For office use only	Team Control Number	For office use only
T1 ________________	**1923231**	F1 ________________
T2 ________________		F2 ________________
T3 ________________	Problem Chosen	F3 ________________
T4 ________________	**C**	F4 ________________

2019
MCM/ICM
Summary Sheet

Modified Lotka-Volterra Model to Analyse and Predict Opioid Use

In order to identify the locations where specific opioid use started and describe the patterns and characteristics of opioid and heroin incidents spread, our team analyze, process and construct models on the drug reports and census profiles. Then, according to the applicable range of the model, we make suggestions to the government in the short and long term respectively.

Specifically, in the data process, we use the K-means algorithm to combine geographically close counties into one cluster. The benefit of clustering is to reduce the influence of the volatility of certain countys drug reports, while maintaining the accuracy of location identification. For census data, every variable can be classified into certain socio-economic factor by its name (e.g. ANCERSTRY including estimate, estimate margin of error, percent, etc.). Since we have no idea whether the estimate number or the percent number provides better information for our modeling, it's reasonable to use the Entropy Weight Method to combine variables into factors for further use.

For Part 1, we build the Logistic Model to identify the start time of specific opioid use in each cluster. We conclude that the Cluster-8 in Pennsylvania first started using heroin drugs and Cluster-3 in Kentucky first use synthetic opioids. However, the Logistic Model doesn't take the spread of heroin incidents into account. Therefore, we add the interaction of different clusters into the LM to get the Lotka-Volterra Model and forecast the trends of next 5 years. We use the upper bound calculated from LM as the drug identification threshold levels, and find Cluster-2 most in West Virginia and Cluster-7 in Pennsylvania will breakthrough threshold in 2018 and 2021. It's urgent for the government to adopt effective measures.

For Part 2, after combining variables into factors, we find certain factors have a strong correlation with the drug reports in different clusters. In order to select the best influence factors for each cluster, a correlation matrix is calculated and contributes to the growth in opioid use and addiction are partly explained. We build linear regression model to regress the natural change rate on factors and modify the change rate in the Lotka-Volterra by taking the weighted average of two results. For the simple Lotka-Volterra Model having chaos phenomenon, the Modified Lotka-Volterra is more stable to forecast the next 50 years trends.

For Part 3, combining the insights from the model, we identify the short-term and long-term policies respectively. For the short term, our main goal is to reduce the natural change rate of each clusters in the Lotka-Volterra. Hence, we suggest the government in each state adopt tough policing policies to crack down on the illicit sale of heroin. For the long term, our main goal is to change the socio-factors in the linear regression model. We suggest the government to increase the regional education attainment and decrease the unmarried rate, which will ultimately reduce the opioid users in the long run.

Finally, for the Chief Administrator, we present a memo to summarize our insights and results identified during model construction.

Contents

1 Introduction

1.1 Background

Opioids, no matter synthetic ones or non-synthetic ones, play an significant role in our daily life. The proper utilization of opioids helps thousands of people get relieved from their pains. However, misuse of them can really destroy the society. Therefore, it is essential and urgent to gain a better understanding of opioids' spread and influencing factors.

Recent years have witnessed the tremendous growth of opioids cases. How this trend will develop remains an national concern. With data of annual drug use and socio-economic factors in five U.S. states over the past few years, we perform data mining and construct a model to figure out the pattern of opioids use and what contributes to it.

1.2 Problem Statement and Analysis

The problem is about the opioid crisis in the U.S. society, offering us a set of data ranging from 2010 to 2017. Our main task is to build models that enable us to identify the locations where specific opioid use started and describe the patterns and characteristics of opioid incidents spread.

First, this is a typical data problem, so we should carefully preprocess the data provided. In the MCM_NFLIS Data, there are 461 counties belongs to 5 states. For most of the counties, the drug reports are fluctuant, even with large amount of missing data. Besides if we combined all the drug reports of certain state, the start of specific opioid use of one county may be covered up by opioid use of other counties. Therefore, in order to increase the stability of patterns without losing the accuracy, we use the K-means algorithm to combine 461 counties into 10 clusters for our further modeling.

While in the U.S. Census socio-economic data, there are about nearly 600 variables ranging from 2010 to 2017, which is too large to use in the model. We notice that every variable can be classified into certain category by its name, like HOUSEHOLDS BY TYPE, MARITAL STATUS and so on. It's reasonable to use Entropy Weight Method to combine 596 variables into 15 factors.

Second, in the early development stage of the heroin incidents, we think such incidents can be treated as individual events and the interaction between different states has not yet formed. So we use the Logistic Model to describe the initial trend of heroin incidents. Through the Logistic Model, we can also get the upper bound of each cluster, which will be used in the next part to determine the threshold levels.

Third, with the spread of incidents, states may start interact with each other and a stable environmental condition for the Logistic Model is no longer met. Therefore, we use the Lotka-Volterra Model to study the spread and characteristics of the reported incidents and forecast the future trend of the heroin incidents. Then we use the Modified Lotka-Volterra Model by adding different factors.

Finally, the Modified Lotka-Volterra Model can provide us both the relationships between clusters and the influence of different factors. By taking all these factors into account, we can identify a possible strategy for countering the opioid crisis and test the effectiveness of this strategy by our model.

The models and their relationships with problems are shown in the folloing figure:

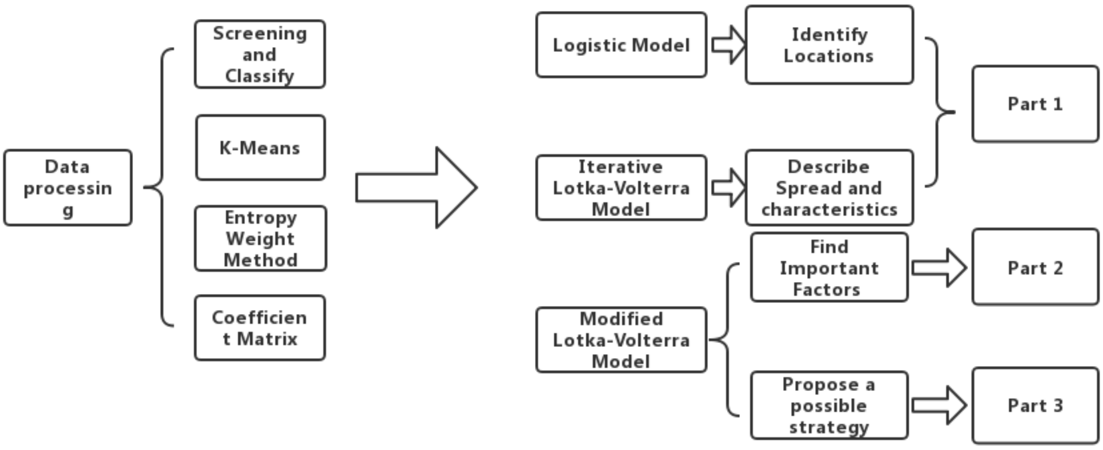


Fig. 1: Overall Flow Graph

2 Assumptions and Notations

2.1 Assumptions

Given the lack of necessary data and limitation of our knowledge, we made the following assumptions to help us perform modeling. These assumptions are the premise for our subsequent analysis:

- The data after our process are correct and robust for our further analysis.
- For the early development stage of the heroin accidents, we assume the incidents growth rate of certain cluster is determined only by itself and the environmental remains stable. So we use the Logistic Model to describe the initial trend of heroin incidents.
- The policy of each state will not change in the future. Because policy change will greatly influence the interaction between clusters and the influence of factors, it's essential for us to assume the policy remain unchanged.
- The influence of the factors to heroin incidents remain unchanged and the factors of each clusters grow linearly.

2.2 Notations

The primary notations used in this paper are listed in Table 1.

Table 1: Notations

Symbol	Definition
$u_k(t)$	Drug reports of Cluster-k at time t
K	Upper Bound of Drug Use
X_t	Factors calculated by Entropy Weight Method
a_k	The natural growth rate of Cluster-k
b_{jk}	The influence factors of Cluster-k to the nature growth rate of Cluster-j

3 Data Processing

We are provided with two types of data, the number of drug reports in categories and socio-economic factors. Both of them are specific to counties. The original data contains a lot of redundant and invalid items, which can seriously affect the accuracy and versatility of our model. Thereby, we first apply some data processing techniques before we construct the model.

3.1 Missing Data Processing

In the worksheet about socio-economic factors, there are many cells that contain no valid information at all. We delete columns and rows which contains a lot of useless cells. For other columns and rows, when invalid items appear, we replace it with the average value of the adjacent numbers.

3.2 Combine variables into a Socio-economic Factor

For every year from 2010 to 2016, there are a list of variables about one single socio-economic factor. For instance, factor ANCERSTRY includes estimate, estimate margin of error, percent, etc. Those variables all reflect some information of the target factor in some aspects, hence how to rank their importance matters.

In information theory, entropy represents the information a sequence contains[3]. So Entropy Weight Method (EWM), which gives each index a weight according to their entropy, is chosen to measure the significance of each index. Before that, we normalize the data given using the following equation.

$$y_i = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \tag{1}$$

where x_i and y_i denotes the original sequence and processed sequence.

