For office use only	Team Control Number	For office use only
T1 _____	**1922115**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**C**	F4 _____

2019
MCM/ICM
Summary Sheet

A Comprehensive Analysis and Prediction of American Opioid Abuse

Summary

The United States is going through a national crisis of opioid abuse, which has great implications on many important sectors of the U.S. economy. In this paper, we are provided the dataset of local drug cases for five states, and we are required to analyze and predict the data as well as identify a possible strategy for countering the opioid crisis.

First, we preprocess the data. In the process of missing value addressing, we deal with the variables lacking a small amount of data by **Cubic Spline Interpolation**.

Second, in order to characterize the spread of opioid reports, we construct some indicators from the dataset: average heroin reports in county, entropy of geographical distribution, entropy of types, etc. Based on the above indicators, we can characterize the spreading tendency.

Third, to identify the next spreading area, we firstly build a spreading model to describe the spreading pattern, then we use clustering to decrease the difficulty for determining parameters in the model. Next, we use **Monte Carlo simulation** to estimate the parameters. Lastly, we use the extimated data to simulate the next spreading area. The results show Calloway County (KY), Cumberland County (VA) and so on will be possible locations where synthetic opioid use might have started.

Fourth, to predict the trends in the future accurately, we adopt **Gaussian Process Regreesion** to predict the abuse population for each state in next five years. Then we define the threshold as the twice of average three-fourths quantile of the data. Based on the threshold, we can identify the severe outbreak areas for heroin where U.S. government should pay special attention on as: PHILADEPHIA (PA), HAMILON (OH), etc; for synthetic opioids, there are CUYAHOGA (PA), CUYAHOGA (OH), etc.

Finally, we use **Gray Correlation Analysis** to find the assoicated socio-economic factors with the trend of opioid use. Then based on these factors, we adopt **Gradient Boostling Decision Tree** to establish a quantitative relationship between them and the abuse population. In the process of determining threshold, we define a "Potential Flashpoint" which are the points where abuse population experiences a dramatic increases. Based on the point, we set the target for controlling the further deterioration of opioid abuse situation is to prevent the further development of abuse population before it reaches the "Potential Flashpoint". To achieve this, U.S. goverment needs to keep the divorce rate less than 31.3%, keep the percentage of college students with associate's degree less than 5.2%.

Keywords: Gaussian Process Regreesion; Gray Correlation Analysis; Gradient Boostling Decision Tree

Contents

1 Introduction

1.1 Background

The United States is going through a national crisis of opioid abuse. Between 1999 and 2014, drug and opioid-involved overdose deaths nearly tripled compared to preceding years [1]. According to experts' speculation, one of the main reasons causing that is because the increased prescribing of opioids for alleviating pain and suffering [2]. As opioid abuse has negative influence on all ages, sexes, ethnic and socioeconomic backgrounds [3], and also has tremendous implications to all sectors of U.S economy, so it is urgent for governments to pay great attention to the spread of overdose and take action to control its further deterioration.

In this problem, the datasets about drug identification results and associated information are provided in five states, and we are required to characterize the infectious features and trends of opioid use as well as predicting the zones where opioids could be severely affected. Also, we need to identify the possible socio-economic factors which may contribute to the development of opioid abuse, and present a possible strategy for countering the opioid crisis.

1.2 Our Goals

Based on the given problems, we set the following goals:

- Use the given data to describe the characteristics and spread of reported opioid incidents for each state
- Develope a model to simulate the spreading process of opioids and predict the next possible locations where specific opioid use might have started in each of the five states
- Build a model to predict the number of opioid abusers, and define drug identification threshold levels, then find severe areas where U.S. goverment should pay special attention based on the threshold
- Identify the possible opioid abusers, and find the association between the U.S. Census socio-economic data and the trend of opioid abuse.
- Modify the model in Part I combining with the social-economic data
- Identify a possible strategy for countering the opioid crisis, and test the effectiveness of this strategy

1.3 Overview of Our Work

This is a typical data analysis problem, so we solve it from the persepctive of statistical analysis. Here is an overview of our work.
First, we preprocess the given data, including missing value addressing, data integration and classification. Specifically, we delete the redundant drugs and divide all the remaining drugs into two categories: heroin and synthetic opioids.
Second, we select some important data from the provided dataset to construct some explanatory variables, then we use them to characterize the spread of the reported opioids over time

for each state. Comparing the explanatory variables of five states, we can find which state has the most serious opioid abuse situation, and the geographical distribution of opioids.
Third, to find where specific opioid use might have started in each of the five states, we firstly build a spreading model of opioid abuse, then we can use k-means clustering to reduce the difficulty of estimating the parameters in the model, lastly we can apply **Monte Carlo Simulation** to determine the parameters and use the parameters to simulate the next spreading area.
Fourth, for predicting the specific abuse population accurately, we adopt **Gaussian Process Regression** (GPR) to predict the abuse population in each county for next five years, then we define the threshold and find the most severe outbreak areas based on the threshold. Last, we use **Grey Correlation Analysis** (GRA) to find the association between the trend of opioid use and socio-economic factors. Then based on this factor, we present **Gradient Boosting Decision Tree** to set a quantitative relationship between the socio-economic factor and abuse population. Lastly, we combine the model from the first two sections to identify the possible strategies for countering the opioid crisis.

2 Asummptions and Notations

2.1 Assumptions

We make three basic assumptions to our model:

1. The opioid abuse reports are only affected by these five states, without considering the effect from other states.
2. We assume that the observed report value for a given year follows a joint normal distribution
3. Ignore the isolation among counties which restrains the spread of opioid
4. The data is authentic and reliable

2.2 Notations

The notation can be seen in table 1.

3 Data Processing

For this problem, there are some incomplete and mismatched data in the large amount of raw data, which may have significant impact on the results of the model and the accuracy of our conclusions, so it is quite necessary to preprocess the data.

3.1 Addressing the Missing Values

The first data file only provides the counties where the number of opioid reports is nonzero, so there are several counties in some certain year are not recorded in data set. There are two

Table 1: Mathematical Notations

Variable	Mathematical Meanings
Us_i	Everage synthetic opioid use in county for each state
Uh_i	Everage heroin use in county for each state
T_i	Entropy of types for synthetic opioid in each state
Gh_i	Entropy of geographical distribution for heroin in each state
σh_i	Standard deviation for the number in different counties of heroin
l	Length-scale parameter in GPR model
K	Covariance matrix
γ	Gray relational coefficient
F	The functional space of all the trees
L	Loss function
r_{mi}	The negative gradient of the variable x in the m-th iteration
N_y^P	The abuse population in P-county in y-year
I_o^P	Outer factor in P-county.
I_i^P	Inner factor index in P-county.
$N_{real}(p, y)$	Real abuse population

reasons for this situation, one is that the data is missing, and the other is that there are no cases in this county this year. For the convenience of building model, we fill these undocumented counties with zero values. There are also a small amount of data missing in provided U.S. socio-economic data, we adopt cubic spline interpolation to fill in the missing values.

3.2 Data Classification and Integration

As we are only required to analyze the synthetic opioid and heroin, so we firstly delete morphine and codeine from the dataset. For further analysis, we divide all the drugs into two categories: synthetic opioid and heroin.

It is also worth mentioning that the provided U. S. socio-ecomomic data has different variable labels between 2010 to 2012 and 2013 to 2016, so we integrate the values for these different years as a whole for these variables.

