of s_t. If $|T| > t_{\frac{\alpha}{2}}(n-2)$,then we consider the series to be unstable. What's more,if $q_s > 0$, we think the series tend to ascend.If $q_s < 0$, we think the series tend to descend. And If $|T| < t_{\frac{\alpha}{2}}(n-2)$, then we consider the series to be unstable.

3.2.4 AR(2) Model

In order to find out the counties whose opioid use rate is ascending and according to the Daniel test theory, we focus on series which satisfy $|T| > t_{\frac{\alpha}{2}}(n-2)$ and $q_s > 0$. However, these series are unstable.

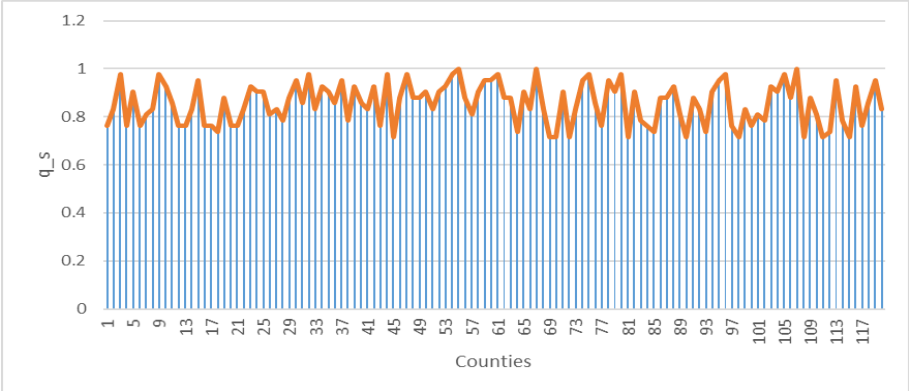

Figure 3: Counties with An Increasing Percentage of Opioid Use in 2010-2016

To get the stable series, we do careful first-order difference calculation

$$b_t = a_{t+1} - a_t \tag{7}$$

about the series $a_t(t = 1, 2, \cdots, 8)$ to get the series $b_t(t = 1, 2, \cdots, 7)$. As is shown from most of counties' scatter diagram and after careful test, we reach the conclusion that the time series is stable. In order to predict b_t, we set up a (AR(2)) model listed as following:

$$y_t = c_1 y_{t-1} + c_2 y_{t-2} + \varepsilon_t \tag{8}$$

where c_1,c_2 are unknown parameters; ε_t is the random disturbance parameter.

New County Number in 2019	New County Number in 2020
39007	39095
39023	42019
39057	42043
39099	
39123	
42055	
42101	
51135	
54107	

Table 3: Counties that may exceed threshold levels in the next two years

The table demonstrates the change in the number of counties with potential opioid overdose risk from 2019 and 2020. It is also clear that there are 10 counties that may surpass threshold level while there are only 3 in the next year, which shows great decline.

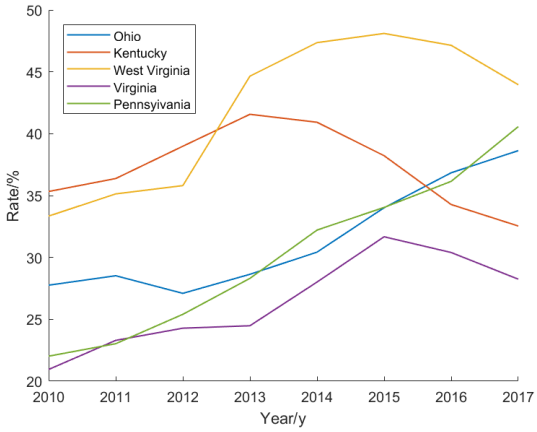

Figure 4: Percentage of Raw Data - Year Line Chart

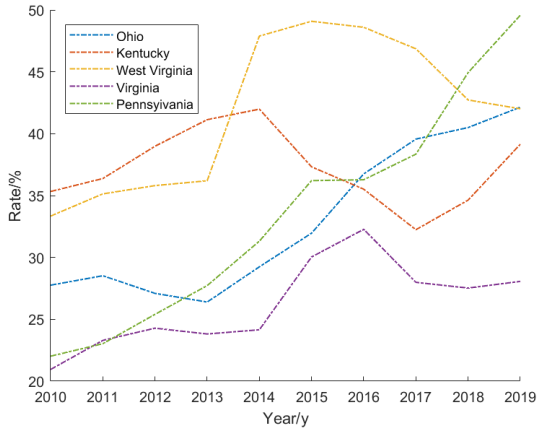

Figure 5: Percentage of Forecasted Data - Year Line chart

From Figure 4 , we can see clearly that West Virgiria reaches the peak in 2016, which means it has overdosed opioid already. However, as can be seen from Figure 5, there

is an upward increase from 2018 to 2019 in Pennsylvanias data, which represents the high probability of overdose in the future.

3.3 Model 3: Pearson Correlation Coefficient Test Model

Considering the invalidity of some data, we tend to make some adjustments to the data. As for the unknown statistics, if the data corresponding to the countys characteristic(census socio-economic data) is uncertain, we use zero as default. If not, we adopt the polynomial interpolation method on the basis of year to do the data fitting for the sake of completing the information. In addition, when dealing with the data which has significant deviation from the standard statistics, we retain the raw data and make simple analysis in the following Sensitivity Analysis section. In order to ensure the accurancy and reliability of the model, we sort out the repeating statistics from 2010 to 2016. Next, we classify the Estimate and Percent into two categories.

As for one kind 'Percent', we tend to make analysis on the relevance between the percentage of using opioid in each county and various characteristics and pick out those whose relevance is high. In this process, we use Pearson correlation coefficient, which aims to measure the linear dependence between two variables. Let r_{XY} denote the Pearson correlation coefficient, and we it can be expressed as:

$$r_{XY} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - (E(X))^2}\sqrt{E(Y^2) - (E(Y))^2}} \tag{9}$$

where $E(X)$ is the sample average value. In most cases,the symbol of the correlation coefficient r_{XY} reflects the direction of correlation and the absolute value represents the degree to which two variables are related to each other. Besides, the absolute value usually will not surpass 1, which means r_{XY} value range between -1 and 1. If $r_{XY} < 0.4$, we say two variables are inrelevant; If $0.4 < r_{XY} < 0.75$, we consider two variables are greatly relevant. And if $0.75 < r_{XY} < 1$, there exists linear dependence between two variables.

According to the discussion above, we pick out 72 characteristics(census socio-economic data) and the most representative Pearson correlation coefficients. That is to say, the absolute value of r_{XY} is relatively large.With the help of MATLAB, we obtain the result as follows:

In fact, we have determined 72 closely related census society. Here we list ten of them. According to the Table 4, the top six show positive correlation while the rest show negative correlation features. Also it is remarkable that households with one or more 65 years and over occupy the headline.