We apply EWM in the way addressed as follows (Using ANCERSTRY as an example):

- Define Estimate, Estimate margin of error, Percent, Percent margin of error of ANCERSTRY is defined as y_j (y_1 means Estimate...). Compute entropy for y_j: $E_j = -\frac{\sum p_i lnp_i}{ln(n)}$, where $p_i = \frac{y_{ji}}{\sum_i y_{ji}}$, E_j represents the entropy for each index.
- Compute weight for Estimate, Estimate margin of error, Percent, Percent margin of error of ANCERSTRY: $W_j = \frac{E_j}{\sum E_j}$, where W_j represents the weight for each index.
- Combine variables using the weight: $X_i = \sum_j W_j * y_{ji}$

So far, we have combined several variables of a factor into one X, and the same can be done to other factors. The newly generated sequence will be reserved for further analysis.

3.3 Cluster Counties of Five States

Since we have detailed number of drug reports, it is not hard to fit a curve to observe the trend of opioid use for each state and for each county. However, we consider this method inappropriate. On the one hand, the sample for only one county is normally fluctuating irregularly, i.e. drug reports of heroin in ACCOMACK, VA was 2, 38, 6 during 2012-2014. It is unnecessary and meaningless to bulid a model when facing such fluctuation. On the other hand, given the large number of all kinds of opioids, even one type change significantly, it is not obvious when viewed as the level of the whole state.

Considering that state and county are not suitable for our research, a unit in between is introduced. We put counties on a two-dimensional plane and regard them as nodes. Then, k-means clustering method[1], which aims to partion nodes into k clusters minimizing the within-cluster sum of squares, was performed. In our case, we take $k = 10$. The distance of two counties can be computed with their latitude and longtitude retrieved from U.S. Census Bureau[2]. The clusters we derived are labeled as Cluster-1, Cluster-2, ..., Cluster-10.

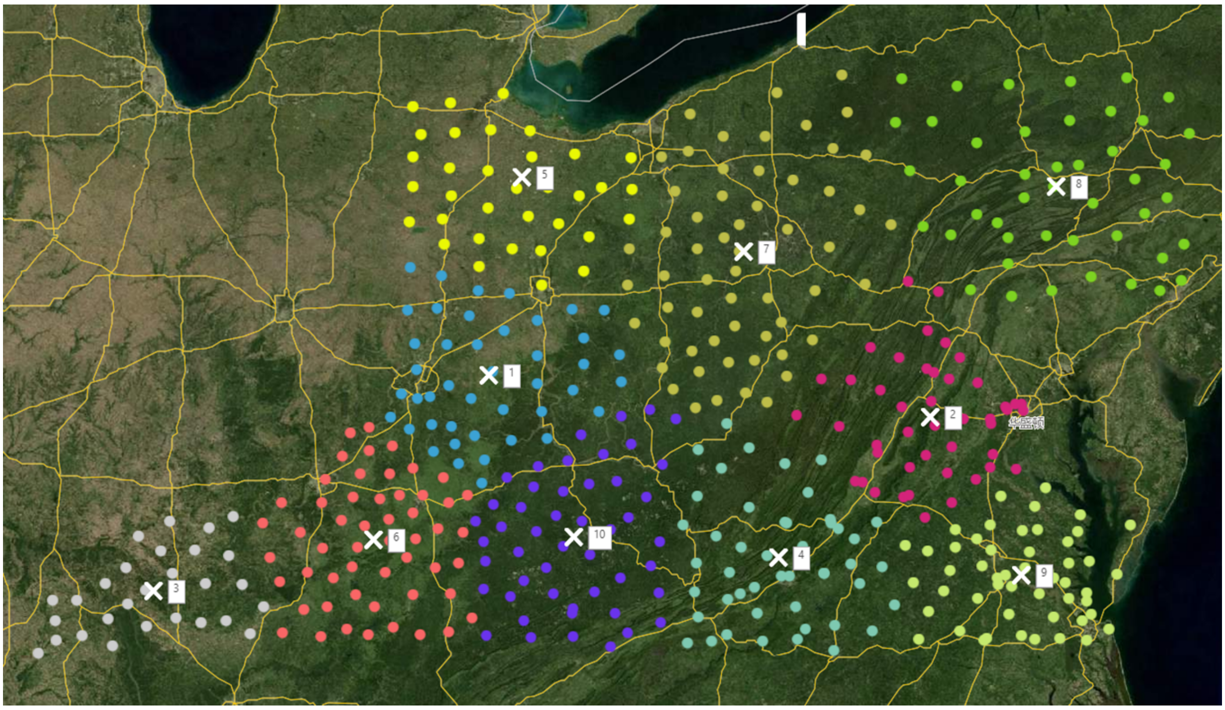


Fig. 2: Clusters got after k-means, points with the same color belong to one cluster

3.4 Data Screening – Drug Classification

In order to describe the dissemination and characteristics of the synthetic opioids and heroin spread sperately, we need to screen the information in MCM_NFLIS_Data.xlsx. By consulting Wikipedia, DrugBank[4] and other websites, provided drugs are roughly classified into synthetic opioids, semi-synthetic opioids, and natural ingredients(Table 2). In view of the topic requirements, we excluded the natural ingredients and drugs other than heroin in semi-synthetic opioids, leaving the drug reports of heroin and synthetic opioids only. As for the synthetic opioid, we find that many drugs have been used in small amounts over the years compared to those used in large numbers. And some drugs only start to appear in the past two years, so we can just ignore them. Finally, we leave about 15 kinds of synthetic drugs that have appeared since 2010 for further research.

Table 2: Opioids and Their Kinds

Kind	Specific Opioids
Natural Ingredients	Codeine, Morphine, Opium
Semi-synthetic Opioids	Dihydrocodeine, Heroine, Opiates, Oxycodone, Oxymorphone, etc.
Synthetic Opioids	ANPP, Buprenophine, Butorphanol, Fentanyl, Meperidine, Methorphan, etc

4 Model Construction

4.1 Identify Locations using Logistic Model

4.1.1 Model Establishment

With the k-means cluster analysis, we acquire data about the synthetic opioid and heroin events in ten clusters. It is clear to identify that the drug use levels in most areas are increasing. If drug use continues to grow at such speed, it will bring about numerous negative effects to the society when it exceeds an certain level, as mentioned in the Problem Background. Therefore, government's relevant departments, including the Drug Enforcement Administration(DEA), will take corresponding measures to control the growing trend of drug use. However, when consumptions reach a high level, the growing rate in usage will reduce. And it is gradually increasing when consumptions are low.

The regular drug-use pattern is similar to the Logistic model of population growth, in which the natural growth rate of the population will be limited by the population. Similarly, the drug-use will be limited by many restrictions including social factors such as government control and natural factors such as population growth. Assuming that drug-use level could be affected by these factors and thus having a ceiling, it can be considered that the growth trend of synthetic opioid and heroin events are subordinated to the Logistic model. Here's the Logistic model:

$$\begin{gathered}\frac{du(t)}{dt} = r(1 - \frac{u(t)}{K})u(t) \\ u(t_0) = u_0\end{gathered} \tag{2}$$

where $u(t)$ represents the drug-use amount as time t, K represents the upper bound of the drug-use level and r means an natrual growth rate of the drug-use amount. The model can be understood that there is no strict control and the drug is alluring at the early stage, which results in the quickly increase of the drug use. As some departments begin to pay attention to this phenomenon and implement policies to control it, the growth rate of drug use begins to decrease. When the drug-use level is close to the allowable upper bound, $\frac{(1-Q)}{K} \approx 0$. At this point, the drug-use level will no longer increase. Considering that our model is discrete in time, we replace $\frac{du(t)}{dt}$ with $N(t+1) - u(t)$.

The following equations gives the solution of logistic regression (Equation (2)).

$$u(t) = \frac{K}{1 + Ce^{-r(t-t_0)}} \tag{3}$$

$$C = \frac{K - u_0}{u_0} \tag{4}$$

The above mathematical formulas and model can be established by programing. We utilize the data of ten clusters to build the model and acquire the logistic regressing curves.

4.1.2 Model Application and Problems Solving

In order to test the effectiveness of our model, we first put the total drug reports into the model for regressing and fitting, and get the following results. From the above clusters mentioned, the drug-use level in some areas is decreasing or fluctuating, we find three unsuitable clusters(Cluster-4, Cluster-5, Cluster-10) in Fig.3 and we think that the trend of these cluster are not suitable in our model. That is to say, other models may be needed for their prediction, which is not considered currently.

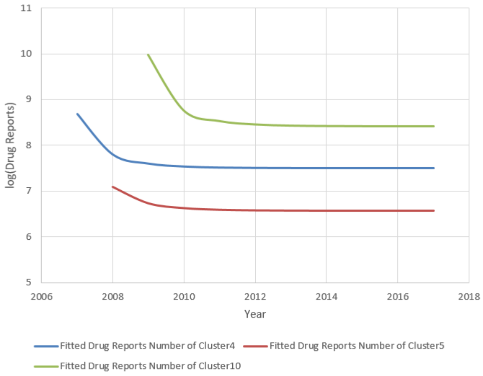


Fig. 3: Total Drug Reports Curves of Unsuitable Clusters

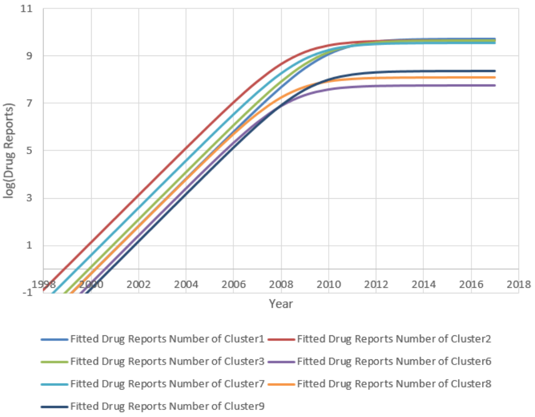


Fig. 4: Total Drug Reports Curves of Suitable Clusters

According to fitted curves of the remaining data, we obtain the curves of drug-use level for 20 years from 2000 to present. Then, we look for intersections of every curve and the time axis. Fig.4 describes most clusters' characteristics of total drug use.