4 Analysis and Prediction of Opioid Reports

4.1 Characteristics of Opioid Reports

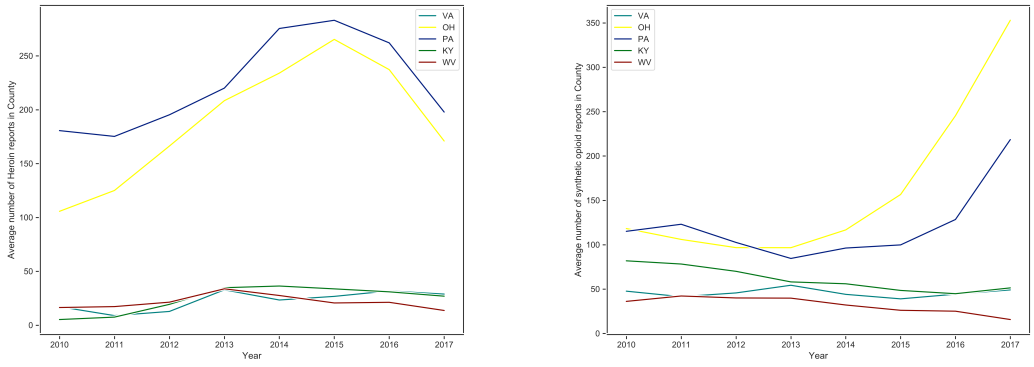
To describe the spread of the reported synthetic opioid and heroin incidents over time, we adopt the descriptive statistics method to show the characteristics of each state. We subdivides the characteristics into three aspects: average opioid use in county, distribution for different types of synthetic opioid, geographical distribution of opioid use. We then analyze each of them in five states individually.

4.1.1 Average Number of Opioid Reports in County

To eliminate the effect of the number of counties in different states, we define average opioid use in county as:

$$Us_i = \frac{Ms_i}{C_i}$$
$$Uh_i = \frac{Mh_i}{C_i}$$

Where Ms_i and Mh_i mean the total synthetic opioid and heroin reports in the state, and C_i means the number of counties in the state.



(a) Heroin (b) Synthetic Opioid

Figure 1: Average Number of Opioid Reports in County

From Figure 1, we can conclude that PA has the biggest heroin abuse population, followed by OH. Also, OH has the biggest synthetic opioid abuse population followed by PA. The reported synthetic opioid cases in OH and PA appears to an ascending tendency. The reported heroin cases in OH and PA offer upgrade firstly then descending latter tendency while WV has shown a tiny fluctuation close-by a certain value.

4.1.2 Distribution for Different Types of Synthetic Opioid

In order to characterize the distribution for the number in different types, we draw the lessons from the definition of entropy in information theory, and define "Entropy of types" as following:

$$T_i = -\sum_{k=1}^{n} p_i log_2 p_i$$

Where p_k means the percentage of one type of synthetic heroin in the total amount. In information theory, when a variable has a more balanced distribution (various values have similar probabilities), then its entropy is bigger, so in this problem, we could let the percentages of

different opioids represent the probability and define a "entropy of types" to describe distribution ratio of different synthetic opioids. When the entropy is large, it means the percentage of different types of synthetic opioids is very close.

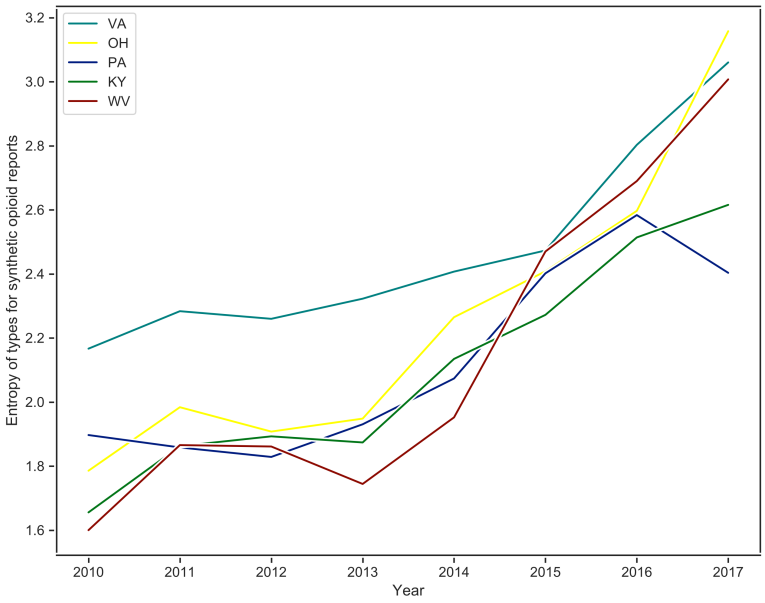


Figure 2: Entropy of Types for Synthetic Opioid

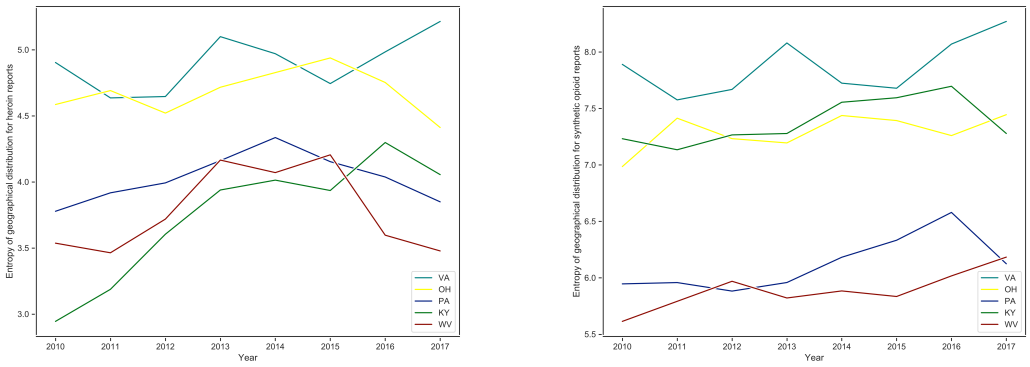
From Figure 2, we can conclude that the proportion of different types is getting more and more balanced in each state, which means at first there may be only several types of synthetic opioids, but with time going, there are more types of synthetic opioids, and their percentage is also more close.

4.1.3 Geographical Distribution of Opioid Reports

For either synthetic opioid or heroin, we characterize the geographical distribution of opioid reports by entropy of geographical distribution.
Let ph_j and ps_j to represent the percentage of heroin and synthetic opioids in each county of the state, then the entropy of geographical distribution is defined as:

$$Gh_i = -\sum_{j=1}^{n} ph_j log_2 ph_j$$

$$Gs_i = -\sum_{j=1}^{n} ps_j log_2 ps_j$$



(a) Heroin (b) Synthetic Opioid

Figure 3: Entropy of Geographical Distribution for Opioids

From 3, we can draw the following conclusions: the number of heroin abusers is more dispersed in differnt counties in VA and OH, and for synthetic opioid abusers, thier number is relatively more dispersed in VA, followed by KY and OH. Also, there is no obvious increasing or decreasing trend over time, which means their geographical distribution is fluctuating over time.

4.2 Prediction of The Next Spreading Area

In order to find the new spreading area where specific opioid use might have started in each of the five states, we firstly consider building a spreading model, which can describe the pattern of the opioid spreading. Then we use clustering to decrease the difficulty for determining parameters, lastly, we use **Monte Carlo simulation** to estimate the parameters.

