For office use only
T1 _____
T2 _____
T3 _____
T4 _____

Team Control Number

1904686

Problem Chosen

C

For office use only
F1 _____
F2 _____
F3 _____
F4 _____

2019 Mathematical Contest in Modeling (MCM) Summary Sheet
(Attach a copy of this page to each copy of your solution paper.)

A solution to opioid abuse problem in Rust Belt

Abstract

We preprocessed the given data using the Moving Average method to make up for the exact data. If the number of missing data in a county is greater than 20% of the count of the sample, we discard the county's data. We used visualizing approaches to make the model and its application results intuitive and easy to understand.

In Part I: we defined the concept of **H(Heat)** to measure the degree of abuse of a certain type of drug in a county. H avoided the influence of develop situation of the state, because H is a ratio.

Then we drew a Heat Map of Virginia showed in figure 3. We visualized the principle of *the spread of the reported heroin incidents(cases) of counties in a state over time*. Moreover, we also visualized the principle of *the spread of the reported heroin incidents(cases) within different state over time*, the result is showed in figure 4.

We established a PCA(Principal Component Analysis) model to find the threshold of PCAS(PCA Score). The situation of each states is different means the thresholds are also different, so we applied PCA model on states one by one. Then we found the threshold and identified *any possible locations where specific opioid use might have started in each of the five states*. The locations are showed in table 2.

We established a SES(Second Exponential Smoothing) model to predict the DIC_{state} (drug identification counts of a state) and $HDIC_{county}$(heat of drug identification counts in a county). DIC_{county}(drug identification counts of a county) is the product of DIC_{state} and $HDIC_{county}$.

In Part II: We filtered some indexes(showed in table 4) based on the ρ(correlation coefficient) between DIC_{KY} and U.S census data, showed in 8. We select the indexes according to the following rules.The first rule is $0.4 \leqslant \rho \leqslant 0.95$. The second rule is census data are combined according to the real theme.

To modified model in Part I, we trained a random forest model from the sample from 2010 to 2015. The parameter of the model: t(count of regression trees) is 103, d(depth of each CART regression tree) is random between 1 and 20 and f(count of features) is random between 1 and 11. Then we evaluated the tree via predicting the DIC_{KY} in 2016, the $RMSE$(root-mean-square error) calculated 52 by equation 12.

In Part III: Based on the important influencing parameters analyzed in Part I and Part II, we identified possible strategies of countering Opioid Crisis of these parameters. Then, we used the SVP(single variable principle) to test the parameters and put them into the model, predicting the impact of these changes on the total amount of opioids in Kentucky. The test result is shoed in table 5. Finally, we gave some possible strategies to counter Opioid Crisis based on the test result.

Last but not least, We analyzed the strength and weakness about our model. Then we wrote a memo see about how to counter Opioid Crisis to the government.

Contents

Memo

From: Team 1904686, MCM 2019
To: The group of Governors
Date: January 28, 2019
Subject: Counter Opioid Crisis in Rust Belt

Dear Governors:

We are honored to offer you our advice, which comes from our research results.

As everyone knows, opioids are extremely harmful to society [7]. It not only harms the health of citizens, but also leads to a series of public safety and economic burden. Opioid drugs are spreading in the black market, and crimes are greatly increased [7]. At the same time, The abuse of opioids will also bring economic burden, in 2009, the economic burden of opioid poisoning was about 20.4 billion dollars. Productivity losses are associated with 89% of this total. Direct medical expenses are 2.2 dollars billion [8]. Therefore, it is necessary to introduce policies to control the spread and use of opioids.

In the first part of the mathematical model, we used known data to establish a heat map of the annual opioids in each state, which can directly tell the distribution characteristics of the drug. By analyzing the changing patterns and traffic routes of the annual distribution of these drugs, we can get a rough drug transfer route. At the same time, we used the SES(Second Exponential Smoothing) model to predict areas where opioids are most likely to be flooded in 2018 and 2019, which will be the basis for our next recommendations.