We can clearly see that Cluster-2, where drug cases have started since 1999, is earlier than any other six clusters and this is exactly our target to find in specific opioids in most clusters. Querying the map of U.S., we know that Cluster-2 represents east of West Virginia area.

Therefore, we further make use of this model to analyze the specific usage of heroin and synthetic opioids screened and acquire the following results.

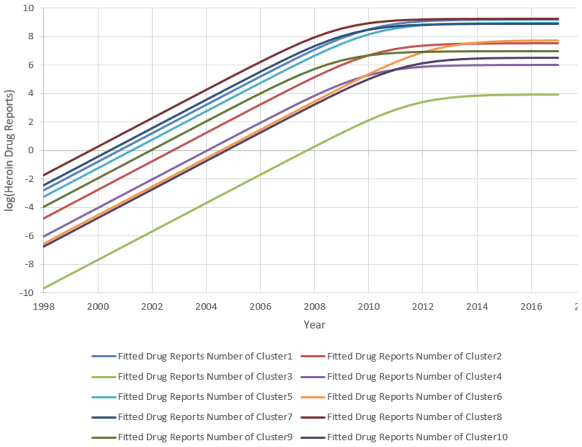


Fig. 5: Drug Reports for Heroin

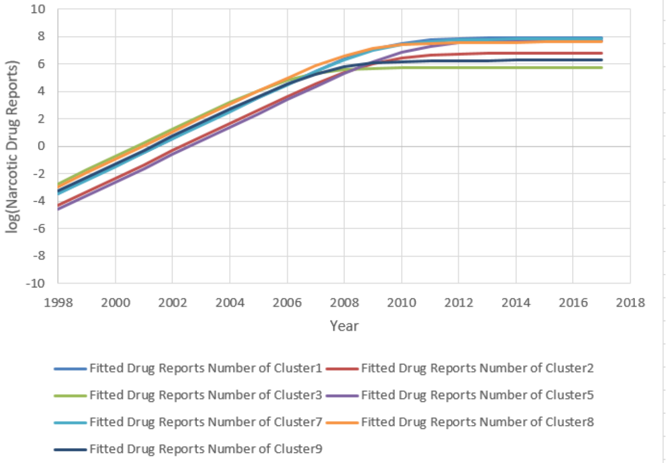


Fig. 6: Drug Reports for Syntheic Opioids

As we can see from Fig.5, the Cluster-8(Pennsylvania) in the ten clusters has already started using heroin drugs before 2000. It can be considered that counties which belonging to Cluster-8 could be the locations where heroin started to appear and use.

In Fig.6, it can be seen that Cluster-1, 3, 7, 8, and 9 all started using drugs around 2001. Because the time is not particularly precise, they may all be the areas where synthetic opioids were used earlier. According to our model, we believe that Kentucky represented by Cluster-3 is the area where the synthetic opioids use have started earlier.

What's more, we use the logistic discrete equation to solve the coefficients of the regression, and obtain the upper bound of drug-use in each region based on this model. Results presented presented in the following table will continue to be used in subsequent models.

Table 3: Upper Bound(UB) for Clusters

Cluster	1	2	3	4	5
UB	44101	22261	1909	602	2840
Cluster	6	7	8	9	10
UB	2840	19022	3330	4318	6084

4.2 Describing the spread and characteristics using Iterative Lotka-Volterra Model

4.2.1 Introduction to Prey-Predator Model

In the last section, we construct Logistic Model to identify the possible locations where specific opioid use might have started. However, as the number of drug reports approaches the carrying capacity u_{max}, the Logistic Model become smooth and can't provide more information about the changing trends. Therefore, in this section, we modify the dynamic Lotka-Volterra equations to model the interactions between different clusters.

The Lotka-Volterra equations are known as the predator-prey equations in a biological system of two (or more) species. In fact, the Lotka-Volterra equations can describe not only the predation relationships between species, but also the competition, parasitism even mutualism relationships. And we think the spread of the reported synthetic opioid and heroin incidents may have a mixed features, which meets the application conditions.

For simplification, let S_1 and S_2 denote, respectively, Species 1 and Species 2. Let $u_1(t)$ and $u_2(t)$ denote, respectively, the measurement of S_1 and S_2 at time t. Let a_1 and a_2 denote the natural growth fate of S_1 and S_2, and b_1 and b_2, respectively, the influence factor of S_1 to the natural growth rate of S_1 and S_2. Similarly, here goes the c_1 and c_2. The Lotka-Volterra equations are non-linear, first-order differential equations defined as follows:

$$u_1'(t) = u_1(t)(a_1 + b_1u_1(t) + c_1u_2(t)) \tag{5}$$

$$u_2'(t) = u_2(t)(a_2 + b_2u_2(t) + c_2u_2(t)) \tag{6}$$

4.2.2 Model Establishment

Next, we extend and apply the above equations to model interactions between the 10 clusters. Let $u_k(t)$ denote the Cluster-k, where k = 1,2,,10. Let a_k denote the change rate of Cluster-k itself, and b_{kj} denote the influence of Cluster-j to Cluster-k. The equations can be written as

$$\begin{gathered} \frac{du_k(t)}{dt} = u_k(t)(a_k + \sum_{j=1}^{9} b_{jk}u_k(t)) \\ k = 1, 2, .., 9 \end{gathered} \tag{7}$$

$$\begin{pmatrix} \frac{du_1(t)}{dt} \\ \frac{du_2(t)}{dt} \\ \vdots \\ \frac{du_9(t)}{dt} \end{pmatrix} = \begin{pmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_9(t) \end{pmatrix} \otimes \begin{pmatrix} a_1 & b_{11} & \dots & b_{19} \\ \vdots & \vdots & \ddots & \vdots \\ a_9 & b_{91} & \dots & b_{99} \end{pmatrix} * \begin{pmatrix} 1 \\ u_1(t) \\ u_2(t) \\ \vdots \\ u_9(t) \end{pmatrix} \tag{8}$$

For the ease of calculation and prediction, we modify the equation as follows. Let $\bar{\boldsymbol{U}}(\boldsymbol{t})$ denote the vector $[\boldsymbol{du_k(t)/dt}]^{\boldsymbol{T}}_{\boldsymbol{1*n}}$, $\boldsymbol{U(t)}$ denote the vector $[\boldsymbol{u_k(t)}]^{\boldsymbol{T}}_{\boldsymbol{1*n}}$, $\boldsymbol{A}$ denote the matrix $[\bar{\boldsymbol{a}}|\bar{\boldsymbol{B}}]$ where $\bar{\boldsymbol{a}} = [\boldsymbol{a_k}]^{\boldsymbol{T}}_{\boldsymbol{1*n}}$, $\bar{\boldsymbol{B}} = [\boldsymbol{b_{kj}}]_{\boldsymbol{n*n}}$.

$$\bar{\boldsymbol{U}}(\boldsymbol{t}) = \boldsymbol{U(t)} \otimes \boldsymbol{A} \times [\boldsymbol{1}|\boldsymbol{U(t)^T}]^{\boldsymbol{T}} \tag{9}$$

We notice the time parameter is discrete. In order to make the equation iterated, we transformed equations as follows:

$$\begin{pmatrix} \frac{1}{u_1(t)}\frac{du_1(t+1)}{dt} \\ \frac{2}{u_2(t)}\frac{du_2(t+1)}{dt} \\ \vdots \\ \frac{3}{u_3(t)}\frac{du_3(t+1)}{dt} \end{pmatrix} = \begin{pmatrix} \frac{u_1(t+1)}{u_1(t)} - 1 \\ \frac{u_2(t+1)}{u_2(t)} - 1 \\ \vdots \\ \frac{u_9(t+1)}{u_9(t)} - 1 \end{pmatrix} = \begin{pmatrix} a_1 & b_{11} & \dots & b_{19} \\ \vdots & \vdots & \ddots & \vdots \\ a_9 & b_{91} & \dots & b_{99} \end{pmatrix} * \begin{pmatrix} 1 \\ u_1(t) \\ u_2(t) \\ \vdots \\ u_9(t) \end{pmatrix} \tag{10}$$

$$\begin{pmatrix} \frac{u_1(t+1)}{u_1(t)} \\ \frac{u_2(t+1)}{u_2(t)} \\ \vdots \\ \frac{u_9(t+1)}{u_9(t)} \end{pmatrix} = \begin{pmatrix} a_1 & b_{11} & \dots & b_{19} \\ \vdots & \vdots & \ddots & \vdots \\ a_9 & b_{91} & \dots & b_{99} \end{pmatrix} * \begin{pmatrix} 1 \\ u_1(t) \\ u_2(t) \\ \vdots \\ u_9(t) \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{11}$$

$$\begin{pmatrix} u_1(t+1) \\ u_2(t+1) \\ \vdots \\ u_9(t+1) \end{pmatrix} = \begin{pmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_9(t) \end{pmatrix} \otimes \left\{ \begin{pmatrix} a_1 & b_{11} & \dots & b_{19} \\ \vdots & \vdots & \ddots & \vdots \\ a_9 & b_{91} & \dots & b_{99} \end{pmatrix} * \begin{pmatrix} 1 \\ u_1(t) \\ u_2(t) \\ \vdots \\ u_9(t) \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right\} \tag{12}$$