Census Socio-economic Characteristics	Correlation Coefficient
Households with one or more people 65 years and over	0.682
High school graduate or higher	0.668
Females 15 years and over(Never married)	0.637
Males 15 years and over(Never married)	0.647
Percent bachelor"s degree or higher	0.667
Householder living alone(65 years and over)	0.628
Males 15 years and over(Now married, except separated)	-0.668
Total population(English)	-0.649
Females 15 years and over(Now married, except separated)	-0.644
Married-couple family(With own children under 18 years)	-0.664

Table 4: Correlation Matrix of the Proportion and Influence Factors of Opioids in 5 States in 2010-2016 (partial)

* Due to space limitations, the criteria selected in this table are the ten largest absolute values among the correlation coefficient matrices.

3.4 Model 4: Difference Equation-Linear Regression Model

3.4.1 Construction of the Model

Now let us get back to AR(2) Model. Combining equation(6) and equation(7), we have:

$$a_{t+1} = (c_1 + 1)a_t + (c_2 - c_1)a_{t-1} - c_2 a_{t-2} + \varepsilon_t \tag{10}$$

This is a difference equation whose feasibility is based on prior data and least squares.

In Model 3, we reach the conclusion that there is 72 variables linearly dependent to opioid use amount by using Pearson correlation coefficients. So we can modify the equation of AR(2) Model as :

$$y_(t, l_1, \cdots, l_k) = \mu_0 + \mu_1 l_1 + \mu_2 l_2 + \cdots + \mu_k l_k + x(t) \tag{11}$$

where $k = 72$ and $\mu_1, \mu_2, \cdots \mu_k$ are k characteristics of census socio-economic data. $x_(t)$ is the coefficient which is only related to t. Owning to the assumption that we only consider how much opioid and heroin account for overall controlled substance rather than the absolute amount of the two prescriptions, the increase in the number of drug users can represent the increase of the percentage of using opioid. Firstly, we consider the number of drug users in the year t. Suppose the probability of one drug user successfully influence others on using drugs in the year of $t-1$ is λ, due to the fact that controlled substance consists of non-opioid substance, we introduce the hysteresis

index γ and obtain the equation:

$$\Delta x(n) = x(n+1) - x(n) = \lambda(\gamma - x(n))x(n) + \delta(n) \tag{12}$$

where

$$\delta(n) \text{ is the remaining substance,} \qquad 0<\gamma<1$$

By searching sufficient materials, we know that Methadone and Buprenorphine are two medications designed to reduce opioid withdrawal symptoms. We also take the factor into account that drug users might commit crime or suicide because of drug addiction. More specifically, we consider that there is occasional factor like drug use leading to accidental occurance. So in order to simplify the model, we predict the possibility of recovery to be s_1 and suppose the possibility of committing suicide and crime is s_2. Then we have:

$$\delta(n) = -s_1 x(n) - s_2 x(n) \tag{13}$$

Combining all the formulas in Model 4 together, we have :

$$y(n) = \mu_0 + \sum_{i=1}^{k} l_i \mu_i + x(n) \tag{14}$$

$$x(n+1) = \lambda(y - x(n))x(n) + (1 - s_1 - s_2)x(n)$$

Then we get the modified model.

3.4.2 Algorithm of the Model

This model is a modification of model 2 which is based on prior data and least square method. The specific algorithm is as follows:

Let $\overrightarrow{L} = (l_1, l_2, \cdots, l_k)^T$, $\overrightarrow{U} = (\mu_1, \mu_2, \cdots, \mu_k)^T$, then we have :

$$y(t, \mu_1, \mu_2, \cdots, \mu_k) = \mu_0 + \overrightarrow{L}^T \overrightarrow{U} + x(t) \tag{15}$$

$$x(t) = \lambda(\gamma - x(t-1))x(t-1) - (s_1 + s_2)x(t-1)$$

where

$\mu_1, \mu_2, \cdots, \mu_k$ are k variables, μ_0 is a constant. In fact, let $\mu_i (i = 1, 2, \cdots, k)$ denote the characteristic of census socio-economic data, they can also be seen as time series and $\overrightarrow{L} = (l_1, l_2, \cdots, l_k)^T$, $\overrightarrow{U_t} = (\mu_1(t), \mu_2(t), \cdots, \mu_k(t))^T$, $y(t-1, \overrightarrow{U}_{t-1}) = \overrightarrow{L}^T \overrightarrow{U}_{t-1} + x(t-1)$, we can get the equation:

$$\begin{aligned} y(t, \overrightarrow{U}_t) = \mu_0 + \overrightarrow{L}^T \overrightarrow{U}_t + \lambda(\gamma - y(t-1, \overrightarrow{U}_{t-1})) + \overrightarrow{L}^T \overrightarrow{U}_{t-1})(y(t-1, \overrightarrow{U}_{t-1}) \\ - \overrightarrow{L}^T \overrightarrow{U}_{t-1}) - (s_1 + s_2)(y(t-1, \overrightarrow{U}_{t-1}) - \overrightarrow{L}^T \overrightarrow{U}_{t-1}) \end{aligned} \tag{16}$$

Let $a(t)$ denote the original value of (6). After the value $\overrightarrow{L}, \overrightarrow{U}_t, \overrightarrow{L}, \overrightarrow{U}_{t-1}$ is certain(Here we have $y(t-1) = a(t-1)$ when we approximate $a(t)$, according to seven groups of data, let $f(L) = \sum_{t=2}^{7} (a(t) - y(t))^2$. To get the minimum value of $f(L)$,we adopt the least squares and get the equation:

$$\nabla f(L) = 0, \qquad i.e. \ \frac{\partial f}{\partial l_i} = 0, \ i = 1, 2, ..., k \tag{17}$$

by which we can calculate related parameter variables.

4 Data Analysis and Model Validation

- Clustering algorithm is adopted in Model 1. Considering the huge data provided, we approximate that such data satisfy the normal distribution. However, in order to make it easier to find counties with a high proportion of opioid use(that is, increase the feasibility of K-means aggregation algorithm), we decide to use logarithmic transformation for the sake of increasing index's convexity. Furthermore, when we set the critical point, we want a consistent standard for each year, so it is inappropriate to simply look for data in two-dimensional coordinates. Instead, we should look for a point with the same horizontal and vertical coordinates. We know the clustering center of K-means clustering algorithm will change, once the samples change. Because of that, when choosing the critical point, we tend to let the distance from any point that belongs to a higher use of opioid to the existing clustering center is less than the distance from the clustering center to the critical point, which means the algorithm of choosing the critical point will no longer depend on the K-means algorithm.

- Model 2 adopts a version of ARMA model under the premise that a proportion of opioid use is only time-dependent and does not add any other assumptions. Obviously, this is a simple fit, but the output is generally good.