The Drug Enforcement Administration (DEA) can allocate more police force to the administrative district where these drugs are about to flood. The sudden increase in the use of opioids may mean the transfer of certain drug dens. Checkpoints should be placed on the poison line as indicated in our analysis report to cut off possible poison routes. At the same time, monitoring of sensitive local venues (bars, etc.) should also be strengthened. Because these places are places where drugs are traded and used more frequently. Strengthen control to reduce crime rate.

The Internet is also a factor we cannot ignore. With the popularity of the Internet, many new sales routes have appeared on the dark network, which makes it difficult for us to track. We must recruit more technicians to counter this means of poisoning.

Lack of control and free sharing of prescription drugs may also be an important reason of the spread of opioids, so it is meaningful to strengthen the construction of laws in areas where opioids are likely to occur in our forecast. Nearly 70% of those who take opioid prescription painkillers do not believe sharing the medications is a felony [9]. Sharing opioids is not a legal act, we must publicize this idea with the public. At the same time, the sale of prescription drugs must be more stringent (including pet medication, because we have known that some addicts will abuse their pets to get prescription drugs) to reduce the circulation of opioids. We should also allocate a portion of the money for training for local doctors. Prescribers should be more cautious when prescribing and take responsibility for introducing patients to the pros and cons

of opioids.

In the mathematical model of the second part, we identified some important factors about the population related to the addiction of opioids, which can affect the use of opioids. We should develop targeted strategies to change these factors.

We should review applicants of immigrants and reject potential drug users and drug traffickers. Care for unmarried people and solitary people to reduce their drug use rate. For minority drug abuse, we provide better medical care and promote the harm of opioids. For the above strategy, we have explained in detail in Part3 and used the model for feasibility study.

In addition, naloxone is provided free of charge in areas where opioids are frequently used. Naloxone, now available as a nasal spray, immediately blocks the deadly respiratory suppression caused by heroin, methadone and narcotic pain pills (like OxyContin, Percodan and Vicodin), and it should be made easily available to first responders, families and those dependent on narcotics and their friends [10]. it is also available as an auto-injector, which allows people without medical training to inject people who have overdosed [11].

1 Introduction

1.1 Background of Opioid Abuse in Rust Belt

The rust belt starts in the central part of New York State and crosses the west through the lower peninsula of Pennsylvania, West Virginia, Ohio, Indiana, and Michigan. New England was greatly influenced by the decline of the industry of the same era. Since the middle of the 20th century, industries in this area, formerly known as the industrial center of the United States, were declining due to various economic factors such as overseas relocation of manufacturing industry, expansion of automation, decline of the United States. Iron and Steel and Coal Industry Although some cities and towns have been successfully adapting themselves to focus on services and high-tech industries, others have seen increasing poverty and declining population, I have not failed as well.[1]

Opioid crisis has been increasing rapidly in the United States and Canada since the late 1990s, with the use of prescription and commercial opioids increasing rapidly and will last for 20 years. The increase in overdose of opioids is very large, and now opioids have saused 49,000 deaths until 2017 and deaths due to overdose of 72,000 drugs in the United States[2]. The long-term usage rate of opioids is increasing worldwide. [3]

It is obviously that if the government cannot take effective measures to deal with the phenomenon, it will cause serious social problems. Because if the abuse of opioids spreads to people with professional skills, it may result in a lack of certain professional human resources. If the proportion of drug addiction in the elderly increases, it will undoubtedly increase the pressure on the health care system. Therefore, the establishment of an opioids research model is of practical significance which can help us solve many important social problems.

1.2 Our Work

1.2.1 In Part I

We mainly finished the following work.

- Preprocess the data, make the data correct and reliable.
- Establish a mathematical model to describe the spread characteristics of the reported synthetic opioid and heroin incidents (cases) in and between the five states and their counties over time.
- Use the model to distinguish possible locations where specific opioid use might have started in each of the five states.
- We built a PCA model(showed in figure 1)and predict some certain place where may occur opioids abuse.