4.3 Construction of Spreading Model

The spreading model is used to describe the spreading pattern of the illegal opioid abuse population. To simplify the problem, we assume the abuse population of any county in next year is depend on the population of this county and other adjacent counties in this year. We express their relationship as following:

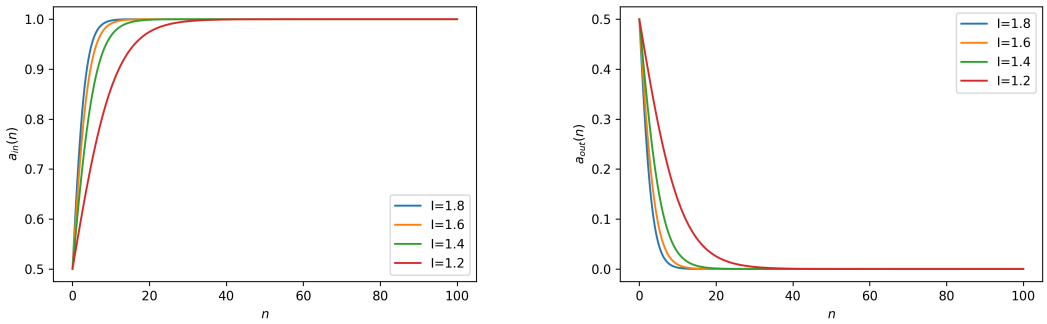
$$N_{y+1}^{P} = I_i^P N_y^p + \sum_{c \in P_{adj}} I_o^c N_y^c \tag{1}$$

Where N_y^P means the abuse population in P-county in y-year, I_o^P and I_i^P are outer factor index and inner factor index in P-county. The inner and outer factor index are all synthetic variables. In this model, in order to balance the effect of the inner factor and outer factor, we can define an activation function as following:

$$a_{in}(n) = (1 + I^{-n})^{-1} \tag{2}$$

$$a_{out}(n) = (1 + I^{n})^{-1} \tag{3}$$

Where I mean the factor index, n is the abuse population, and $a_{in}(n)$ and $a_{out}(n)$ means the activation function of inner and outer factor respectively. From Figure 4, we can find by doing so, we can offset the impact of external large values on the outcome of the results and increase the impact of internal factors on a small inner population base. It is quite reasonable in the realistic model, as usually, the inner factor plays a dominant role even when the internal population has a small base.



(a) Inner Factor (b) Outer Factor

Figure 4: Outer Factor

After combining the activation function, we define I_o^P and I_i^P as:

$$I_i^P = I_i^P + (1 + (I_{i-act}^P)^{-N_y^p})^{-1} \tag{4}$$

$$I_o^P = I_o^P + (1 + (I_{o-act}^P)^{N_y^p})^{-1} \tag{5}$$

Where I_i^P and I_{i-act}^P are two inner subfactor index, I_o^P and I_{o-act}^P are two outer subfactor index need to be determined.
For 461 counties, it is hard to determine all the outer and inner factor index. So, we apply the clustering to find the counties with similar pattern. For them, we use the same index for each county.

4.3.1 Data Clustering by K-means

The total opioid report number varies widely from place to place. We need to do the data normalization before clustering the time series of abuse population. In this case, we apply the min-max normalization

$$N_y' = \frac{N_y - min(N)}{max(N) - min(N)} \tag{6}$$

Then we apply k-means clustering to the data, and we treat the time series of abuse population as eight dimensions to group the different counties. The final clustering result graph is in 8 dimensions, we can project it into two dimensions as shown in 5.

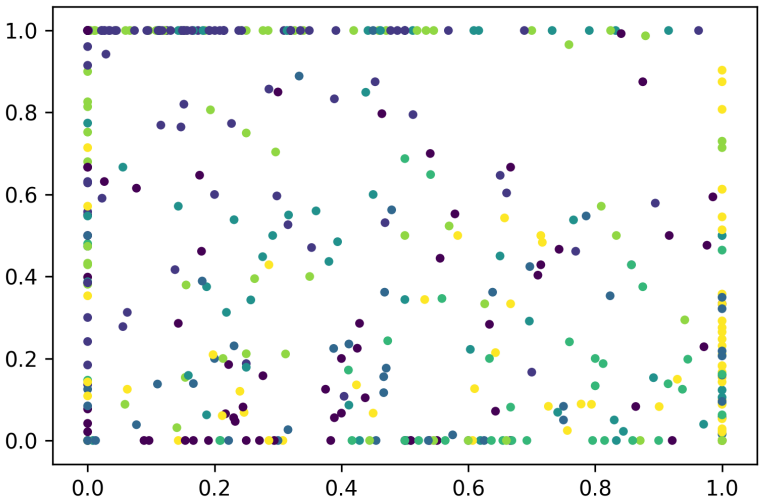


Figure 5: Clusting Result shown in Two dimensions

4.3.2 Estimate the Parameters by Monte Carlo Simulation

After clustering the county, we use the **Monte Carlo Simulation** based on the spreading model to estimate the parameters. For using the initial state N^P_{2010}, we do Monte Carlo simulation with randomly chosen factor index I^P_o and I^P_i and estimate the N^P_y in the following years. Specifically, let N^P_y represents the estimated abuse population based on the index factor (I^P_i and I^P_i), $N_{real}(p, y)$ is real abuse population, then our target is:

$$\arg\min_{I^c_o, I^c_i} \sum_{y \in Y} \sum_{p \in C} |N^p_y - N_{real}(p, y)| \tag{7}$$

After simulating the factor index, we can compare our simulation results with real results as can be seen in Figure 6 and 7, we can ensure that our model is basically in line with the actual situation.

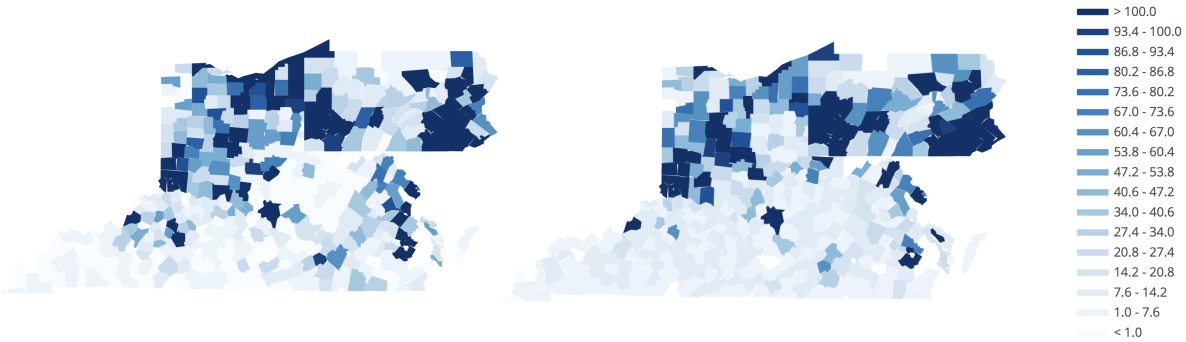


Figure 6: Real Result (Left) and Simulation Result (Right) of Heroin Abuse Population

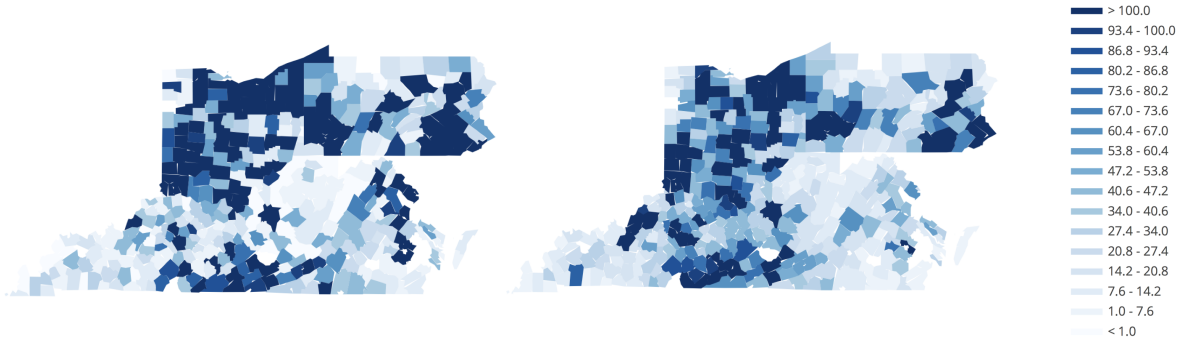


Figure 7: Real Result (Left) and Simulation Result (Right) of Synthetic Abuse Abuse Population

Lastly, we substitute the population data in 2017 (N^P_{2017}) into the model with estimated factor index to estimate N^P_{2017}, we can find the next spreading area where opioid use might have started given in 2.