And we get the final iterative equation as follows:

$$\boldsymbol{U(t+1)} = \boldsymbol{U(t)} \otimes \{\boldsymbol{A} * [\boldsymbol{1}|\boldsymbol{U(t)}^T]^T + 9 * [1]\} = \boldsymbol{g(U(t))} \tag{13}$$

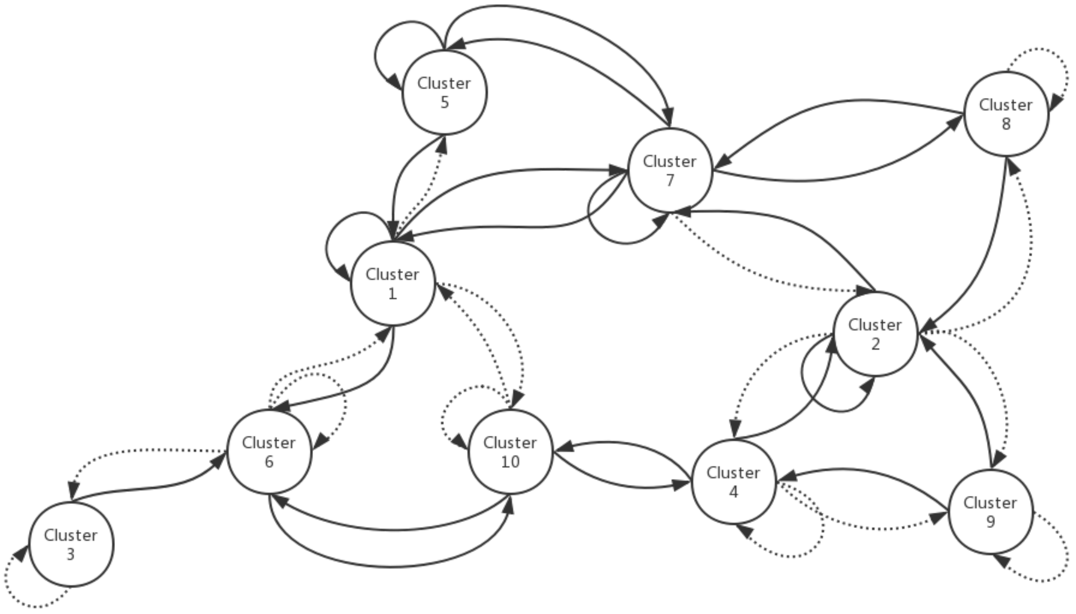
In order to get the iterative matrix $\boldsymbol{A}$, we use the least square method to estimate. For Cluster-i, the vector $[a_i, b_{i1}, ..., b_{i9}]$ has 10 parameters, while we only have 8 equations ranging from 2010 to 2017, which cant meet the condition of least square $n > k$. Therefore, we use the interpolation method to increase sample points. For simplicity, we take the median of any two adjacent sample points as the interpolation value.

$$\boldsymbol{U}(\boldsymbol{t} + \frac{1}{2}) = \frac{\boldsymbol{U}(\boldsymbol{t}+1) + \boldsymbol{U}(\boldsymbol{t})}{2} \tag{14}$$

4.2.3 Result Analysis

After implementing the above mathematical model by programming, we obtain the coefficient matrix of regression equation, which represents the effects of clusters self-influence and mutual influence. According to the positive or negative values of the matrix, we can conclude the spread feature of the drug usage over time.

The positive coefficient value indicates the influence is to promote the drug-use level, contrary to the negative value which represents the impact of growth restricting. So we have the following relationship network figure, where the solid line indicates that the effect is positive and the dotted line indicates that the effect is negative.



--→ Negative Effect
⟶ Positive Effect

Fig. 7: Clusters' Self-influence and Mutual Influence

We draw the total drug reports for clusters in the next 5 years as follows:

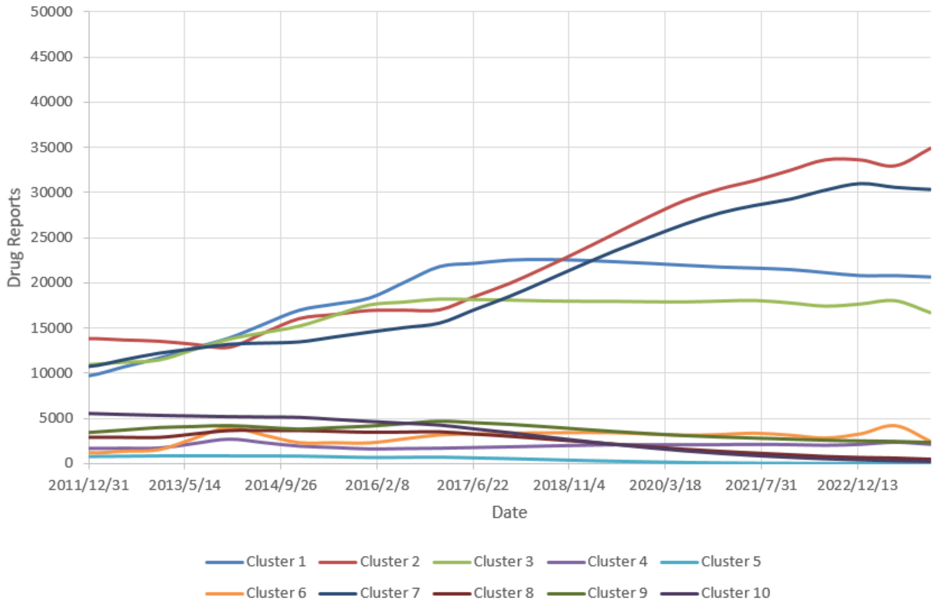


Fig. 8: Trend of the Next 5 Years

According to the logistic model, we have obtained upper bound for drug-use level in each region, which we approximately consider as the drug threshold level for each region, making an analogy with the maximum environmental capacity in biology. If drug usage exceeds the threshold level, the drug use in this region might become uncontrollable thus resulting in an unprecedented social crisis.

In the above figure, we know that the number of drug use in some areas is decreasing over time or being stable at a relatively low value. So it can be considered that drug usage will not exceed its threshold level in the next few years, and the government does not need to be concerned about these areas, hence we no longer make their curves. Now we focus on clusters with clearly upward drug-use trends. The first goal is to find the intersection of the drug threshold level and the rising curve in order to make a judgement on whether government should take actions. The following figures clearly demonstrates what has been mentioned.

We can see that both Cluster-2 and Cluster-7, which represent east of West Virginia and part of Pennsylvania, exceed the drug threshold levels. Cluster-2 will reach the threshold level 2018, which indicates that local government should take measures as much quickly as possible. And Cluster-7 will reach the threshold a few years later, around 2021. Also, it shows that the rising trend still maintains for the following years. So it can be considered that drug use may be uncontrolled in these two regions when the levels are exceeded. Thus, it is urgent for the government to adopt effective measures in advance.

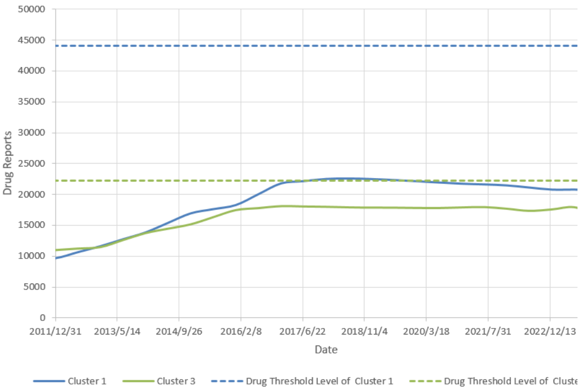


Fig. 9: Threshold levesl of Cluster-1 and Cluster-3

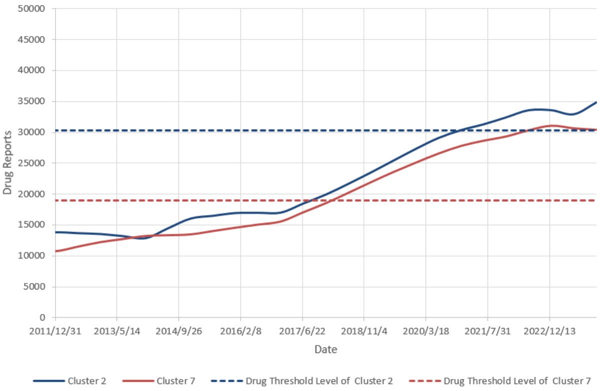


Fig. 10: Threshold Levels of Cluster-2 and Cluster-7

4.2.4 The Failure of the Model

However, the Modified Lotka-Volterra Model we construct is an unstable differential equations. As time goes on, the differential equations system show chaotic phenomena. Every 20 years, all the values of 10 clusters fluctuate severely and then suddenly converge to 0. Fig.11 depicts the trend of opioid and heroin incidents in the 10 clusters for the next 50 years.