- In model 3, Pearson correlation coefficient test is used to find out the relatively linear correlation and prepare for the subsequent modification of the model.In addition, we use the PCA algorithm to revalidate the correlation between census socio-economic data and the proportion of opioid use.
- The value of $\overrightarrow{U}_t$ is needed to predict $y(t)$ in the future in the model above. However, they are usually unknown to us. So in order to predict the value of $\overrightarrow{U}_t$ in the future, we can make full use of AR(2) model in Model 2. After that, we can predict y(t).
- Notice that the value of variables L,U has been determined, we can only consider λ,s_1, s_2. By enhancing education among teenagers to raise their awareness of drug's potential danger, λ can be greatly reduced. Besides, expanding the scale of investment in Methadone and Buprenorphine can be helpful to reduce the value of s_1. Also strengthening the law supervision and reduce the crime rate leads to a smaller value of s_2.
- Through analysis on the characteristics of census socio-economic data, we have some new discoveries.If the number of households with one or more 65 year old people increases, it is shown that the possibility of one's taking drugs will increase correspondingly. In other words, population ageing has positive correlation with the probability of drug abuses. On the contrary, with the increase of females who are married, the drug users will be less than it used to be.

5 Sensitivity and Stability Analysis

5.1 Sensitivity Analysis of K-means Clustering Algorithm

For model 1, we process the data first before analyzing it. We can adjust the function that processes the data to get a slightly different classification. Figures 6 and Figures 7 below are two different ways of dealing:

The method of Figure 6 is a direct classification of the original data, the output cluster is almost divided by the tangent plane. In general, for binary variables, this approach is desirable. But the method we used in Model 3.1.2 is no longer applicable. On the other hand, for example, the two data $(40, 50)^T$ and $(50, 40)^T$ are at the boundary of the center of the cluster, and the fluctuation is too large within two years, so it is not easy to determine the drug identification threshold level.

The method of Figure 7 uses a quadratic function $-0.01x^2+x$, which uses a concave

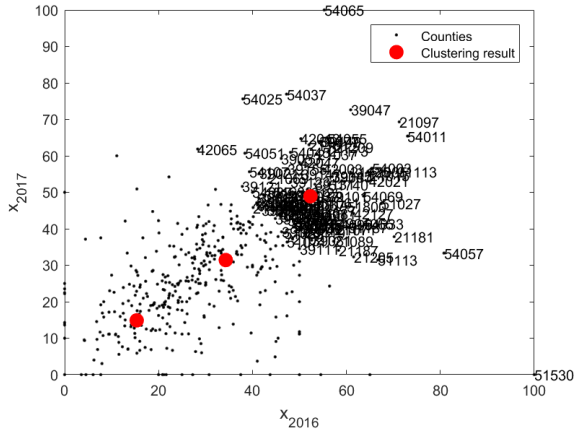

Figure 6: K-means Cluster Analysis Graph 1

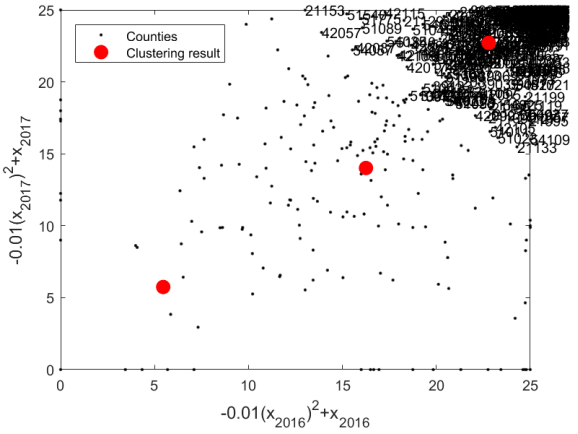

Figure 7: K-means Cluster Analysis Graph 2

function to process the data. Obviously, this type of classification divides most counties into categories that overuse opioid. This is not conducive to distinguishing between counties that overuse opioid.

Combined with the above analysis, we can see that the convex function is more conducive to distinguishing the data, and it is also more able to concentrate the data on the line of $y = x$, which is easier to determine the drug identification threshold level.

5.2 Stability Analysis

Suppose we neglect the variation of $\mu_1, \mu_2, \cdots, \mu_k$, the equation (18) can been seen as a difference equation approximately. And it can also be expressed as :

$$y(t) = -\lambda y(t-1)^2 + (\lambda y - \overrightarrow{L}^T \overrightarrow{U}) - s_1 - s_2)y(t-1) + (1 + s_1 + s_2 - \lambda\gamma)\overrightarrow{L}^T \overrightarrow{U} \quad (18)$$

where the right of the equation is a quadratic function about $y(t-1)$, we denote it as $h(y(t-1))$, and the left is a function about $y = x$. Firstly, make a coordinate transformation and then we have:

$$y^*(t) = -\lambda y^*(t-1)^2 + Y^* y^*(t-1) = h^*(y^*(t-1)) \quad (19)$$

where Y^* is the parameter of linear item and $h^*(y^*(t-1))$ is a quadratic equation. As is shown in the figure below:

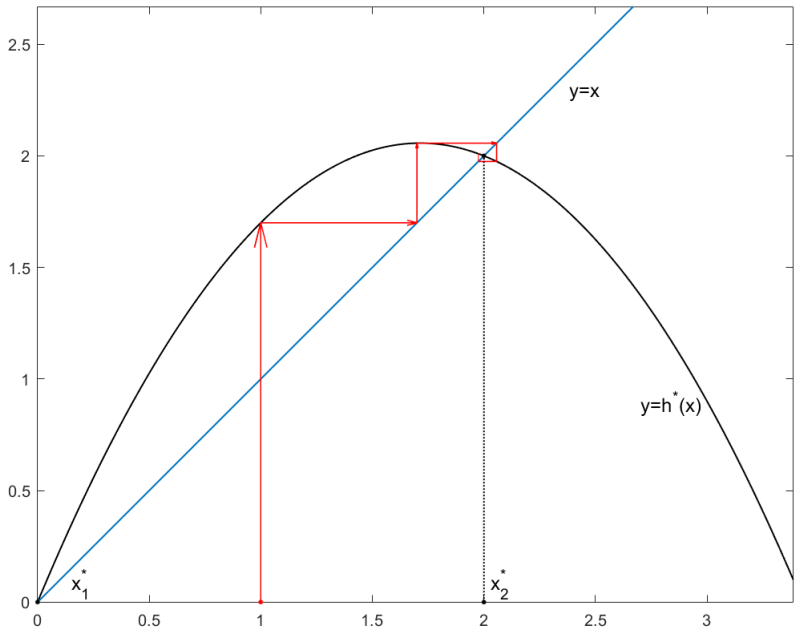

Figure 8: the Percentage of Opioid Use Iteration Graph

As long as the preliminary point is greater than x_1^* in value and converges toward a value of the point x_2^*. Here we require that:

$$\triangle = (\lambda\gamma - \overrightarrow{L}^T\overrightarrow{U} - s_1 - s_2)^2 + 4\lambda(1 + s_1 + s_2 - \lambda\gamma) > 0 \tag{20}$$

Let two roots of $h(x) = 0$ be x_1 and x_2 ($x_1 < x_2$) which correspond with x_1^* and x_2^*.Moreover, they satisfy:

$$x_1 \geq 0 \quad , \quad x_2 < 1 \tag{21}$$

from which we can get the data ranges of $\gamma, \lambda, s_1, s_2$ .In this case, we consider y(t)in Model 4 is stable when it is not influenced by $\mu_1, \mu_2, \cdots, \mu_k$.