Table 2: Add caption

Synthetic	Heroin
Calloway County, KY	Accomack County, VA
Cumberland County, VA	Carroll County, OH
Mineral County, WV	Carlisle County, KY
Montour County, PA	Graves County, KY
Paulding County, OH	Greene County, VA
Wirt County, WV	Patrick County, VA
Northampton County, VA	Scott County, VA
-	Vinton County, OH

4.4 Identification of Severe Outbreak Areas

4.4.1 Construction of Multi-kernel Based GPR Model

As the above model has a good simulation of the overall spreading trend, it can predict the new spreading area well, but it lacks accuracy for the prediction of the specific abuse population. In a bid to obtain a constructive foreknowledge of U.S government concerns in the future and offer related constructive suggestion for them, we are prone to constitute a more accurate model to vaticinate the number of the reported synthetic opioid and heroin incidents in the future.
First, we perform **exploratory data analysis** (EDA) from the raw data to understand the profile of the given data. Based on the strength of the EDA, we generalize several stringent requirements that must be satisfied by the model:

- The model needs to identify the fundamental trends of time series given by raw data.

- The model should be endowed with the capability that adaptable to data intrinsic instability.
- The model needs to be able to solve the problem that the linear model cannot well-fit the data with random noise.

Then take into these factors into consideration, plus we cannot simply know given time series of abuse population is quadratic, cubic, or even nonpolynomial, we put **Gaussian Process Regression** (GPR) to use to construct our forecasting model. The basic idea of GPR is that the process is regarded as the extension of multi-dimensional Gaussian distribution to infinite dimension [5]. The model of GPR could be expressed as equation (8).

$$y = f(x) + N(0, \sigma_n^2) \tag{8}$$

where $f(x)$ is a Gaussian process that is assumed to be given a priori based on data as quation (9).

$$f(x) \sim GP(0, K) \tag{9}$$

Where K is called kernel. To increase the robustness of the model, we need to get rid of the traditional method of using kernel function homogeneously. Particularly, due to the fact that the effectiveness of GPR is very dependent on the choice of covariance function and hyperparameters, so we choose three kernels respectively to make a final prediction. This greatly improves the shortcoming of the single kernel function for the final prediction result. These kernel functions include RBF kernel, Rational quadratic kernel and Matern kernel shown in equation 10, 11 and 12. Among the hyperparameters of Matern kernel, $\nu = \frac{3}{2}$ are popular choices for learning functions that are not infinitely differentiable but at least once.

$$K_{RBF}(x_i, x_j) = exp(-\frac{1}{2}d(\frac{x_i}{l}, \frac{x_j}{l})^2) \tag{10}$$

$$K_{Rational}(x_i, x_j) = (1 + \frac{d(x_i, x_j)^2}{2\alpha l^2})^{-\alpha} \tag{11}$$

$$K_{MATERN}(x_i, x_j) = \sigma^2 \left(1 + \gamma\sqrt{3}d(x_i/l, x_j/l)\right) exp\left(-\gamma\sqrt{3}d(x_i/l, x_j/l)\right) \qquad \nu = \tfrac{3}{2} \tag{12}$$

where l is a length-scale parameter that is larger than 0 and α is a parameter that can be trained to well-adapt the GPR model.

Consider training sample $y = [y_1, y_2, ..., y_n]$, $X = [x_1, x_2, ..., x_n]$ for samples y_* to be predicted. The vector of all random variables follows a multivariate gaussian distribution of zero means. Zero mean normalization is as follows:

$$y' = \frac{y - mean(y)}{std(y)} \tag{13}$$

where $std(y)$ means the standard variance of y and $mean(y)$ is the mean of y. For value $x*$, to predict $y*$, we can get:

$$\begin{bmatrix} \mathbf{y} \\ y^* \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}) \tag{14}$$

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & ... & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & ... & k(x_2, x_n) \\ ... & ... & ... & ... \\ k(x_n, x_1) & k(x_n, x_2) & ... & k(x_n, x_n) \end{bmatrix}$$

$$K_* = \begin{bmatrix} k(x_*, x_1) & k(x_*, x_2) & ... & k(x_*, x_n) \end{bmatrix}$$

$$K_{**} = k(x_*, x_*)$$

The relationship between the multivariate Gaussian distribution and the Gaussian process is that GP defines the function, while the Gaussian distribution describes the vector composed of random variables of the training samples and the test samples.

According to the conditional distribution of multivariate Gaussian distribution mentioned above, we can get equation (15)

$$y^* | \mathbf{y} \sim \mathcal{N}(K_* K^{-1} \mathbf{y}, K_{**} - K_* K^{-1} K_*^T) \tag{15}$$

Therefore, we could consider $y^* \sim N(0, \sigma_*^2)$ where,

$$\sigma_* = \sqrt{K_{**} - K_* K^{-1} K_*^T + \sigma} \tag{16}$$

Where σ is the noise level.

By adopting a multi-kernel based GPR model, we fit the time series of heroin cases and synthetic opioid cases in five states separately. GPR model could not only determine a certain value, but we could also demonstrate the 95% confident intervals of each kernel-based GPR model in the figure. The dotted lines in these figures show 30 simulated Gaussian process curves, which more intuitively explain the confidence interval.

4.4.2 Predict the Development Trend of Opioid Abuse in the Future

Take Ohio for instance, figure 8 9 10 shows the prediction of the heroin abuse population, and Figure 11 12 13 shows the prediction of the synthetic opioid abuse population in Ohio by the multi-kernel in the next five years based GPR model. As we can see, the prediction of the synthetic opioid population is slightly different in different kernel functions. By contrast, the prediction of heroin abuse population in Ohio shows a similar tendency. From the information given above, we could say that in the next five years, heroin abuse in Ohio will show a relatively flat trend while synthetic opioid use will keep going up significantly. In other words, for Ohio, the government should be concerned about the use of synthetic opioids and should tighten control on synthetic opioids as soon as possible.

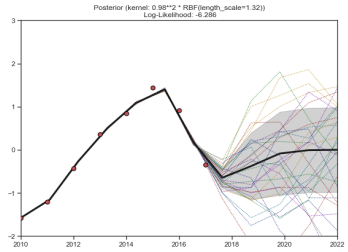


Figure 8: RBF

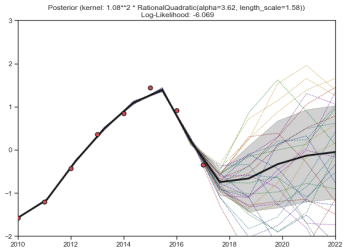


Figure 9: Rational

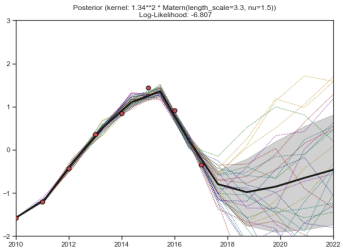


Figure 10: Matern

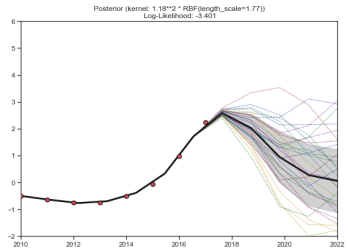


Figure 11: RBF

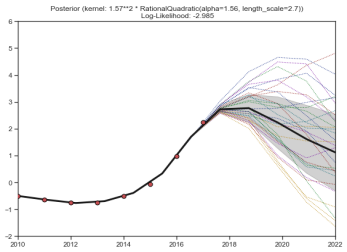


Figure 12: Rational

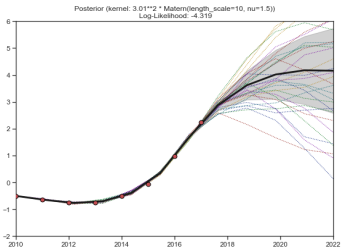


Figure 13: Matern

4.4.3 Identify the Severe Outbreak Areas

For every county in each state, we are going to apply multi-kernel based GPR model to each of them for the sake of finding if it is imperative for U.S. government to carry out related measures to control over the undesirable situation.