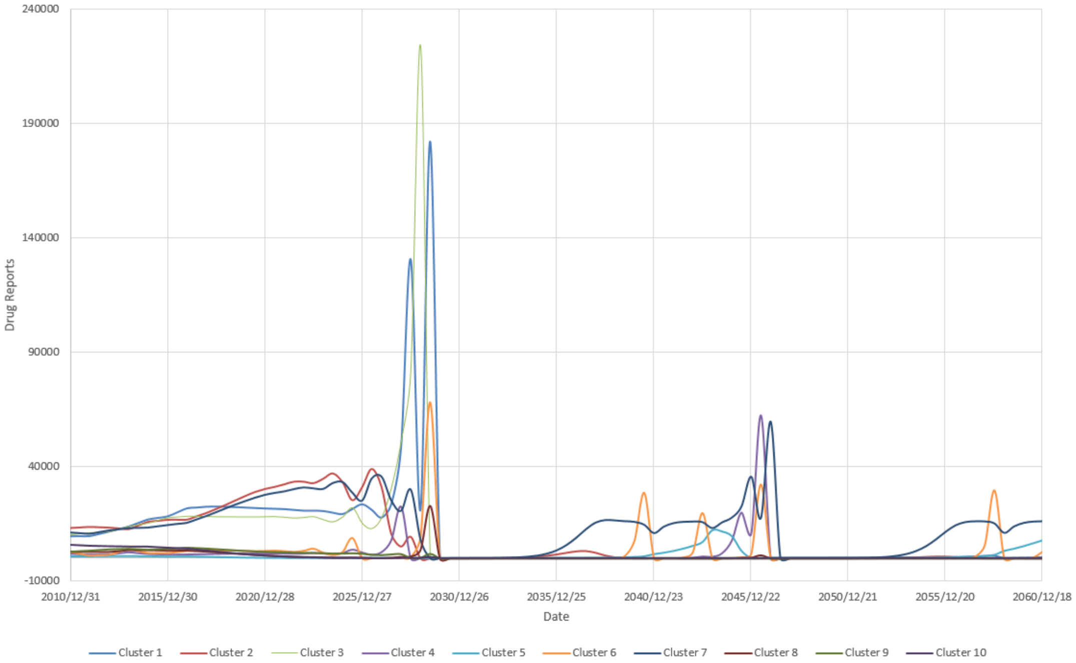


Fig. 11: Trend of the Next 50 Years

The reasons for the model failure are as follows:

1. We use the least square method to estimate the iterative matrix A by samples from 2010 to 2017. However, in the reality, the interactions between different clusters may change with time.
2. The patterns and characteristics we derive from samples may continue only for a short time. For Equation 5, we set the right-hand side of the system to 0, to get 4 equilibrium points

$$A_1(0,0)$$
$$A_2(0,\frac{c_1a_2-a_1c_2}{b_1c_2-c_1b_2})$$
$$A_3(\frac{b_1a_2-a_1b_2}{c_1b_2-b_1c_2},0)$$
$$A_4(\frac{b_1a_2-a_1b_2}{c_1b_2-b_1c_2},\frac{c_1a_2-a_1c_2}{b_1c_2-c_1b_2})$$

The first three equilibrium points contain the coordinate 0, which contrasts the reality that each species shouldn't extinct. The reasonable equilibrium point A_4 should meet the conditions:

$$\begin{vmatrix} a_1 & a_2 \\ c_1 & c_2 \end{vmatrix} * \begin{vmatrix} c_1 & c_2 \\ b_1 & b_2 \end{vmatrix} > 0$$
$$\begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix} * \begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix} > 0$$

The matrix $\bar{B}$ may not satisfy the corresponding conditions.

4.3 Considering Factors Using Modified Iterative Lotka-Volterra Model

In the last section, we build the Iterative Lotka-Volterra Model(LVM) to research the interactions between different clusters. However, the LVM can only be applied in a short range of time. In this section, we modify the LVM by adding attribute characters from the U.S. Census socio-economic data.

The U.S. Census socio-economic data has nearly 500 more variables ranging from 2010 to 2016, which is too large to use. In the data process section, we use entropy weight method to further classify variables into 15 factors. The time series of 15 factors in Cluster 1 are listed as follows.

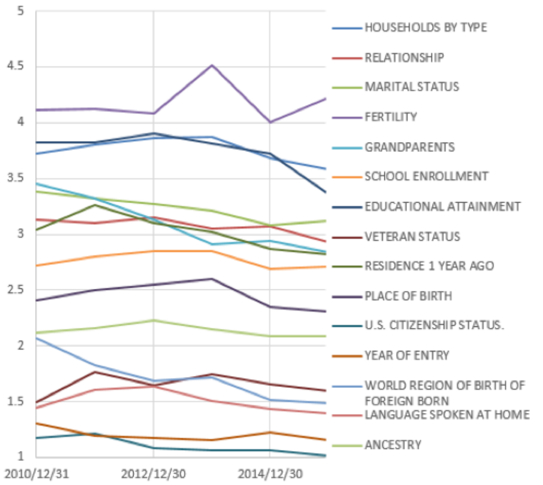


Fig. 12: Time series of 15 factors in Cluster 1

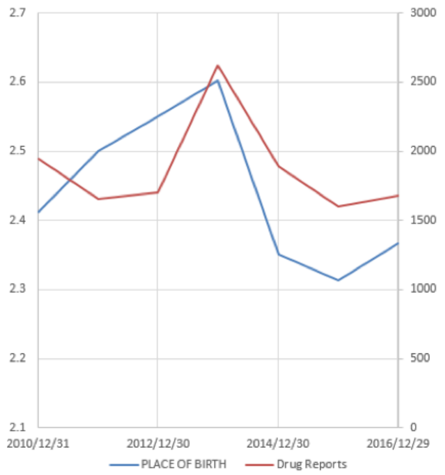


Fig. 13: Factors and Drug Reports in Cluster-1

As we can see from Fig.12, part of those factors are strongly correlated. So certain factors can provide as much information as we need in the model construction. By further process, we find factors 'PLACE OF BIRTH' of Cluster-1 has a similar trend with the Drug Reports of Cluster-1. These findings inspire us that certain factors may help to forecast the trend of Drug Reports. In order to select the best factors to forecast, we calculator the correlation coefficient between every 15 factors and every 10 clusters. The results are put in the Appendix.

Fig.14 demonstrates the major influential factors. We observe that the influencing factors differ by regions, based on the locations of five states and ten clusters in the map. 14. So we categorize these influencing factors on a geographical basis, which are presented in 14.

It is obvious that immigration plays a decisive role in Cluster-2, 8, and 9 near the East Coast. This is not difficult to understand, because many immigrants started locating in the east coast of the United States ever since Columbus discovered the Americas. The immigration culture and degree also determine their preference and use of drugs to some extent. At the same time, the drug-use level in the east of West Virginia area is related to veterans, which can be speculated in the light of the fact that military power concentrates in its vicinity - Washington, DC.

The main influence factor in the central region is the level of education. Since there are various institutions of high level in these regions (Cluster-1, 5, 7, 10), we can conclude that education is also closely related to drug use. As forwestern regions of the five states, family composition could be the main reason. Based on common knowledge, there is a certain relationship between drug crime, family background and growing environment, the result is convincing.

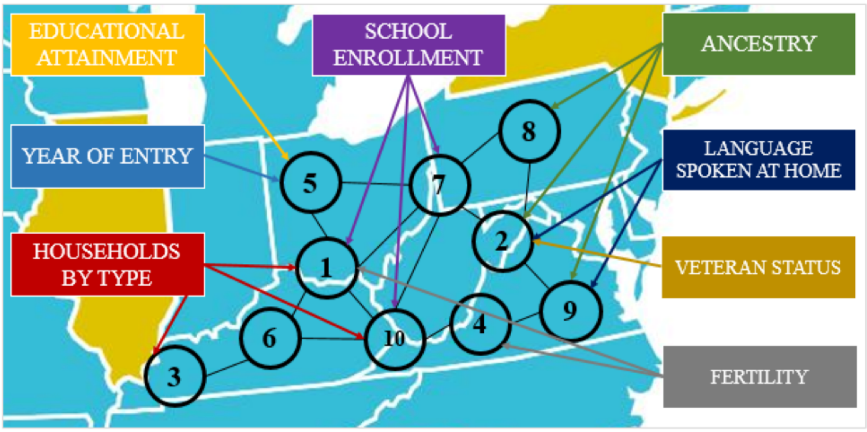


Fig. 14: Major Influential Factors

Since the Iterative Lotka-Volterra Model doesnt take the attribute into account, we can modify the model by adding attribute characters from socio-economic factors. For each Cluster, we select the factors most correlated to drug reports (including positive and negative correlation coefficient). And build linear regression model to calculate the change rate of $u_k(t)$ as follows:

$$\begin{aligned}\frac{du_k(t)}{dt} &= a + bX_t^k \\ &= h(X_t^k) \\ k &= 1, 2, ..., 9\end{aligned} \tag{15}$$

In the LVM, we calculate the change rate of $u_k(t)$ by

$$\begin{aligned}\frac{du_k(t)}{dt} &= u_k(t-1)(a_k + \sum_{j=1}^{9} b_{kj} u_k(t-1)) \\ &= f(u_k(t-1))\end{aligned} \tag{16}$$

By formula 15, we get the change rate of $u_k(t)$ by using the attributes of Cluster-k. By formula 16, we get the change rate of $u_k(t)$ by spreading the interactions between different Clusters. To combine both of the influences, we modify the change rate of $u_k(t)$ by taking the weighted average of two results.

$$\frac{du_k(t)}{dt} = \lambda(t)h(X_t^k) + (1-\lambda(t))f(u_k(t-1)) \tag{17}$$

Because in the short run, the interactions among different clusters remain stable. While in the long run, the macro attributes are more reliable to forecast. Hence the weighted parameter $\lambda(t)$ should satisfy the following properties.

$$\begin{aligned}\lambda(t) &\in (0, 1) \\ \lambda'(t) &> 0\end{aligned} \tag{18}$$

where $\lambda(t) = \frac{ln(t)}{1+ln(t)}$.