6 Strengths and Weaknesses

6.1 Strengths

Our outlined model has particular strengths, unavailable in other models.

- Visualization: We analyze the correlation between variables coming to the conclusion that they are nearly independent. Besides, we carefully construct cluster-

ing analysis and adopt the method of logarithmic transformation. Such practice brings simplification and helps to highlight the distinction the overdose of opioid and proper drug use.

- Flexibility: In Model 2, we make the accurate prediction by using APMA mode. The method to determine the critical point is ingenius which incorporates geometry. It also apply to vast majority of clustering analysis.
- Innovation and Practicality: We fully make use of the idea in the iteration of the difference equation.

6.2 Weaknesses

It is important to consider the weaknesses of our model that may be worth addressing in the future.

- Data Limitations: In Model 4, it is hard to define the coefficient. Furthermore, they may be the function concerning other dependable variables.
- Predicted Limitation: In the models, we made the prediction based on AR(p) model, but the AR (p) model only applies to long-term predictions.
- Data Imperfection: When we use the Pearson correlation coefficient to detect the correlation of data, some data are very different from the overall data, which may be due to other factors, but we are limited by time and just ignore it without processing.

References

[1] Shoukui Si, Zhaoliang Sun. *Mathematical Modeling Algorithm and Application*; National Defense Industry Press: Beijing, China, 2017; pp.409-412,418-421.

[2] Jianpin Zhu. *Applied Multivariate Statistical Analysis*; Science Press: Beijing, China, 2006; pp.62-92.

[3] Qiyuan Jiang, JinXing Xie, Jun Ye. *Mathematical Model*; Higher Education Press: Beijing, China, 2018; pp.180-187.

[4] Jinwu Zhuo. *Application of MATLAB in Mathematical Modeling*; Beihang University Press: Beijing, China, 2017; pp.8-10,48-91.

[5] Li Yibin. Considering the Socio Economic Elements of Indian Terrorism. *South Asian Studies*, 2018(2); pp.139-160.

[6] Wikipedia: Pearson correlation coefficient. 2019.1.27. `https://en.wikipedia.org/wiki/Pearson_correlation_coefficient`

Appendix

This MATLAB program is used to solve Part1.