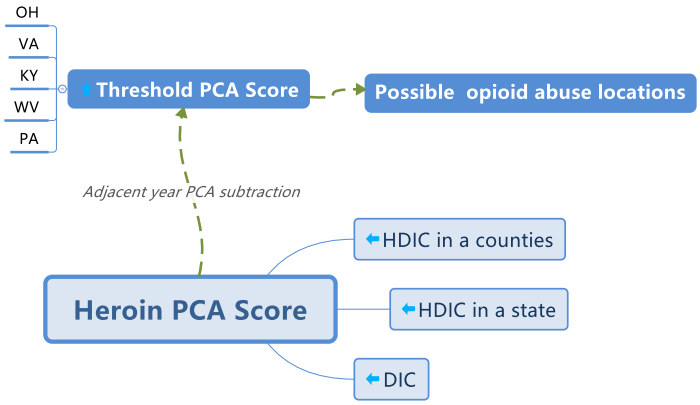


Figure 1: PCAmodel

1.2.2 In Part II

In this part we treat the opioids abuse as the results, it composed by many infectors and contributors. We found the main causes of the opioids and how they affect the opioids abuse.

- We collected information from demographic data and found the main population in drug abuse.
- We calculated the correlation coefficient between the possible impact factors and drug abuse indicators, and find the most relevant factors on drug abuse.

1.2.3 In Part III

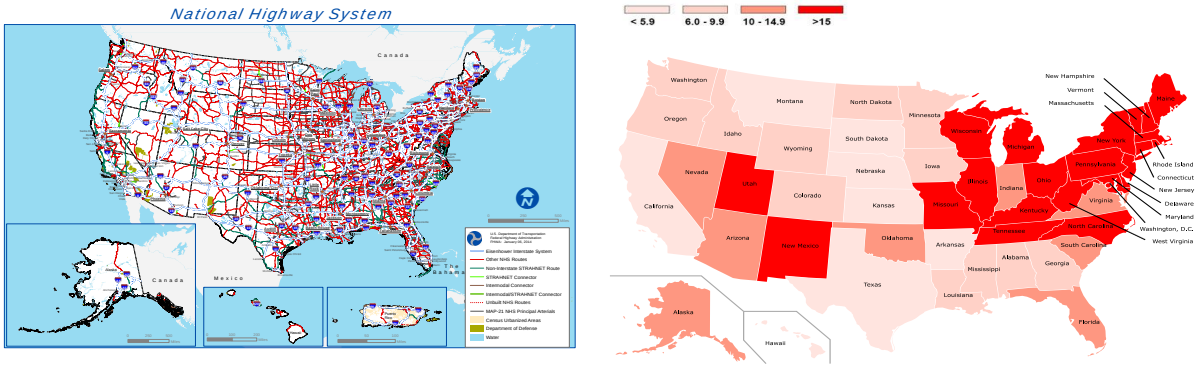
- Based on the former parts, we identified some possible strategies for countering the opioid crisis. Using the strategies and the models, we evaluated our work and found some constructive improvements.
- Give some advice to the government to prevent the adverse consequences of the abuse of opioids. If the result of drug identification comes to a threshold level, the government of the state should take actions to stop the bad tendency.

2 Assumptions and Notations

2.1 Assumptions

- The population of each county does not suddenly change. The proliferation of drugs and the abuse of drugs is a gradual process of change, which is a prerequisite for predicting change.

- For these five states all lay in inland, we assume that the main mode of transportation of drugs is land transport, rather than airline or waterway. In addition, the roads in these five states are well developed. It is showed in figure 2a. Furthermore, aviation security is very strict, so drugs are not easily transported by air. And the five states are in close proximity to each other. The cost of airline is unacceptable.



(a) Possible counties (b) The opioid use in 2017 of USA

Figure 2: Maps Associated to the problem

- Do not consider the effects of other states adjacent to the five states on opium abuse in these five states. According to figure 2b, the opioid abuse in these five states is far more serious than other states.[4] Therefore, we assume that there is no significant correlation between opium abuse in these five states and other states.

2.2 Notations

The overall situation notations defined in table 1.

Table 1: Notations

Symbol	Definition
DIC	Drug Identification Counts.
DIC_t	Drug Identification Counts in year t.
DIC_{state}	State Drug Identification Counts.
DIC_{county}	County Drug Identification Counts.
Y_t	The year t
$PCAS$	PCA score
X	A random variables in statistics

3 Data Preprocessing

For big data problems, there are usually abnormal data and missing data. These data are fatal to the stability of the model, so data cleanup is necessary. We divided the problem into two types.