By calculating, we get the next 50 years of drug reports in different clusters.

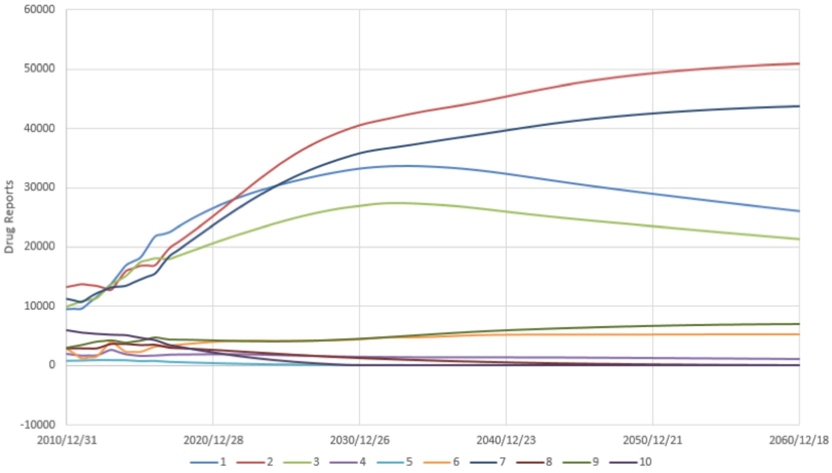


Fig. 15: Trend of the Next 50 Years

5 Toward the Policy

In this section, we will identify a possible strategy for countering the opioid crisis and use our model to test the effectiveness of this strategy. Before that, its necessary to have a close look at the historical policy to opioid users in the U.S.

5.1 Historical Policies towards Opioid Users

In U.S., the drug control started from 1940s, both use and sale of opioids are under supervision. Policicians were dedicated to enforce policies that impose penalties for drug users. Those users may feel pressure from every aspects, the government agencies, laws, and treatment programs[5]. We list major policies in the following paragraphs.

Some states started to limit needles and syringes for people without prescriptions from doctors ever since 1940s. To prevent heroin addiction, the needles/syringes possession was made a crime similar to heroin possession. This policy made herion addicts share needles with each other, which stimulated the increase of HIV and AIDS infection rate.

From late 1960s, the U.S. government started to treatment addicts with new medicine, particularly methadone maintenance treatment(MMT). It was proved that MMT was useful for drug users and decreased heroin use, and drug-related crimes were also reduced. The government also constructed communities for therapy, which supplied houses, abstinence-based recovery programs for people addicted to drugs[4].

For those doesn't obey the drug control laws, police officers will impose sanctions accordingly. Public chose to take "Zero tolerance" attitude toward heroin users in view of heroin's harm. Therefore, both drug buyers and sellers took several measures to escape the inspection of the police. The sellers, made this activity unknown for as much people as they can, including neighbors, family members and even other heroin users[5].

Because heroin and other opioids often have a high cost, it is normally not covered by the insurance, since the U.S. health system is mainly funded privately. And People who are jobless may face difficulties when using the income transfrom system, many drug users were even removed from the system in the 1990s. That means that they don't have stable income, job and even residence[5].

Above all, the explanations of the current level which opioid use got to are various and complex. Some policies may have side effect, therefore the government should carefully select policy.

5.2 Short-term Policy: Reducing the Natural Growth Rate of Major Domain

In Section 4.2, we use the Lotka-Volterra equations to simulate the interactions between different clusters. This model prerequisite the interaction is stable and thus not considering changes. So it is only suitable for short-term prediction. We utilize it to guide our short-term policy.

For each Cluster-i, leaving out other clusters' influence, it has a natural growth rate a_i of itself. The first idea is to force decrease a_i for regions. For instance, enforcing laws to penalize non-prescription use of drugs more strongly, from Fig.16, the dash line represents drug-use of restricting a_i by 20%, and the effect is significant.

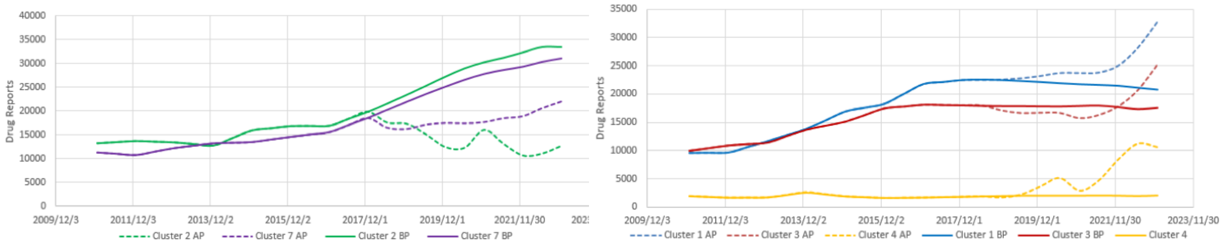


Fig. 16: Drug Reports Change Before Policy(BP) and After Policy(AP) in Policy-Implemented Areas(left) and Other Areas(right)

However, Fig.16 simultaneously demonstrates that when one area implements policies, adjacent areas without proper measures will face boom of drug-use crisis. This can be explained that people influenced in areas policy-implemented area will migrate there and more drugs may flow there after. Thereby, policies should be considered nationally rather than regionally, such as limiting opioids import, enforcing laws and more relevant policies. Fig.17 proves the practicality.

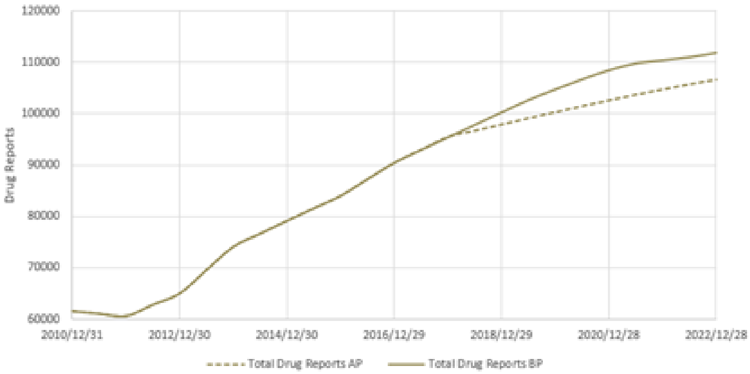


Fig. 17: Decrease a_i of All Clusters by 10%

5.3 Long-term Policy: Increasing Education Attainment and Married Rate

Compared to the short-term policy, long-term policy could be some measures that influence the whole society and then regulate the drug-use level. In order to prove out view, we use the improved model to adjust the regional education factor and family factor (marriage rate, single female proportion and so on). When we change the value of above two factors by increasing them by 0.6 percentage, we obtain the dashed line presented in the Fig.18. It is clearly that these two factors can reduce the usage of drugs by more than 10%, therefore we can believe that increasing the popularity of education, as well as encouraging marriage are conducive to affect and control the US drug-use level.

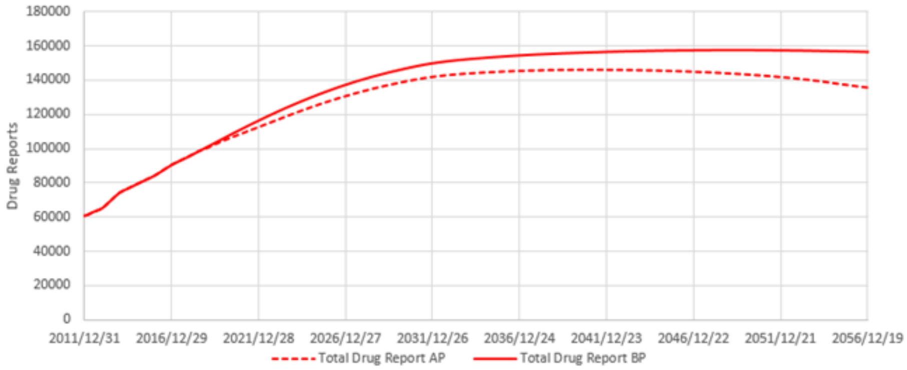


Fig. 18: Policies in the Long Term

6 Conclusion

6.1 Strengths

- In the process of Drug Reports data, our model takes the idea of clustering, which overcomes the disadvantages of analyzing either the whole state or one single county.
- In the process of Census socio-economic data, we use the Entropy Weight Method to maximum compressed variables on the basis of maintaining certain economic logic.
- In order to identify where specific opioid use, we construct the Logistic Model for simplicity. And we get maximum number from the model to further determine threshold levels for caution. While forecasting its future patterns and characteristics, we use and the Modified LVM to consider both the interaction and factors influence.