Threshold The setting of the threshold is particularly important, and due to the limitations of the data itself, in this part, we only roughly define the threshold. As all opioids are divided into heroin and synthetic opioids, so they have two different thresholds for each of them. For each county, we find three-fourths quantile of abuse population in each county for every subclass of opioid. And then we take the average of these three quartiles, denoted by $\mathscr{T}$. Thresholds are chosen based on this way rather than using averages or quantiles directly because the data vary greatly, they cannot describe the characteristics of the data well. Instead, this method characterizes data in crosswise and lengthways, which is more comprehensive.

After calculation, we find the threshold of heroin is $\mathscr{T} = 424$, and the threshold of synthetic opioid is $\mathscr{T} = 364$.

Severe Outbreak Areas According to the predicted values, U.S. government need to pay special attention to those counties whose abuse reports will exceed $\mathscr{T}$ in the future. For heroin, these counties include PHILADELPHIA (PA), HAMILTON (OH), ALLEGHENY (PA), etc. For synthetic opioid, these counties include HAMILTON (OH), CUYAHOGA(PA), CUYAHOGA (OH), etc. Other areas that need to be focused by the government have been placed in the appendix.

4.5 Sensitivity Analysis

We inquire into the sensitivity of a certain parameter(ν in Matern kernel) in our multi-kernel based GPR model. The reason why the sensitivity test is only carried out for this parameter is that for other parameters, they have been continuously optimized in the process of model construction, while ν is artificially determined according to experience. As can be seen in Figure 14, we change ν from 1.0 to 2.0. For each of them, the forecast curve could witness an firstly moves upward and then go through a gentle descent. Although the prediction in near 2023 varies most greatly, the prediction in 2040 tend to be the same. Such trends reveal a phenomenon in which the parameter ν is more sensitive to short-term predictions than to long-term ones. Additionally, With the increase of parameters, the prediction result of the model and its 95% confidence interval tends to a larger value. The inherent elasticity of the model also gives the U.S. government better options. In particular, if the government has more stringent requirements for a certain region, the value of ν can be deliberately set higher, which enables the government to take measures to prepare for the possible outbreak timely, and vice versa.

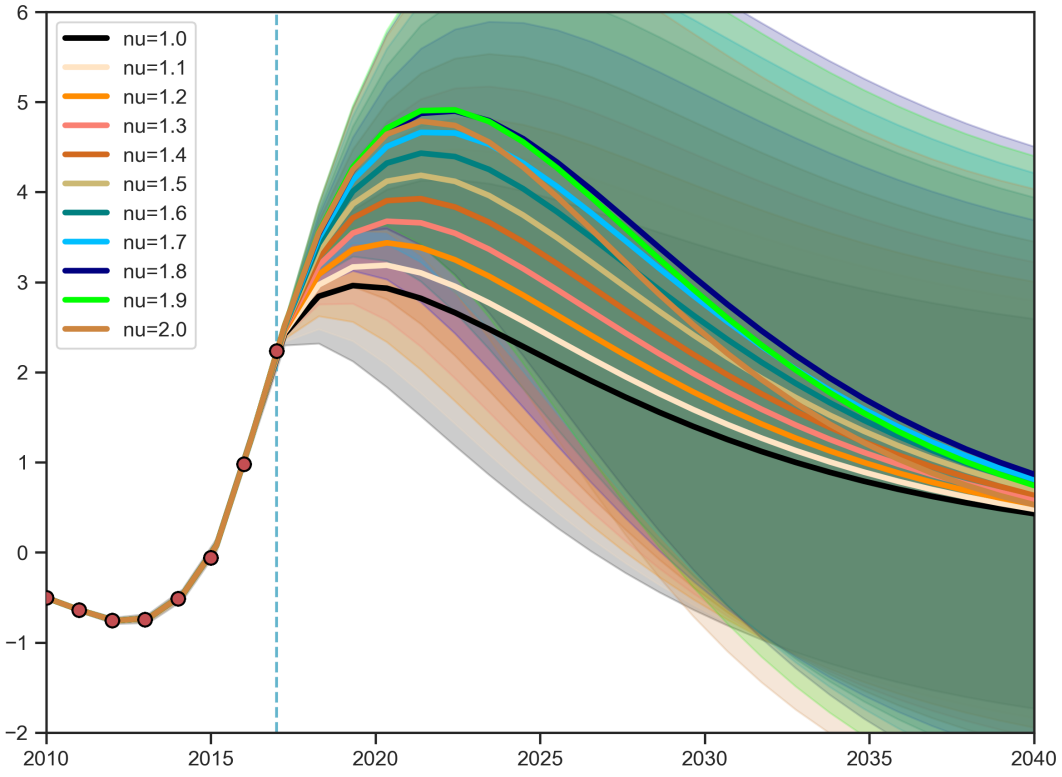


Figure 14: Sensitivity of ν

5 Model Modification Based on Socio-economic Data

5.1 Correlation Analysis with Socio-economic data

5.1.1 Construction of GRA Model

In order to find what socio-economic factors influence the opioid abuse, we firstly need to address the U.S. Census socio-economic data. After EDA of the data, we found that the data provided mainly includes the statistical data of various groups in a certain county, for example, different age groups and different group status (married, unmarried or divorced). However, some statistics have the problems of mismatch between the label and description. Also, there is some instability of time series, which brought difficulties and challenges to make accurate analysis. For such a knotty problem, we firstly address the mismatch problems, then apply the Gray Relational Analysis(GRA) to bring the influencing factor to light. GRA is firstly developed by Julong Deng in 1982 [4], and the advantage of GRA is that it analyzes according to the development trend, so there is no excessive requirement for the sample size. In addition, there is no need for a typical distribution, and the amount of calculation is relatively small. The results are also in good agreement with the results of qualitative analysis.

The process of establishing GRA model is as follows: define the data sequence reflecting the behavior characteristics of the system is called *reference sequence* while the data sequence of factors that affect the behavior of a system is called *comparison sequence*. Set the reference sequence as $Y = \{Y(k)|k = 1, 2, ..., n\}$ and comparison sequence as $X_i = \{X_i(k)|k = 1, 2, ..., n\}, (i = 1, 2, ..., m)$. Because the data in each factor column in the system may be different in dimension, it is not easy to compare or to get a correct conclusion when comparing. Therefore, we should carry out dimensionless processing of data as follow,

$$x_i(k) = \frac{X_i(k)}{X_i(l)}, k = 1, 2, ..., n; i = 0, 1, 2, ..., m$$

Then, we calculate the interrelationship between $y(k)$ and $x_i(k)$ and denoted as $\gamma(y(k), x_i(k))$. The calculation is as equation (17)

$$\gamma(y(k), x_i(k)) = \frac{\min_i \min_k |y(k) - x_i(k)| + \xi \max_i \max_k |y(k) - x_i(k)|}{|y(k) - x_i(k)| + \xi \max_i \max_k |y(k) - x_i(k)|} \tag{17}$$

Where ρ is called *resolution ratio*,and for best resolution, we set $\rho = 0.5$. Finally we could obtain Gray relational coefficient:

$$\gamma(Y, X_i) = \frac{1}{n} \sum_{k=1}^{n} \gamma(y(k), x_i(k)) \tag{18}$$

5.1.2 Influential Factors for Opioid Abuse

In order to find the possible influential factors, we subdivide the socio-economic data into three aspects: population composition, household type and educational at-

tainment. Household type intends to find the possible connection between the family structure and opioid abuse, and educational attainment

Population Composition

Population composition aims at identifying the most likely types of population which abuse opioids. Specifically, we use **Grey Correlation Analysis(GRA)** to get the correlation between the opioid abuse population and different groups of people including marriage status, veteran status, etc.