3.1 Processing of abnormal data

If a value in a set of data is more than twice the standard deviation of the average, we call it the abnormal value. We use box plot analysis to detect anomalous data and replace it with the mean of the adjacency data of the anomaly data.

3.2 Processing of missing data

If this indicator has 20% or more of data missing, we will discard the data directly. If this indicator has 20% and the following is missing, we use the moving average window method to make up missing data. Since since we only have 8 years data, our sliding window is 1.

We use Simple Moving Average (SMA).

$$SMA = \frac{1}{n}\left(p_1 + p_2 \cdots + p_n\right) \tag{1}$$

In this problem,if DIC_t is missing, we use equation (2) to make up it.

$$DIC_t = \frac{1}{2}\left(DIC_{t-1} + DIC_{t+1}\right) \tag{2}$$

4 Part I

4.1 Spread Principle of synthetic drug incidents (cases)

Since we need to analyze the spread of synthetic opioid, we must eliminate the DIC samples of non-synthetic drugs. For example, cocaine is not a kind of synthetic drug[5], so the DIC of it should be discarded.

4.1.1 In A State

By referring to 2a, we can find that these states are close together, which is the base of giving *the spread of the reported synthetic opioid and heroin incidents (cases) in and between the five states and their counties over time*.

Because of the differences in population, area, and other factors in each state and county, it is not objective to use DIC for comparison. Therefore, we use a ratio to

measure the degree of abuse of a kind of drug in a county. We call it $HDIC$(heat of DIC).

Taking Virginia (VA) as an example, we give a detailed method for determining *the spread of the reported heroin incidents (cases) in and between counties in each state over time*. The $HDIC$ of heroin $HDIC_{heroin}$ in a county is calculated by equation (3)

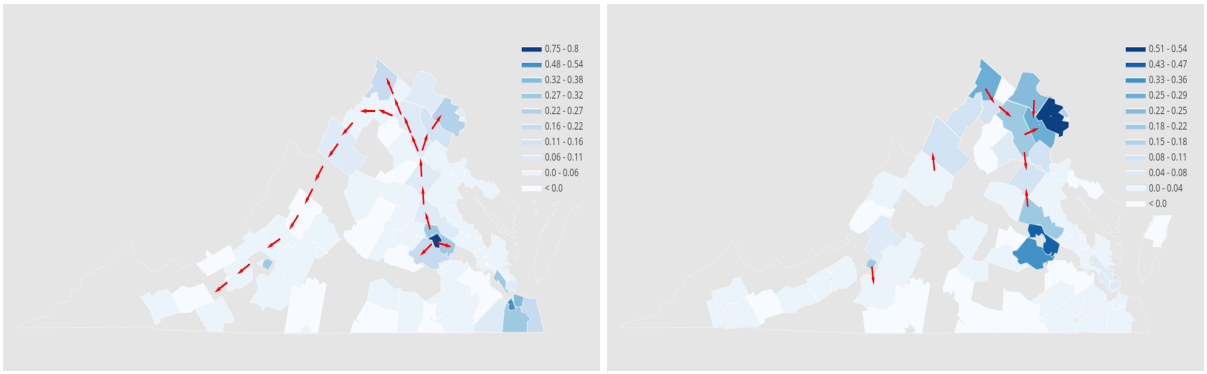
$$HDIC_{county}\left(herion\right) = \frac{DIC_{county}\left(heroin\right)}{DIC_{state}} \tag{3}$$

We plotted the maps of $HDIC_{heroin}$ from 2010 to 2017 in each state, which is showed in figure 3. The red arrows in the heat map represent the spread of the reported synthetic opioid and heroin incidents (cases) in each counties of Virginia over time. Other analysis results are presented in the Appendix A.1.