6.2 Weaknesses

- The cluster number we set is still too small to identify the accurate locations.
- The factors we get by Entropy Weight Method may lose its economic meaning, and the strong correlation between certain factor and cluster may due to chance. And there may exist excessive data mining.
- We assume the early development of heroin incidents have a stable environment meeting the Logistic Model's conditions. But with the incidents expanding, the condition may no longer sustain.
- Logistic regression model is suitable for most of the clusters data, but there are still 1-2 clusters which has weak correlation with any factor.

Memo

To: Chief Administrator, NFLIS, DEA
From: Modelers from Team 1923231
Subject: Insights and Results During Modeling
Date: Jan 28, 2019

Dear Administrator, we are honored to inform you our insights after conducting data analysis and modeling.

The misuse of opioids has become an national crisis, hence it is urgent to fight against it. We are especially interested in this problem and determined to contribute. Using the dataset provided by NFLIS, we aim to address the insights and recommendations for you.

Initially, we use Logistic Regression Model to characterize the trend of opioids use in early stage and identify the maximum number of possible use, which can be used to determine the threshold. Then we use Lotka-Volterra Model to take different clusters' interaction into account. It can be used to forecast the use trend in the future. Further, we take socio-economic factors into consideration and find out factors that are decisive for every region.

Secondly, during we analyze the dataset, we have the following insights:

- The influencing factors for different regions differ greatly.
- In the short term, controlling the drug use by force in one region may lead to an increase in the neighboring region.
- In the long term, educational level and family environment such as marital status have paramount effects on drug use.

Thirdly, based on our research, we have some recommendations for DEA:

- Reduce the possibility of making regional policies, take national action instead.
- Promote the educational level and family satisfication of the public, which can be achieved by developing the economic and others.

References

[1] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

[2] Gazetteer Files - Geography - U.S. Census Bureau. `https://www.census.gov/geo/maps-data/data/gazetteer.html`

[3] Shannon, C. E. (1948). A mathematical theory of communication. Bell system technical journal, 27(3), 379-423.

[4] DrugBank. `https://www.drugbank.ca/drug`

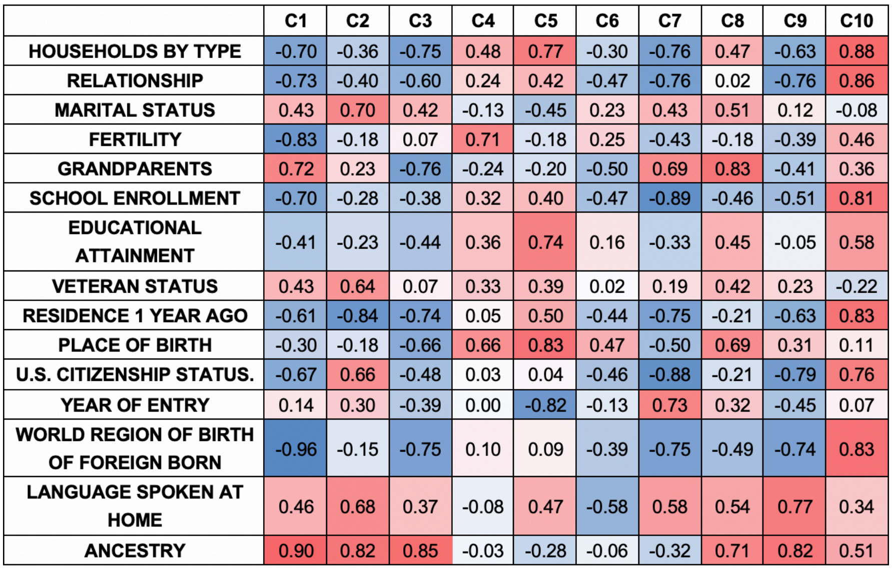
[5] Johnson, B. D., Maher, L., & Friedman, S. R. (2001). What public policies affect heroin users?. Journal of Applied Sociology, 14-49.

[6] ZHANG Yu, GAO Kening, CHEN Mo, YU GA (2018). Method of Link Prediction Combining Network Structure and Node Attributes. Journal of Frontiers of Computer Science and Technology.

[7] Carrieri, M. P., Amass, L., Lucas, G. M., Vlahov, D., Wodak, A., & Woody, G. E. (2006). Buprenorphine use: the international experience. Clinical Infectious Diseases, 43(Supplement_4), S197-S215.

Appendices

Appendix A Figures



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
HOUSEHOLDS BY TYPE	-0.70	-0.36	-0.75	0.48	0.77	-0.30	-0.76	0.47	-0.63	0.88
RELATIONSHIP	-0.73	-0.40	-0.60	0.24	0.42	-0.47	-0.76	0.02	-0.76	0.86
MARITAL STATUS	0.43	0.70	0.42	-0.13	-0.45	0.23	0.43	0.51	0.12	-0.08
FERTILITY	-0.83	-0.18	0.07	0.71	-0.18	0.25	-0.43	-0.18	-0.39	0.46
GRANDPARENTS	0.72	0.23	-0.76	-0.24	-0.20	-0.50	0.69	0.83	-0.41	0.36
SCHOOL ENROLLMENT	-0.70	-0.28	-0.38	0.32	0.40	-0.47	-0.89	-0.46	-0.51	0.81
EDUCATIONAL ATTAINMENT	-0.41	-0.23	-0.44	0.36	0.74	0.16	-0.33	0.45	-0.05	0.58
VETERAN STATUS	0.43	0.64	0.07	0.33	0.39	0.02	0.19	0.42	0.23	-0.22
RESIDENCE 1 YEAR AGO	-0.61	-0.84	-0.74	0.05	0.50	-0.44	-0.75	-0.21	-0.63	0.83
PLACE OF BIRTH	-0.30	-0.18	-0.66	0.66	0.83	0.47	-0.50	0.69	0.31	0.11
U.S. CITIZENSHIP STATUS.	-0.67	0.66	-0.48	0.03	0.04	-0.46	-0.88	-0.21	-0.79	0.76
YEAR OF ENTRY	0.14	0.30	-0.39	0.00	-0.82	-0.13	0.73	0.32	-0.45	0.07
WORLD REGION OF BIRTH OF FOREIGN BORN	-0.96	-0.15	-0.75	0.10	0.09	-0.39	-0.75	-0.49	-0.74	0.83
LANGUAGE SPOKEN AT HOME	0.46	0.68	0.37	-0.08	0.47	-0.58	0.58	0.54	0.77	0.34
ANCESTRY	0.90	0.82	0.85	-0.03	-0.28	-0.06	-0.32	0.71	0.82	0.51

Fig. 19: coefficients between Factors and Clusters

Appendix B The Source Codes

This python program implements k-means clustering.

Program 1: k-means.py

```
from numpy import *
import time
from math import radians, cos, sin, asin, sqrt
import matplotlib.pyplot as plt


def calculate_dis(vec1,
                vec2):
    """
    Calculate the distance between two places
    """
    # print(vec1)
    # print(vec2)
    lat1 = vec1[0]
    lon1 = vec1[1]
    lat2 = vec2[0]
    lon2 = vec2[1]
    lon1, lat1, lon2, lat2 = map(radians, [lon1, lat1, lon2, lat2])

    # haversine
    dlon = lon2 - lon1
    dlat = lat2 - lat1
    a = sin(dlat/2)**2 + cos(lat1) * cos(lat2) * sin(dlon/2)**2
    c = 2 * asin(sqrt(a))
    r = 6371 # the radius of earth
    return c * r * 1000 # return in meters

# init centroids with random samples
def initCentroids(dataSet, k):
    numSamples, dim = dataSet.shape
    centroids = zeros((k, dim))
    for i in range(k):
        index = int(random.uniform(0, numSamples))
        centroids[i, :] = dataSet[index, :]
    return centroids

# k-means cluster
def kmeans(dataSet, k):
    numSamples = dataSet.shape[0]
    # first column stores which cluster this sample belongs to,
    # second column stores the error between this sample and its centroid
    clusterAssment = mat(zeros((numSamples, 2)))
    res = zeros((numSamples, 1))
    clusterChanged = True

    centroids = initCentroids(dataSet, k)

    while clusterChanged:
        clusterChanged = False
        ## for each sample
        for i in range(numSamples):
            minDist  = 100000000000.0
            minIndex = 0
            ## for each centroid
            ## step 2: find the centroid who is closest
            for j in range(k):
                distance = calculate_dis(centroids[j, :], array(dataSet[i])
                    [0]) #convert to array
                if distance < minDist:
                    minDist  = distance
                    # print("minDist", minDist)
                    minIndex = j
                    # print("minindex", minIndex)

            ## step 3: update its cluster
            if clusterAssment[i, 0] != minIndex:
                clusterChanged = True
                clusterAssment[i, :] = minIndex, minDist**2
                res[i] = minIndex

        ## step 4: update centroids
        for j in range(k):
            pointsInCluster = dataSet[nonzero(clusterAssment[:, 0].A == j)
                [0]]
            centroids[j, :] = mean(pointsInCluster, axis = 0)
    return centroids, clusterAssment, res
```

This python program implements entropy weight method.