Program 1: `Part1.m`

```
%%
%read data and preliminary processing of data
clc,clear;
yearData=xlsread('MCM_NFLIS_Data.xlsx','Data','A2:A24063');
DeltaMat=ones(24062,1);
yearData=yearData-2009*DeltaMat;%2010-2017 mapping 2-9
fipsData=xlsread('MCM_NFLIS_Data.xlsx','Data','D2:F24063');
reportsData=xlsread('MCM_NFLIS_Data.xlsx','Data','H2:J24063');
reportsData(:,4)=reportsData(:,1)./reportsData(:,2)*100;
reportsData(:,5)=reportsData(:,1)./reportsData(:,3)*100;
xlswrite('rate.xlsx',reportsData);
%%
%x:Year;y:County;z:DrugReports/TotalDrugReportsCounty,5 datasets for 5 states
rateOhioCounty=zeros(9,840);
rateKentuckyCounty=zeros(9,840);
rateWVirginiaCounty=zeros(9,840);
rateVirginiaCounty=zeros(9,840);
ratePennsyivaniaCounty=zeros(9,840);
rateState=zeros(8,5);
%%
for i=1:24062
    switch fipsData(i,1)
        case 39 %Ohio
            rateOhioCounty(1,fipsData(i,2))=39*1000+fipsData(i,2);
            rateOhioCounty(yearData(i,1)+1,fipsData(i,2))...
                =rateOhioCounty(yearData(i,1)+1,fipsData(i,2))...
                +reportsData(i,4);
            rateState(yearData(i,1),1)=rateState(yearData(i,1),1)...
                +reportsData(i,5);
        case 21 %Kentucky
            rateKentuckyCounty(1,fipsData(i,2))=21*1000+fipsData(i,2);
            rateKentuckyCounty(yearData(i,1)+1,fipsData(i,2))...
                =rateKentuckyCounty(yearData(i,1)+1,fipsData(i,2))...
                +reportsData(i,4);
            rateState(yearData(i,1),2)=rateState(yearData(i,1),2)...
                +reportsData(i,5);
        case 54 %West Virginia
            rateWVirginiaCounty(1,fipsData(i,2))=54*1000+fipsData(i,2);
            rateWVirginiaCounty(yearData(i,1)+1,fipsData(i,2))...
                =rateWVirginiaCounty(yearData(i,1)+1,fipsData(i,2))...
                +reportsData(i,4);
            rateState(yearData(i,1),3)=rateState(yearData(i,1),3)...
```

```
                +reportsData(i,5);
        case 51 %Virginia
            rateVirginiaCounty(1,fipsData(i,2))=51*1000+fipsData(i,2);
            rateVirginiaCounty(yearData(i,1)+1,fipsData(i,2))...
                =rateVirginiaCounty(yearData(i,1)+1,fipsData(i,2))...
                +reportsData(i,4);
            rateState(yearData(i,1),4)=rateState(yearData(i,1),4)...
                +reportsData(i,5);
        case 42 %Pennsylvania
            ratePennsyivaniaCounty(1,fipsData(i,2))=42*1000+fipsData(i,2);
            ratePennsyivaniaCounty(yearData(i,1)+1,fipsData(i,2))...
                =ratePennsyivaniaCounty(yearData(i,1)+1,fipsData(i,2))...
                +reportsData(i,4);
            rateState(yearData(i,1),5)=rateState(yearData(i,1),5)...
                +reportsData(i,5);
    end
end
%%
rateOhioCounty(:,all(rateOhioCounty==0,1))=[];
rateKentuckyCounty(:,all(rateKentuckyCounty==0,1))=[];
rateWVirginiaCounty(:,all(rateWVirginiaCounty==0,1))=[];
rateVirginiaCounty(:,all(rateVirginiaCounty==0,1))=[];
ratePennsyivaniaCounty(:,all(ratePennsyivaniaCounty==0,1))=[];
%% Cluster analysis
x=[rateOhioCounty(8:9,:),rateKentuckyCounty(8:9,:),rateWVirginiaCounty(8:9,:)...
    ,rateVirginiaCounty(8:9,:),ratePennsyivaniaCounty(8:9,:)];
x=(-log10(80-x)+2)';
%x=x';
%x=(-0.01.*x.^2+x)';
z=zeros(3,2);
z1=zeros(3,2);
z=x(1:3,1:2);
while 1
    count=zeros(3,1);
    allsum=zeros(3,2);
    badset=[];
    badsetPos=1;
    for i=1:461 % For 461 samples i, calculate the distance to 3 cluster centers
        temp1=sqrt((z(1,1)-x(i,1)).^2+(z(1,2)-x(i,2)).^2);
        temp2=sqrt((z(2,1)-x(i,1)).^2+(z(2,2)-x(i,2)).^2);
        temp3=sqrt((z(3,1)-x(i,1)).^2+(z(3,1)-x(i,2)).^2);
        if(temp1<temp2&&temp1<temp3)
            count(1)=count(1)+1;
            allsum(1,1)=allsum(1,1)+x(i,1);
            allsum(1,2)=allsum(1,2)+x(i,2);
            badset(badsetPos,:)=x(i,:);
            badsetPos=badsetPos+1;
        elseif(temp2<temp1&&temp2<temp3)
```

```
            count(2)=count(2)+1;
            allsum(2,1)=allsum(2,1)+x(i,1);
            allsum(2,2)=allsum(2,2)+x(i,2);
        elseif(temp3<temp1&&temp3<temp2)
            count(3)=count(3)+1;
            allsum(3,1)=allsum(3,1)+x(i,1);
            allsum(3,2)=allsum(3,2)+x(i,2);
        else
            assert(0);%impossible to reach here
        end
    end
    z1(1,1)=allsum(1,1)/count(1);
    z1(1,2)=allsum(1,2)/count(1);
    z1(2,1)=allsum(2,1)/count(2);
    z1(2,2)=allsum(2,2)/count(2);
    z1(3,1)=allsum(3,1)/count(3);
    z1(3,2)=allsum(3,2)/count(3);
    if(z==z1)
        break;
    else
        z=z1;
    end
end
%% County clustering results display
figure(1);
disp(z1);
plot(x(:,1), x(:,2),'k.')
label=[rateOhioCounty(1,:),rateKentuckyCounty(1,:),rateWVirginiaCounty(1,:)...
    ,rateVirginiaCounty(1,:),ratePennsyivaniaCounty(1,:)];
badsetPos=1;
badsetNum=size(badset,1);
for ii=1:461
    if x(ii,:)==badset(badsetPos,:)
        text(x(ii,1)+0.005,x(ii,2),num2str(label(ii)));
        badsetPos=badsetPos+1;
        if(badsetPos>badsetNum)
            break;
        end
    end
end
hold on
plot(z1(:,1),z1(:,2),'ko',...
    'LineWidth',1,...
    'MarkerSize',9,...
    'MarkerEdgeColor','r',...
    'MarkerFaceColor',[1,0,0])
%xlabel('-lg(80-x_{2016})+2','fontsize',12);
%ylabel('-lg(80-x_{2017})+2','fontsize',12);
```

```
%xlabel('x_{2016}','fontsize',12);
%ylabel('x_{2017}','fontsize',12);
%xlabel('-0.01(x_{2016})^2+x_{2016}','fontsize',12);
%ylabel('-0.01(x_{2017})^2+x_{2017}','fontsize',12);
legend('Counties','Clustering result');
hold off
%% Output the county clustering result to excel
resPos=1;
for i=1:461
    if x(i,:)==badset(resPos,:)
        resultCounties(resPos,1)=label(i);
        resPos=resPos+1;
        if(resPos>badsetNum)
            break;
        end
    end
end
xlswrite('Possible_Counties.xlsx',resultCounties);
%% Analyze clusters for each state based on county-level clustering results
state=rateState(7:8,:);
state=(-log10(80-state)+2)';
for i=1:5
    temp1=sqrt((z1(1,1)-state(i,1)).^2+(z1(1,2)-state(i,2)).^2);