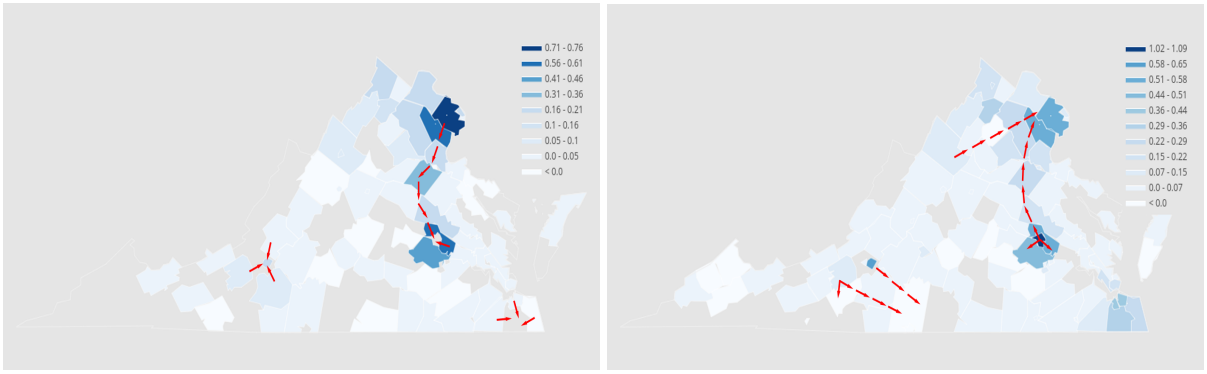
4.1.2 Between Five States

The spread of the reported heroin incidents (cases) between the five states over time is expressed by red arrows in figure 4. The red arrows show the possible drug routes. The *the spread routes of the reported heroin incidents (cases) between the five states over time* are follows.

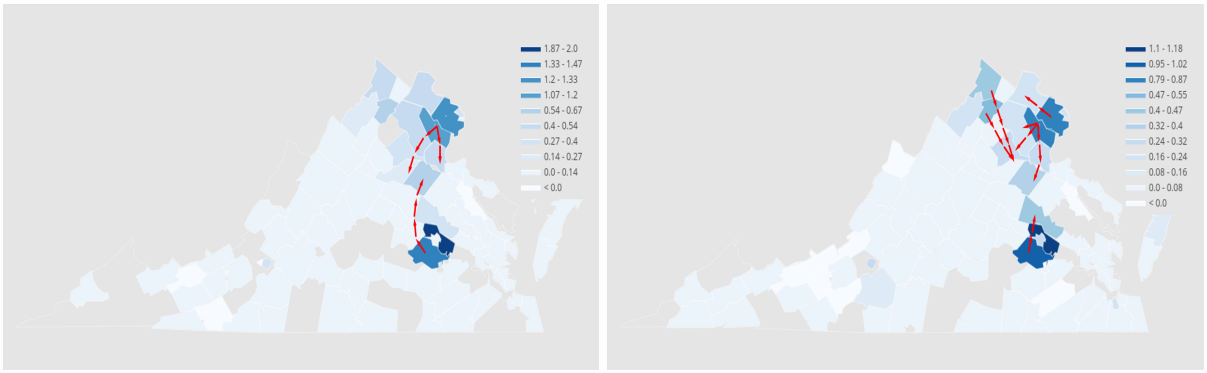
- *2010:* Mainly transferred within Pennsylvania.
- *2011:* Mainly from Ohio to Kentucky.
- *2012:* Mainly dispersed from Virginia.
- *2013:* Mainly from Ohio to Kentucky.
- *2014:* Mainly from Pennsylvania and Kentucky to West Virginia.
- *2015:* Mainly from Pennsylvania and Kentucky to West Virginia. Similar to route in 2014.
- *2016:* Mainly from Pennsylvania to West Virginia.