Program 2: entropy-weight.py

```
import numpy as np
import pandas as pd
```

```
def entropy(dataset):
    n, k = np.shape(dataset)
    maximum = np.max(dataset, axis=0) #minimum in column
    minimum = np.min(dataset, axis=0)
    data = (dataset - minimum) * 1.0 / (maximum - minimum)
    col_sum = np.sum(data, axis=0)
    data = data / col_sum
    a = data * 1.0
    a[np.where(data==0)]=0.0001
    e = (-1.0/np.log(n))*np.sum(data*np.log(a), axis=0) # e 1*k
    w = (1 - e) / np.sum(1 - e)
    recodes = np.sum(dataset * w, axis=1)
    return recodes
```

Program 3: Logistic Model

```
# logistic
def logistic_functions(df,name):
    dg = pd.DataFrame()
    dg = dg.reindex(range(1998, 2018))
    alpha_list=list(range(0,10))
    for k in range(0, 10):
        print(k)
        y = np.array(df[df['cluster'].isin([k])]['DrugReports'].values)
        x = np.array(range(0, 8))
        y1 = 1 / y
        y1 = y1.reshape(-1, 1)
        x1 = np.exp(-x)
        x1 = x1.reshape(-1, 1)

        model = LinearRegression()
        model.fit(x1, y1)
        alpha = model.intercept_[0]
        beta = model.coef_[0][0]
        x0 = list(range(-12, 8))
        y0 = [1 / (alpha + beta * np.exp(-i)) for i in x0]
        alpha_list[k]=1/alpha

        x0 = range(1998, 2018)
        y0 = np.log(y0)

        dg['Fitted␣Drugreports␣Number␣of␣Cluster' + str(k + 1)] = list(y0)
        dg.loc[2010:2018, 'Actual␣Drugreports␣Number␣of␣Cluster' + str(k +
           1)] = np.log(y)
```

Program 4: LVM Model

```
dg=pd.DataFrame()
dg=dg.reindex(columns=range(0,10))
dg=dg.reindex(pd.date_range('20101231','20171231',freq='6M'))

for i in pd.date_range('20101231','20171231',freq='Y'):
    dh=df[df['date'].isin([i])]
    dg.loc[i,:]=list(dh['DrugReports'])
for i in range(2011,2018):
    dg.loc[pd.datetime(year=i,month=6,day=30),:]=dg.loc[pd.datetime(year=i
        -1,month=12,day=31),:]/2+dg.loc[pd.datetime(year=i,month=12,day=31)
        ,:]/2
```

```
matrix=np.zeros((10,11))
for j in range(0,10):
    y=dg[j].pct_change(1)[1:15].values
    x=dg.iloc[0:14,0:10].values
    y=y.reshape(-1,1)
    x=x.reshape(-1,10)

    model = LinearRegression()
    model.fit(x, y)
    alpha = list(model.intercept_)
    beta = list(model.coef_[0])
    alpha.extend(beta)
    matrix[j,:]=alpha

dg=dg.reindex(pd.date_range(’20101231’,’20601231’,freq=’6M’))
dg.index=range(1,102)

for i in range(16,102):
    list_b=[1.0]
    list_b.extend(dg.loc[i-1,:])
    result=np.dot(matrix,list_b)

    result1=(result+1)*np.array(dg.loc[i-1,:])
    for k in range(0,10):
        if result1[k]<0:
            result1[k]=1
    dg.loc[i, :]=result
```

Program 5: Modified LVM Model

```
dg=pd.DataFrame()
dg=dg.reindex(columns=range(0,10))
dg=dg.reindex(pd.date_range(’20101231’,’20171231’,freq=’6M’))

for i in pd.date_range(’20101231’,’20171231’,freq=’Y’):
    dh=df[df[’date’].isin([i])]
    dg.loc[i,:]=list(dh[’DrugReports’])
for i in range(2011,2018):
    dg.loc[pd.datetime(year=i,month=6,day=30),:]=dg.loc[pd.datetime(year=i
       -1,month=12,day=31),:]/2+dg.loc[pd.datetime(year=i,month=12,day=31)
       ,:]/2

dg=dg.reindex(pd.date_range(’20101231’,’20601231’,freq=’6M’))
dg.index=range(1,102)
dg2=pd.DataFrame()
dg2=dg2.reindex(range(1,102))

dh=pd.DataFrame()
dh=dh.reindex(range(1,102))

file=’C:\\Users\Administrator\Desktop\\\\3spread+feature\Data\\
   forecast_entropy.xlsx’
df1=pd.read_excel(file,sheet_name=’0’)
dh[’0.1’]=df1[’WORLD␣REGION␣OF␣BIRTH␣OF␣FOREIGN␣BORN’]
dh[’0.1’][16:]=dh[’0.1’][16:]
# dh[’0.2’]=df1[’ANCESTRY’]
# dh[’0.3’]=df1[’FERTILITY’]

df1=pd.read_excel(file,sheet_name=’1’)
```

```
dh[’1.1’]=df1[’RESIDENCE␣1␣YEAR␣AGO’]
# dh[’1.2’]=df1[’ANCESTRY’].pct_change(1)

df1=pd.read_excel(file,sheet_name=’2’)
# dh[’2.1’]=df1[’GRANDPARENTS’].pct_change(1)
dh[’2.1’]=df1[’ANCESTRY’]
dh[’2.1’][16:]=dh[’2.1’][16:]

df1=pd.read_excel(file,sheet_name=’3’)
dh[’3.1’]=df1[’FERTILITY’]
dh[’3.1’][16:]=dh[’3.1’][16:]*0.994
# dh[’3.2’]=df1[’PLACE OF BIRTH’].pct_change(1)

df1=pd.read_excel(file,sheet_name=’4’)
# dh[’4.1’]=df1[’YEAR OF ENTRY’].pct_change(1)
dh[’4.1’]=df1[’HOUSEHOLDS␣BY␣TYPE’]

df1=pd.read_excel(file,sheet_name=’5’)
dh[’5.1’]=df1[’LANGUAGE␣SPOKEN␣AT␣HOME’]
# dh[’5.2’]=df1[’FERTILITY’].pct_change(1)

df1=pd.read_excel(file,sheet_name=’6’)
dh[’6.1’]=df1[’SCHOOL␣ENROLLMENT’]
dh[’6.1’][16:]=dh[’6.1’][16:]
# dh[’6.2’]=df1[’GRANDPARENTS’].pct_change(1)

df1=pd.read_excel(file,sheet_name=’7’)
dh[’7.1’]=df1[’GRANDPARENTS’]
# dh[’7.2’]=df1[’WORLD REGION OF BIRTH OF FOREIGN BORN’].pct_change(1)

df1=pd.read_excel(file,sheet_name=’8’)
# dh[’8.1’]=df1[’U’].pct_change(1)
dh[’8.1’]=df1[’ANCESTRY’]

df1=pd.read_excel(file,sheet_name=’9’)
dh[’9.1’]=df1[’HOUSEHOLDS␣BY␣TYPE’]
# dh[’9.2’]=df1[’VETERAN STATUS’].pct_chang
matrix=np.zeros((10,11))
for j in range(0,10):
    y=dg[j].pct_change(1)[1:15].values
    x=dg.iloc[0:14,0:10].values
    y=y.reshape(-1,1)
    x=x.reshape(-1,10)

    model = LinearRegression()
    model.fit(x, y)
    alpha = list(model.intercept_)
    beta = list(model.coef_[0])
    alpha.extend(beta)
    matrix[j,:]=alpha

# dh.to_excel(’C:\\Users\Administrator\Desktop\\\ARIMA\Data\logistic\\
   number.xlsx’)

matrix2=np.zeros((10,2))
for j in range(0,10):
    print(j)
    # y=dg[j].pct_change(1)[1:15].values
    y = dg[j][1:15].values
    x=dh.loc[1:14,str(j)+’.1’].values
    y=y.reshape(-1,1)
    x=x.reshape(-1,1)
    model = LinearRegression()
    model.fit(x, y)
    alpha = list(model.intercept_)
    beta = list(model.coef_[0])
    alpha.extend(beta)
    matrix2[j,:]=alpha
    dg2[j]=alpha[0]+model.coef_[0]*dh[str(j)+'.1']
    dg2[j]=dg2[j].pct_change(1)

for i in range(16,102):
    print(i)
    list_b=[1.0]
    list_b.extend(dg.loc[i-1,:])
    result1=np.dot(matrix,list_b)
    result2=np.array(dg2.loc[i,:])
    # result_12=result1/4+result2*3/4
    result_12=result1/(np.log(i)+1)+result2*np.log(i)/(np.log(i)+1)
    # print(result1)
    print(result2+1)
    # print(result_12)
    # print(result_12+1)
    result3=(result_12+1)*np.array(dg.loc[i-1,:])
    for k in range(0,10):
        if result3[k]<0:
            result3[k]=1
    dg.loc[i, :]=result3
dg.columns=['1','2','3','4','5','6','7','8','9','10']
dg.index=pd.date_range('20101231','20601231',freq='6M')
dg['TOTAL']=dg[['1','2','3','4','5','6','7','8','9','10']].sum(axis=1)
dg['TOTAL'].plot()
plt.show()
print(dg['TOTAL'])
dg.to_excel('C:\\Users\Administrator\Desktop\\\3spread+feature\Data\
   Model_spread+feature(log)(policy).xlsx')
```