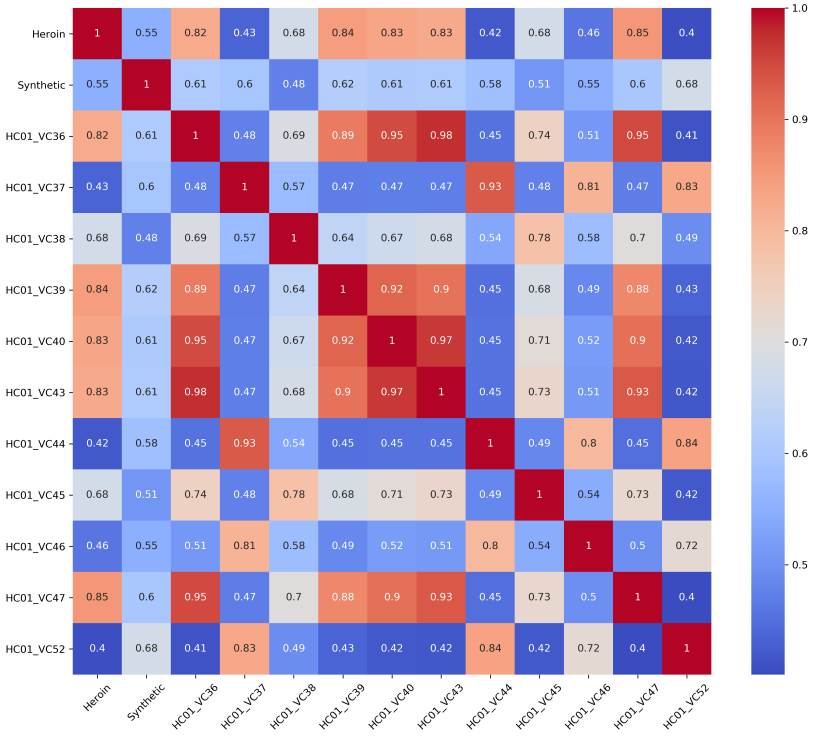


Figure 15: Grey Correlation Heat Map of Population Composition

From Figure 15, we can conclude that heroin abuse is strongly correlated with divorced women ($\gamma = 0.85$), widowed men ($\gamma = 0.84$), divorced men($\gamma = 0.83$), unmarried women($\gamma = 0.83$); Synthetic abuse is relatively highly correlated with unmarried women (widowed, divorced, and never married) who had a birth in the past 12 months ($\gamma = 0.68$). From which we can identify the main opioid users as divorced people, widowed men, unmarried women and unmarried women who have just given birth. We can infer their possible reason of abusing opioid despite knowing dangers is to dispel the unhappy depression of marriage.

Household Type

From Figure 22 given in appendix, we can conclude heroin abuse is strongly correlated with female householders without husband ($\gamma = 0.84$), population in households ($\gamma = 0.83$) and those non-family households whose householders live alone ($\gamma = 0.82$).

It is also highly correlated with male householders without wife ($\gamma = 0.81$) and the number of non-family households ($\gamma = 0.81$). Synthetic opioid abuse is relatively highly correlated with male householders who has no wife and their children are under 18 years old($\gamma = 0.65$). We can infer that those family householders without partners are more likely addicted to opioid, and a beatific family can effectively reduce the incidence of opioid abuse.

Educational Attainment

From Figure 21 given in appendix, we can conclude that heroin abuse is strongly related to college student with associate degree ($\gamma = 0.81$), graduate or professional degree ($\gamma = 0.8$), and bachelor's degree ($\gamma = 0.79$). Synthetic opioid abuse is relatively highly related to college students with bachelor's degree ($\gamma = 0.64$)and graduate or professional degree ($\gamma = 0.64$). Also, it is highly correlated with some colleges with no degree ($\gamma = 0.79$). We can infer that opioid are very popular among college students, especially in some colleges with poor education quality.

5.2 GBDT: Analysis of Regression Based on Gradient Boosting Machine

In previous model, as we are not provided with socio-economic data, so we can only predict the trend of abuse population based on the data itself. By doing so, it will cause the increase of subjectivity of probelm solving such as determining the threshold. Also, as it is not combinated with socio-economic data, so it is hard to identify the possible influential factors for opioid abuse as well as identify the possible strategy for countering the opioid crisis. In this section, we put emphasis on several representative socio-economic factors that might influence the heroin or synthetic opioid use in these five states and quantitatively describe their impact on eventual heroin and synthetic opioid use. Specifically, we will tackle this issue mainly from three aspects including educational attainment, residence change and marital status. For all three aspects, we choose the indicators of percentages.

5.3 Model Construction of GBDT

We now try to find the quantity relationship between the factors and opioid use. By analyzing the data as a whole, we find that the range of these percentages is relatively small, while the the range of corresponding opioid use is opposite. Therefore, a linear regression model cannot fit the data well. Additionally, although the nonlinear regression model (e.g. GPR) might fit this type of data well, its expression is implicit, which is somewhat challenging for the establishment of index threshold. Based on these restricted condition, we have to constitute a model that can well solve the above problems and should be provided with the following characteristics:

- It should be able to accommodate the various dimensions of the data.
- The model need to guarantee the ability to process nonlinear data.
- The prediction accuracy can also be high with relatively little time for adjust parameter.

- This model should be able to use some robust loss functions, making it very robust against outliers.

Based on these factors, we will use **Gradient Boosting Decision Tree(GBDT)** to solve this question. GBDT algorithm can be regarded as an addition model composed of K trees [6] that is:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{19}$$

where F is the functional space of all the trees. The boosting tree uses the addition model and the forward distribution algorithm to realize the learning optimization process, that is, a series of weak classifiers are obtained through iteration, and then the corresponding strong learners are obtained through different combination strategies. In the iteration of GBDT, it is assumed that the strong learner obtained in the previous round is $f_{t-1}(x)$, the corresponding loss function is $L(y, f_{t-1}(x))$. So the goal of the new iteration is to find a weak classifier $h_t(x)$ which could enable loss function $L(y, f_{t-1}(x) + h_t(x))$ to reach the minimum. In this problem, we choose loss function as:

$$L = \frac{1}{2}[y_i - f(x_i)]^2 \tag{20}$$

To solve the problem of general fitting method, Freidman proposed gradient boosting algorithm [6], which uses the approximation method of the **The Fastest Descent Approximation** which uses the negative gradient value of the loss function in the current model as the approximate value of the residual of the boosting algorithm in the regression problem as equation (21)

$$y_i - F(x_i) = -\frac{\partial \sum_i L(y_i, F(x_i))}{\partial F(x_i)} \tag{21}$$

The specific implementation steps of GBDT are as follows. Firstly, initialize the weak classifier and estimate a constant value that minimizes the loss function. At this time, the tree has only one root node.

$$f_0(x) = argmin_c \sum_{i=1}^{N} L(y_i, c) \tag{22}$$

Then, we perform M iterations of following steps:

(1) For $i = 1, 2, ..., N$, calculate the negative gradient value of the loss function in the current model and use it as a residual estimate

$$r_{mi} = -[\frac{\partial L(y, f(x_i))}{\partial f(x_i)}]_{f(x)=f_{m-1}(x)}$$

(2) Fit a regression tree for r_{mi}, then leaf node region of the m-th tree can be obtained and denoted as $R_{mj}, j = 1, 2, ..., J$

(3) For $j = 1, 2, ..., J$, calculate:

$$c_{mj} = argmin_c \sum_{x \in R_{mj}} L(y_i, f_{m-1}(x_i) + c)$$

It means the loss function is minimized by using linear search to estimate the value of leaf node region.