    temp2=sqrt((z1(2,1)-state(i,1)).^2+(z1(2,2)-state(i,2)).^2);
    temp3=sqrt((z1(3,1)-state(i,1)).^2+(z1(2,2)-state(i,2)).^2);
    if(temp1<temp2&&temp1<temp3)
        fprintf(...
        'The analysis results for the %d state belong to the first category\n'...
        ,i);
    elseif(temp2<temp1&&temp2<temp3)
        fprintf(...
        'The analysis results for the %d state belong to the second category\n'...
        ,i);
    elseif(temp3<temp1&&temp3<temp2)
        fprintf(...
        'The analysis results for the %d state belong to the third category\n'...
        ,i);
    end
end
%% Filter the county that meets the condition of the difference equation
CountyMat=[rateOhioCounty,rateKentuckyCounty,rateWVirginiaCounty...
    ,rateVirginiaCounty,ratePennsyivaniaCounty];
CountyMat=CountyMat';
%Retain the index of the original element in the sorted sequence
[~,MatIndex]=sort(CountyMat(:,2:9),2);
MatIndex=[CountyMat(:,1),MatIndex];
n=8;
arr=1:n;
```

```
q_s=[];
for i=1:461
    q_s=cat(1,q_s,1-6./(n*(n^2-1))*sum((arr-MatIndex(i,2:9)).^2));
end
T=q_s.*sqrt(n-2)./sqrt(1-q_s.^2);
judge=[MatIndex(:,1),T,q_s];
deleteIndex=abs(judge(:,2))<2.4469|judge(:,3)<0;
judge(deleteIndex,:)=[];
%% Solve the difference equation for 120 counties and save ...
%% the estimated values and coefficients of 2019 and 2020.
solvingPos=1;
judgeLen=size(judge,1);
prevData=[];
compData=[];
for i=1:461
    if(CountyMat(i,1)==judge(solvingPos,1))
        differ=diff(CountyMat(i,2:9));
        model=ar(differ,2,'ls');
        differHat=predict(model,differ');
        differHat(end+1:end+2)=forecast(model,differ',2);
        aHat=[CountyMat(i,2),CountyMat(i,2:9)+differHat(1:end-1)'];
        aHat=[aHat,aHat(end)+differHat(end)];
        tempData=[CountyMat(i,1),aHat(9:10),model.a(2:3)];
        tempData2=[CountyMat(i,12:9),aHat(1:8)];
        prevData=[prevData;tempData];
        compData=[compData;tempData2];
        solvingPos=solvingPos+1;
        if(solvingPos>judgeLen)
            break;
        end
    end
end
%% Solve the difference equations for the five states, find out ...
%% the states that are at risk in the future, and then draw
figure(2);
hold on
yearCur=2010:1:2017;
xlabel('Year/y');
ylabel('Rate/%');
for col=1:5
    plot(yearCur,rateState(:,col),'-','LineWidth',1);
end
legend('Ohio','Kentucky','West␣Virginia','Virginia','Pennsylvania',...
    'Location','northwest');
hold off
figure(3);
hold on;
xlabel('Year/y');
```

```
ylabel('Rate/%');
for col=1:5
    rawStateData=rateState(:,col);
    stateDiff=diff(rawStateData);
    modelState=ar(stateDiff,2,'ls');
    stateHat=predict(modelState,stateDiff);
    stateHat(end+1:end+2)=forecast(modelState,stateDiff,2);
    preState(1)=rawStateData(1);
    for yearState=2:9
        preState(yearState)=rawStateData(yearState-1)+stateHat(yearState-1);
    end
    preState(10)=preState(9)+stateHat(9);
    yearPre=2010:1:2019;
    plot(yearPre,preState,'-.','LineWidth',1);
end
legend('Ohio','Kentucky','West␣Virginia','Virginia','Pennsylvania'...
    ,'Location','northwest');
hold off
%% Find the threshold for the county
figure(4);
maxDistLog=-inf;
for i=1:46
    if badset(i,1)<z1(1,1)&&badset(i,2)<z1(1,2)
        tempDist=sqrt((z1(1,1)-badset(i,1)).^2+(z1(1,2)-badset(i,2)).^2);
        if tempDist>maxDistLog
            maxDistLog=tempDist;
        end
    end
end
axisxLog=(z1(1,1)+z1(1,2)-sqrt((z1(1,1)+z1(1,2))^2-2*(z1(1,1)^2...
+z1(1,2)^2-maxDistLog^2)))/2;
axisx=80-10^(2-axisxLog);
ansPoint=[axisx,axisx];
plot([0,axisxLog,axisxLog,0,0],[0,0,axisxLog,axisxLog,0],'-','LineWidth',1);
hold on;
plot([0,z1(1,1),z1(1,1),0,0],[0,0,z1(1,2),z1(1,2),0],'k:','LineWidth',1);
theta=pi:2*pi/3600:3/2*pi;
Circlex=z1(1,1)+maxDistLog*cos(theta);
Circley=z1(1,2)+maxDistLog*sin(theta);
plot(Circlex,Circley,'r','Linewidth',1);
textx=[real(axisxLog)+0.01,0.01,real(axisxLog)+0.01,real(z(1,1))+0.01...
    ,real(z(1,1))+0.01,real(z(1,1))+0.01,real(z(1,1)-maxDistLog),0.01,0.01];
texty=[real(axisxLog)+0.01,real(axisxLog)+0.02,0.02,0.02,real(z(1,2))...
    -maxDistLog)+0.02,real(z(1,2))+0.02,real(z(1,2))+0.02,real(z(1,2))...
    +0.02,0.02];
textstring={'A(0.5131,0.5131)','B','C','D','E','F','G','H','O'};
text(textx,texty,textstring);
xlabel('-lg(80-x_{2016})+2','fontsize',12);
```

```
ylabel('-lg(80-x_{2017})+2','fontsize',12);
%% Filter counties and years that exceed the critical value of the...
predicted values
exceedCounty2019=[];
exceedCounty2020=[];
for i=1:120
    if(prevData(i,2)>axisx)
        exceedCounty2019=[exceedCounty2019;prevData(i,:)];
    end
    if(prevData(i,3)>axisx)
        exceedCounty2020=[exceedCounty2020;prevData(i,:)];
    end
end
%% Screening collections of counties that are currently in a ...
%% dangerous state, counties that are not currently in danger but...
%% are in danger, and the first two counts
for i=1:size(exceedCounty2019,1)
    if(exceedCounty2019(i,3)<=45||exceedCounty2019(i,3)>=90)
        exceedCounty2019(i,:)=[];
    end
    if (i>=size(exceedCounty2019,1))
        break;
    end
end
for i=1:size(exceedCounty2020,1)
    if(exceedCounty2020(i,2)<=45||exceedCounty2020(i,2)>=90)
        exceedCounty2020(i,:)=[];
    end
    if(i>=size(exceedCounty2020,1))
            break;
    end
end
predSet=union(exceedCounty2019,exceedCounty2020,'rows','sorted');
predSet=predSet(:,1);
badsetPos=1;
curSet=[];
for ii=1:461
    if x(ii,:)==badset(badsetPos,:)
        curSet=[curSet;CountyMat(ii,1)];
        badsetPos=badsetPos+1;
        if(badsetPos>badsetNum)
            break;
        end
    end
end
allPossibleSet=union(predSet,curSet,'sorted');
diffSet=setdiff(predSet,intersect(predSet,curSet));
%% Output data which will be useful in the second question
```

```
xlswrite('DrugRate.xlsx',rateKentuckyCounty,'Sheet1');
xlswrite('DrugRate.xlsx',rateOhioCounty,'Sheet2');
xlswrite('DrugRate.xlsx',ratePennsyivaniaCounty,'Sheet3');
xlswrite('DrugRate.xlsx',rateVirginiaCounty,'Sheet4');
xlswrite('DrugRate.xlsx',rateWVirginiaCounty,'Sheet5');
```

This MATLAB program is used to reorganize the data for Part2

Program 2: `DataReorganization.m`

```
%%
clc,clear;
[ACS2010Num,ACS2010Text]=xlsread(...
    'ACS_10_5YR_DP02\ACS_10_5YR_DP02_with_ann.xlsx','B1:WA466');
[ACS2011Num,ACS2011Text]=xlsread(...
    'ACS_11_5YR_DP02\ACS_11_5YR_DP02_with_ann.xlsx','B1:WA466');
[ACS2012Num,ACS2012Text]=xlsread(...
    'ACS_12_5YR_DP02\ACS_12_5YR_DP02_with_ann.xlsx','B1:WA466');
[ACS2013Num,ACS2013Text]=xlsread(...
    'ACS_13_5YR_DP02\ACS_13_5YR_DP02_with_ann.xlsx','B1:WA466');
[ACS2014Num,ACS2014Text]=xlsread(...
    'ACS_14_5YR_DP02\ACS_14_5YR_DP02_with_ann.xlsx','B1:WA465');
[ACS2015Num,ACS2015Text]=xlsread(...
    'ACS_15_5YR_DP02\ACS_15_5YR_DP02_with_ann.xlsx','B1:WA465');
[ACS2016Num,ACS2016Text]=xlsread(...
    'ACS_16_5YR_DP02\ACS_16_5YR_DP02_with_ann.xlsx','B1:WA465');
ACS2010Num(:,2)=[];ACS2010Text(:,2)=[];ACS2010Text(1,:)=[];
ACS2011Num(:,2)=[];ACS2011Text(:,2)=[];ACS2011Text(1,:)=[];
ACS2012Num(:,2)=[];ACS2012Text(:,2)=[];ACS2012Text(1,:)=[];
ACS2013Num(:,2)=[];ACS2013Text(:,2)=[];ACS2013Text(1,:)=[];
ACS2014Num(:,2)=[];ACS2014Text(:,2)=[];ACS2014Text(1,:)=[];
ACS2015Num(:,2)=[];ACS2015Text(:,2)=[];ACS2015Text(1,:)=[];
ACS2016Num(:,2)=[];ACS2016Text(:,2)=[];ACS2016Text(1,:)=[];
%% Preliminary processing of data, seeking public county...
%% and public characteristics
%% Delete each table of non-public counties
ACS2010Num(372,:)=[];ACS2010Text(373,:)=[];
ACS2011Num(372,:)=[];ACS2011Text(373,:)=[];
ACS2012Num(372,:)=[];ACS2012Text(373,:)=[];
ACS2013Num(372,:)=[];ACS2013Text(373,:)=[];
%% Output results
%% Output results to EXCEL
xlswrite('ACSNum.xlsx',ACS2010Num,1);
xlswrite('ACSNum.xlsx',ACS2011Num,2);
xlswrite('ACSNum.xlsx',ACS2012Num,3);
xlswrite('ACSNum.xlsx',ACS2013Num,4);
xlswrite('ACSNum.xlsx',ACS2014Num,5);
xlswrite('ACSNum.xlsx',ACS2015Num,6);
xlswrite('ACSNum.xlsx',ACS2016Num,7);
```

```
xlswrite('ACSText.xlsx',ACS2010Text,1);
xlswrite('ACSText.xlsx',ACS2011Text,2);
xlswrite('ACSText.xlsx',ACS2012Text,3);
xlswrite('ACSText.xlsx',ACS2013Text,4);
xlswrite('ACSText.xlsx',ACS2014Text,5);
xlswrite('ACSText.xlsx',ACS2015Text,6);
xlswrite('ACSText.xlsx',ACS2016Text,7);
```

This MATLAB program is used to solve Part2

Program 3: `Problem2.m`

```
%%
clc,clear;
[~,textData1]=xlsread('ACSText.xlsx','Sheet1');
numData1=xlsread('ACSNum.xlsx','Sheet1');
[~,textData2]=xlsread('ACSText.xlsx','Sheet2');
numData2=xlsread('ACSNum.xlsx','Sheet2');
[~,textData3]=xlsread('ACSText.xlsx','Sheet3');
numData3=xlsread('ACSNum.xlsx','Sheet3');
[~,textData4]=xlsread('ACSText.xlsx','Sheet4');
numData4=xlsread('ACSNum.xlsx','Sheet4');
[~,textData5]=xlsread('ACSText.xlsx','Sheet5');
numData5=xlsread('ACSNum.xlsx','Sheet5');
[~,textData6]=xlsread('ACSText.xlsx','Sheet6');
numData6=xlsread('ACSNum.xlsx','Sheet6');
[~,textData7]=xlsread('ACSText.xlsx','Sheet7');
numData7=xlsread('ACSNum.xlsx','Sheet7');
%% Create a three-dimensional matrix textData and numData
countyData=numData1(:,1);
nameData=textData4(1,2:597)';
textData1(1,:)=[];textData2(1,:)=[];textData3(1,:)=[];textData4(1,:)=[];
textData5(1,:)=[];textData6(1,:)=[];textData7(1,:)=[];
textData(:,:,1)=textData1(:,2:597);numData(:,:,1)=numData1(:,2:597);
textData(:,:,2)=textData2(:,2:597);numData(:,:,2)=numData2(:,2:597);
textData(:,:,3)=textData3(:,2:597);numData(:,:,3)=numData3(:,2:597);
textData(:,:,4)=textData4(:,2:597);numData(:,:,4)=numData4(:,2:597);
textData(:,:,5)=textData5(:,2:597);numData(:,:,5)=numData5(:,2:597);
textData(:,:,6)=textData6(:,2:597);numData(:,:,6)=numData6(:,2:597);
textData(:,:,7)=textData7(:,2:597);numData(:,:,7)=numData7(:,2:597);
%% Create a Data Structure
for page=1:7
    tempNumPageData=-ones(463,149);
    vecColNum=1:4:593;
    tempNumPageData=numData(:,vecColNum,page);
    for col=1:149
        for row=1:463
            if isnan(tempNumPageData(row,col))
                tempNumPageData(:,col)=-ones(463,1);
```

```
                break;
            end
        end
    end
    numOrgdData(:,:,page)=tempNumPageData;
    tempPerPageData=-ones(463,149);
    vecColPer=3:4:595;
    tempPerPageData=numData(:,vecColPer,page);
    for col=1:149
        for row=1:463
            if tempPerPageData(row,col)>100||isnan(tempPerPageData(row,col))
                tempPerPageData(:,col)=-ones(463,1);
                break;
            end
        end
    end
    perOrgdData(:,:,page)=tempPerPageData;
end
%% Spline spline interpolation for each column
year=1:7;
for col=1:149
    for row=1:463
        tempNumVecy=reshape(numOrgdData(row,col,:),1,7);
        tempPerVecy=reshape(perOrgdData(row,col,:),1,7);
        indexNumVecx=1:7;
        indexPerVecx=1:7;
        NumId=(tempNumVecy==-1|tempNumVecy==0);
        PerId=(tempPerVecy==-1|tempPerVecy==0);
        tempNumVecy(NumId)=[];
        indexNumVecx(NumId)=[];
        tempPerVecy(PerId)=[];
        indexPerVecx(PerId)=[];
        if length(tempNumVecy)>1
            resNumVec(1,1,:)=reshape(interp1(indexNumVecx,tempNumVecy...
                ,year,'spline'),1,1,7);
        elseif length(tempNumVecy)==1 %Keep this constant
            resNumVec(1,1,:)=reshape(tempNumVecy(1)*ones(7,1),1,1,7);
        else
            resNumVec(1,1,:)=reshape(zeros(1,7),1,1,7);
        end
        if length(tempPerVecy)>1
            resPerVec(1,1,:)=reshape(interp1(indexPerVecx,tempPerVecy...
                ,year,'spline'),1,1,7);
        elseif length(tempPerVecy)==1 %Keep this constant
            resPerVec(1,1,:)=reshape(tempPerVecy(1)*ones(7,1),1,1,7);
        else
            resPerVec(1,1,:)=reshape(zeros(1,7),1,1,7);
        end
```

```
        numOrgdData(row,col,:)=resNumVec;
        perOrgdData(row,col,:)=resPerVec;
    end
end
%% Pearson product difference correlation coefficient calculation
rateDrug=xlsread('DrugRate.xlsx','Sheet1','A1:DP9');
rateDrug=[rateDrug,xlsread('DrugRate.xlsx','Sheet2','A1:CJ9')];
rateDrug=[rateDrug,xlsread('DrugRate.xlsx','Sheet3','A1:BO9')];
rateDrug=[rateDrug,xlsread('DrugRate.xlsx','Sheet4','A1:EA9')];
rateDrug=[rateDrug,xlsread('DrugRate.xlsx','Sheet5','A1:BC9')];
rateDrugpos=1;
perPearRes=[];
perPearResTmp=[];
finalset=intersect(countyData',rateDrug(1,:));
finalsetId=1;
for i=1:463
    while i<=463&&finalsetId<=460&&countyData(i,1)~=finalset(1,finalsetId)
        i=i+1;
    end
    while rateDrugpos<=461&&finalsetId<=460&&rateDrug(1,rateDrugpos)...
            ~=finalset(1,finalsetId)
        rateDrugpos=rateDrugpos+1;
    end
    perPearResTmp=[];
    perPearMat=reshape(perOrgdData(i,:,:),149,7);
    perPearMat=perPearMat';
    for j=1:149
        perPearResTmp(1,j)=corr(rateDrug(2:8,rateDrugpos),perPearMat(:,j));
    end
    perPearResTmp=perPearResTmp';
    perPearRes=[perPearRes,perPearResTmp];
    rateDrugpos=rateDrugpos+1;
    finalsetId=finalsetId+1;
    if i==463
        break;
    end
end
%% Output result
perPearRes=[finalset;perPearRes];
xlswrite('Percentage_Pearson_Result.xlsx',perPearRes,1);
nameData=nameData(vecColPer,:);
xlswrite('Percentage_Pearson_Result.xlsx',nameData,2);
%% Calculate and filter features that meet high positive...
%% and negative correlation conditions
perPearResPosi=[perPearRes(1,:);perPearRes(2:150,:)>0.3];
perPearResNegi=[perPearRes(1,:);perPearRes(2:150,:)<-0.3];
vecPos=sum(perPearResPosi==1,2);
vecNeg=sum(perPearResNegi==1,2);
```

```
index=1:149;
vecPos(1,:)=[];vecNeg(1,:)=[];
vecPos=[index',vecPos];vecNeg=[index',vecNeg];
vecPos=sortrows(vecPos,-2);vecNeg=sortrows(vecNeg,-2);
%%
delIndexPos=vecPos(:,2)<150;
delIndexNeg=vecNeg(:,2)<150;
vecPos(delIndexPos,:)=[];
vecNeg(delIndexNeg,:)=[];
lenPos=size(vecPos,1);
lenNeg=size(vecNeg,1);
delIndexPos2=[];delIndexNeg2=[];
for i=1:lenPos
    for j=1:lenNeg
        if(vecPos(i,1)==vecNeg(j,1))
            if(vecPos(i,2)>=vecNeg(j,2))
                delIndexNeg2=[delIndexNeg2;j];
            elseif(vecPos(i,2)<vecNeg(j,2))
                delIndexPos2=[delIndexPos2;i];
            end
        end
    end
end
vecPos(delIndexPos2,:)=[];
vecNeg(delIndexNeg2,:)=[];
%%
matPos=zeros(29,462);
matNeg=zeros(43,462);
perPearRes(isnan(perPearRes))=0;
for row=1:29
    matPos(row,1:460)=perPearRes(vecPos(row,1)+1,:);
    matPos(row,461)=mean(perPearRes(vecPos(row,1)+1,:));
    matPos(row,462)=median(perPearRes(vecPos(row,1)+1,:));
end
for row=1:43
    matNeg(row,1:460)=perPearRes(vecNeg(row,1)+1,:);
    matNeg(row,461)=mean(perPearRes(vecNeg(row,1)+1,:));
    matNeg(row,462)=median(perPearRes(vecNeg(row,1)+1,:));
end
matPos=matPos';
matNeg=matNeg';
xlswrite('matPos_And_Neg.xlsx',matPos,1);
xlswrite('matPos_And_Neg.xlsx',matNeg,2);
%% Output calculation result
xlswrite('Pos_and_Neg_Charac_Table.xlsx',nameData,1);
xlswrite('Pos_and_Neg_Charac_Table.xlsx',vecPos,2);
xlswrite('Pos_and_Neg_Charac_Table.xlsx',vecNeg,3);
```