(4) Update the regression tree:

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} c_{mj} I(x \in R_{mj})$$

In the end, we could get the final output of the model:

$$\hat{f(x)} = f_M(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I(x \in R_{mj})$$

In addition, GBDT model can also rank the importance of features well. The basic idea is that if a certain node for judging a feature is farther away from the leaf node, the feature will be more important, and vice versa.

5.4 Identification of Severe Outbreak Areas

In this section, we still split opioid into two types: heroin and synthetic opioid. Among the socio-economic factors given by the supplementary data set, we select three main categories: educational attainment, residence change, and marital status. In each categories, there exists multiple subfeatures. Therefore, for a certain category, we initially identify the order of importance of subfeatures. Take synthetic opioid for instance. Figure 16 17 18 shows the importance of subfeatures for three categories in synthetic opioid use. From Figure 16, we know the percentage of population with residence change and whose population without residence change have the greatest influence on the use of synthetic opioid. For educational attainment, we know the percentage of associate's degree and that of educational attainment only in high school graduate affect the final result most. In terms of marital status, the proportion of those who are not married and divorced have the greatest impact on the final outcome. We will conduct in-depth analysis on these indicators.

In terms of marital status, we focus on two most important subfeatures: the proportion of unmarried people and divorced people. We then carry out GBDT regression for these two variables separately with synthetic opioids, and the results are shown in the Figure 19 and 20. From these figures, we can clearly get the influence of the proportion of unmarried or divorced people on the use of synthetic opioids. Specifically, we can get several points that are observed dramatic increases and its corresponding

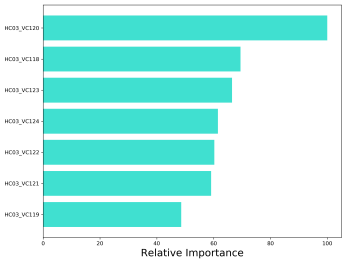


Figure 16: Importance in residence

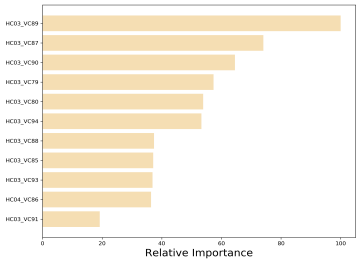


Figure 17: Importance in education

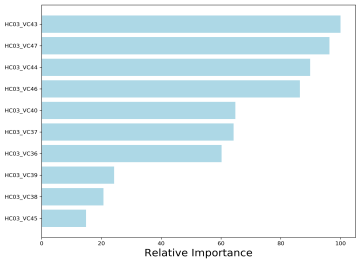


Figure 18: Importance in marriage

y-axis coordinate (Synthetic opioid in use) are defined as "Potential Flashpoint" which means when the number of synthetic opioid use reaches in this point, a slight percentage change in the corresponding feature element could cause a huge increase in opioid use. However, there also exist the situation that when this scenario occures, the slight change of corresponding feature could also cause a reduce. That is the reason why we call it "potential". Based on "Potential Flashpoint", thresholds can be determined objectively. For example, when synthetic opioid use reaches 97.6 per year, a little increase in proportion of divorced will cause an surge. Similarly, when synthetic opioid use attain 40.3, the little change of proportion of unmarried could arise a shoot up. In this premise, the thresholds can be obtained from the same analysis of other factors. The thresholds obtained by different factors are given in the table 3.

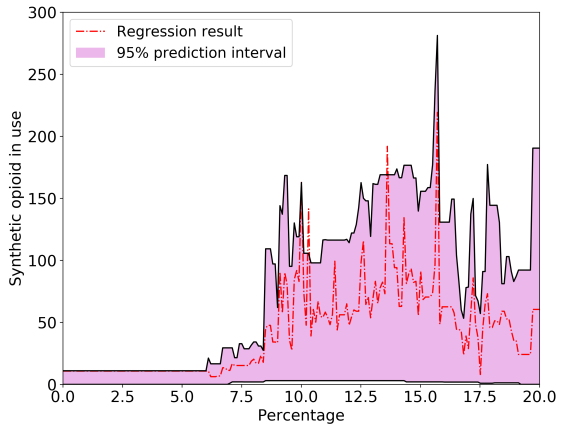


Figure 19: Divorced

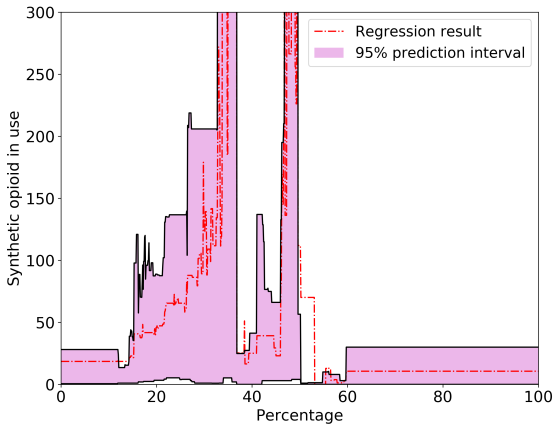


Figure 20: Never married

5.5 Strategies for Countering the Opioid Crisis

In this part, we will combine the model from the first two sections to to identify the possible strategies for counting the opioid crisis. Back to Figure 19 and 20, we can

Table 3: Potential Flashpoint with different index

Category	Percentage	Synthetic oipoid Potential flashpoint	Heroin Potential flashpoint
Marriage	Divorce	97.6	98.4
	Unmarried	40.3	62.3
	Widowed	42.7	47.9
Education	Associate's degree	37.2	72.1
	High school graduates	21.3	22.5
Residence	VC120	61.8	49.6
	VC118	53.2	50.1

VC120: RESIDENCE 1 YEAR AGO - Different house in the U.S. - Same county
VC118: Percent; RESIDENCE 1 YEAR AGO - Same house

get several points that are observed dramatic changes of abuse population. Based on that, the target of our strategy for controlling the further deterioration of drug abuse situation is to prevent the further development of abuse population before it reaches the potential flashpoint. According to this, thresholds for the percentage of subfeatures can be determined objectively by the abscissa of potential flashpoint. By calculation, the divorce rate in the region is required to be less than 8.3% or in 17.6% $\sim$ 20.1% ,then this county or state can avoid opioid outbreak. Similarly, the proportion of those who are unmarried needs to be kept below 31.3% or controlled in [39.2%, 47.7%]. In this premise, the intervals of those socio-economic factors can be obtained from the same analysis. The intervals for subfactors to prevent further deterioration of drug abuse are given in the table 4 and 5. It is important to mention that under those presented strategies, there is 95% chance of succcess to achieve the target of controlling the further deterioration of drug abuse situation, as we only attain the 95% confidence interval of the parameters.