This MATLAB program is used to draw the Percentage of Opioid Use Iteration Graph

Program 4: `Problem2.m`

```
%%
clc,clear;
x=-0.2:0.001:3.5;
a=-0.7;
yQua=a*x.^2+(1-2*a).*x;
plot(x,yQua,'k-','LineWidth',1);
hold on
axis equal
yLin=x;
plot(x,yLin,'-','LineWidth',1);
x=1;y=0;deltax=0;deltay=a*x^2+(1-2*a)*x;
quiver(x,y,deltax,deltay+0.19,'r','LineWidth',0.8);
x=x+deltax;y=y+deltay;deltax=y-x;deltay=0;
quiver(x,y,deltax+0.08,deltay,'r','LineWidth',0.8);
x=x+deltax;y=y+deltay;deltax=0;deltay=a*x^2+(1-2*a)*x-y;
quiver(x,y,deltax,deltay+0.04,'r','LineWidth',0.8);
x=x+deltax;y=y+deltay;deltax=y-x;deltay=0;
quiver(x,y,deltax+0.04,deltay,'r','LineWidth',0.8);
x=x+deltax;y=y+deltay;deltax=0;deltay=a*x^2+(1-2*a)*x-y;
quiver(x,y,deltax,deltay-0.01,'r','LineWidth',0.8);
x=x+deltax;y=y+deltay;deltax=y-x;deltay=0;
quiver(x,y,deltax-0.006,deltay,'r','LineWidth',0.8);
x=x+deltax;y=y+deltay;deltax=0;deltay=a*x^2+(1-2*a)*x-y;
quiver(x,y,deltax,deltay+0.004,'r');
xux=2.*ones(101,1);xuy=0:0.02:2;
plot(xux,xuy,'k:','LineWidth',1);
plot(2,2,'k.','MarkerSize',9);
plot(0,0,'k.','MarkerSize',9);
plot(2,0,'k.','MarkerSize',9);
plot(1,0,'r.','MarkerSize',9);
point1x=3;point1y=a*point1x^2+(1-2*a)*point1x;
text(point1x-0.3,point1y,'y=h^*(x)','fontsize',12);
point2x=2.3;point2y=2.3;
text(point2x+0.08,point2y,'y=x','fontsize',12);
point3x=0.15;point3y=0.085;
text(point3x,point3y,'x^*_1','fontsize',12);
point4x=2.03;point4y=0.085;
text(point4x,point4y,'x^*_2','fontsize',12);
```