Table 4: The percentage range that the impact factors need to be controlled(SYNTHETIC OPIOID)

Category	Subclass	Synthetic oipoid Less than(%)	Synthetic oipoid Controlled in(%)
Marriage	Divorce	31.3	[17.6, 20.1]
	Unmarried	8.3	[39.2,47.7]
	Widowed	5.7	[11.7, 16.2]
Education	Associate's degree	5.2	[7.5, 9.2]&[10.1, 12.5]
	High school graduates	13.3	[37.2, 55.7]
Residence	VC120	9.7	-
	VC118	-	[83.2,93.4]

VC120: RESIDENCE 1 YEAR AGO - Different house in the U.S. - Same county
VC118: Percent; RESIDENCE 1 YEAR AGO - Same house

For the governments of these five states, there are several policies that can be released. For preventing the divorce, more education should be given to new couples to make them respect the sanctity of marriage. Also, the government should introduce

Table 5: The percentage range that the impact factors need to be controlled(HEROIN)

Category	Subclass	Heroin	
		Less than(%)	Controlled in(%)
Marriage	Married(except separated)	29.1	[35.3, 44.5]&[47.6, 60.3]
	Unmarried	19.1	[38.2, 44.7]
Education	College or graduate school(enrollment)	9.7	[10.2, 16.3]&[23.3, 29.6]
	High school(enrollment)	-	[9.3, 16.0]
Residence	VC120	9.8	-
	VC118	-	[86.2, 93.3]

VC120: RESIDENCE 1 YEAR AGO - Different house in the U.S. - Same county
VC118: Percent; RESIDENCE 1 YEAR AGO - Same house

some marriage laws or policies to punish derailment in marriage. In the aspect of education, the government could consider slightly lowering the enrollment threshold for students, so that most students can obtain a bachelor's degree or above which can reduce the opioid crisis to some extent. For residence change, the authorities should enhance the proportion having permanent residence, they also need to control the floating population at a certain level.

6 Strengths and Weaknesses

6.1 Strengths

(1) **Proper data normalization and classification.** For the clustering and spreading model, proper data normalization help to do the classification better. After classification, the spreading model results better as it represents the diversity of the local situation.

(2) **Our model develops and expands the classical GPR model.** Different from the traditional GPR which use a single kernel function to predict the target function, our model applies three kernel functions to fit the target function together and the results are averaged. The reason of this approach is that there are a large number of time series with different characteristics in the data, and for a specific time series, we cannot directly select the most suitable kernel function, so it is reasonable to select multiple kernel functions to make the model more robust.

(3) **Our model has strong professional theoretical support.** In both GPR and GBDT models, we adopted a combination of subjective and objective methods to determine the evaluation threshold and were able to find the factors that had the greatest impact on the results.

(4) **Our model possess portability and result is provided with visualization.** The government can make the model more usable by changing some of the parameters in our model according to its actual needs. In addition, the results of our model are visualized, which makes the results easier to understand.

6.2 Weaknesses

(1) **We do not take full advantage of the given data.** In the second part of the given data, we divided the nearly 600 variables into several categories, and then analyzed the whole category. Only some of the most important variables were extracted according to the algorithm model.

(2) **We modeled only for five states, and this is not representative of the United States as a whole.** Due to the limitation of data, we only modeled for VA, OH, PA, KY, and WV, and only analyzed the interaction between them, without considering the influence of other regions.

References

[1] Rudd, Rose A. "Increases in drug and opioid-involved overdose deaths—United States, 2010–2015." MMWR. Morbidity and mortality weekly report 65 (2016).

[2] Rummans, Teresa A., M. Caroline Burton, and Nancy L. Dawson. "How good intentions contributed to bad outcomes: the opioid crisis." Mayo Clinic Proceedings. Vol. 93. No. 3. Elsevier, 2018.

[3] Sullivan, A. "The opioid epidemic is this generation's AIDS crisis." New York Mag (2017).

[4] Liu, Sifeng, Yingjie Yang, and Jeffrey Forrest. Grey data analysis. Springer: Berlin, Germany, 2017.

[5] Rasmussen, Carl Edward, and Christopher KI Williams.

[6] Gaussian process for machine learning. MIT press, 2006. Friedman, Jerome H. "Stochastic gradient boosting." Computational Statistics & Data Analysis 38.4 (2002): 367-378.

Appendices

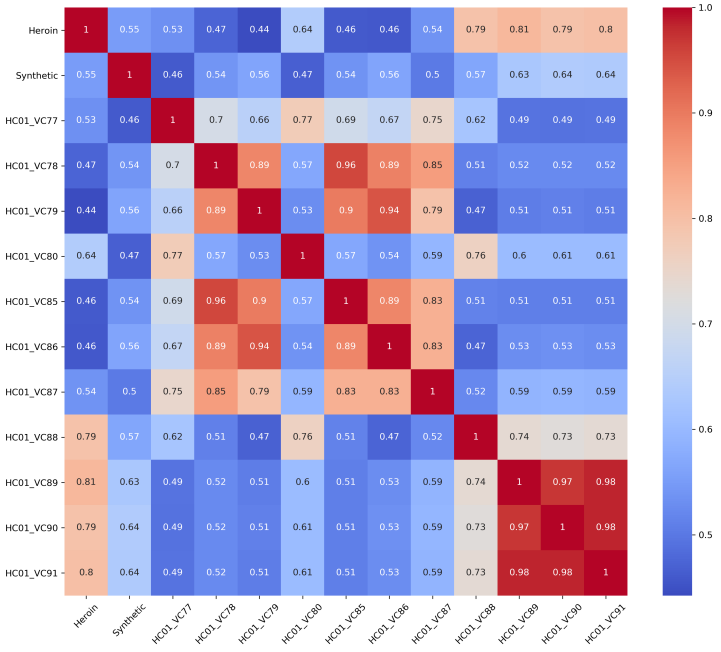


Figure 21: Grey Correlation Heat Map of Educational Attainment

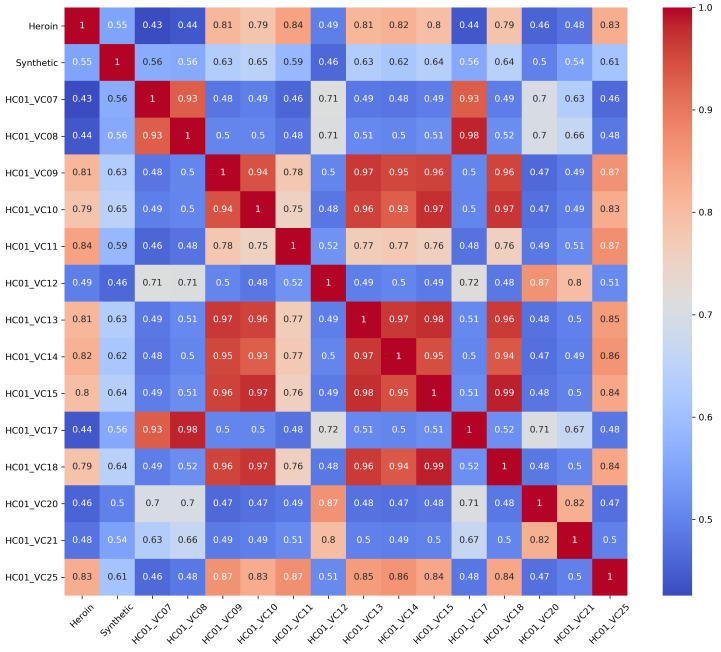


Figure 22: Grey Correlation Heat Map of Household Type

Heroin	Synthetic Opioid
PA, ALLEGHENY	PA, ALLEGHENY
PA, BUCKS	OH, BUTLER
OH, BUTLER	OH, CUYAHOGA
OH, CUYAHOGA	PA, DELAWARE
PA, DELAWARE	VA, FAIRFAX
PA, DAUPHIN	PA, DAUPHIN
OH, FRANKLIN	OH, CLARK
OH, HAMILTON	KY, FAYETTE
VA, HENRICO	OH, FRANKLIN
KY, JEFFERSON	OH, HAMILTON
OH, LAKE	VA, HENRICO
PA, LANCASTER	KY, JEFFERSON
PA, LUZERNE	KY, KENTON
OH, MONTGOMERY	OH, LAKE
PA, PHILADELPHIA	PA, LANCASTER
PA, YORK	OH, LORAIN
-	PA, LUZERNE
-	OH, MAHONING
-	OH, MONTGOMERY
-	PA, PHILADELPHIA
-	OH, SUMMIT
-	OH, TRUMBULL
-	OH, STARK
-	PA, YORK

Table 6: Counties where the government should adopt measures as soon as possible