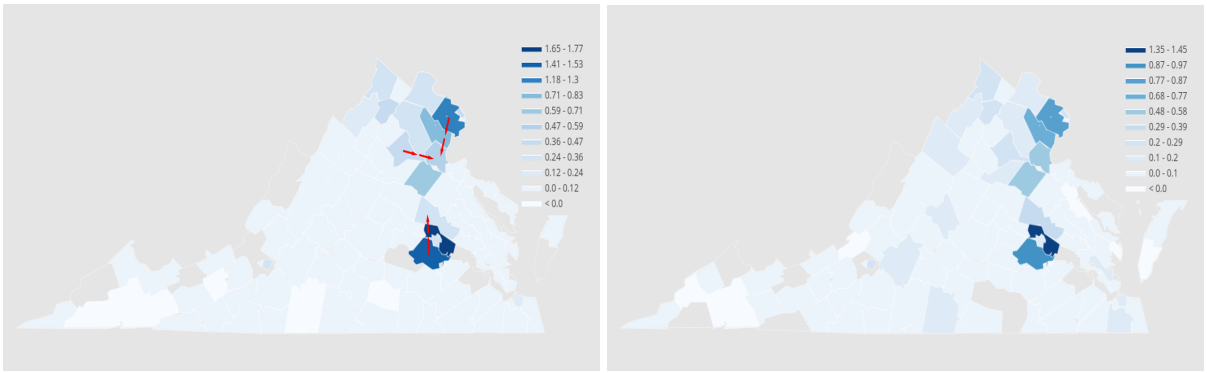
(a) 2010 (b) 2011



(c) 2012 (d) 2013

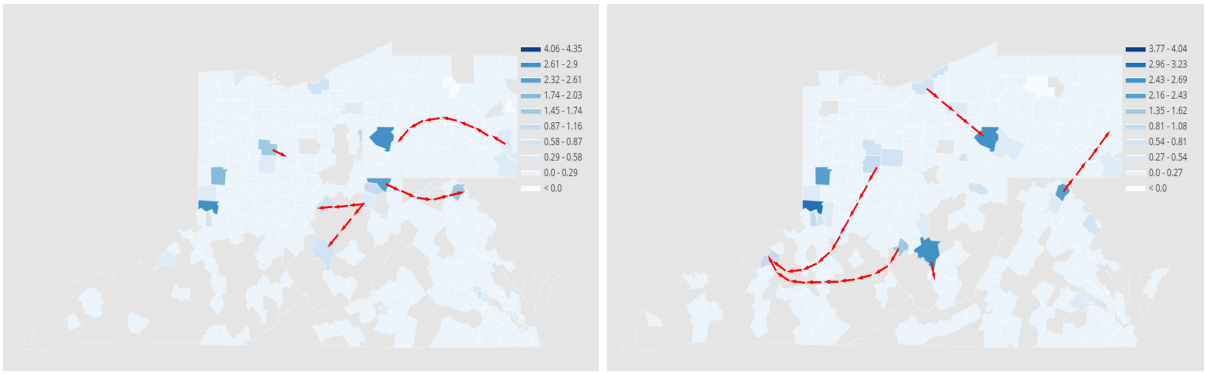


(e) 2014 (f) 2015

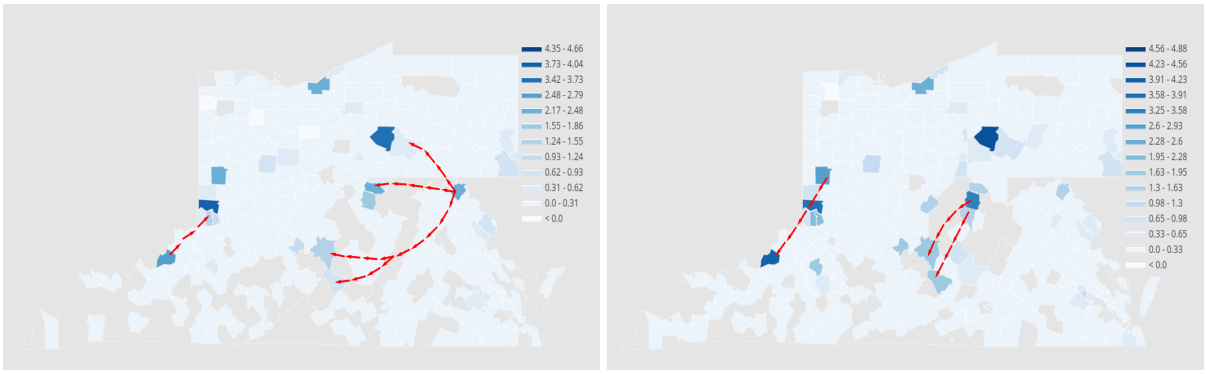


(g) 2016 (h) 2017

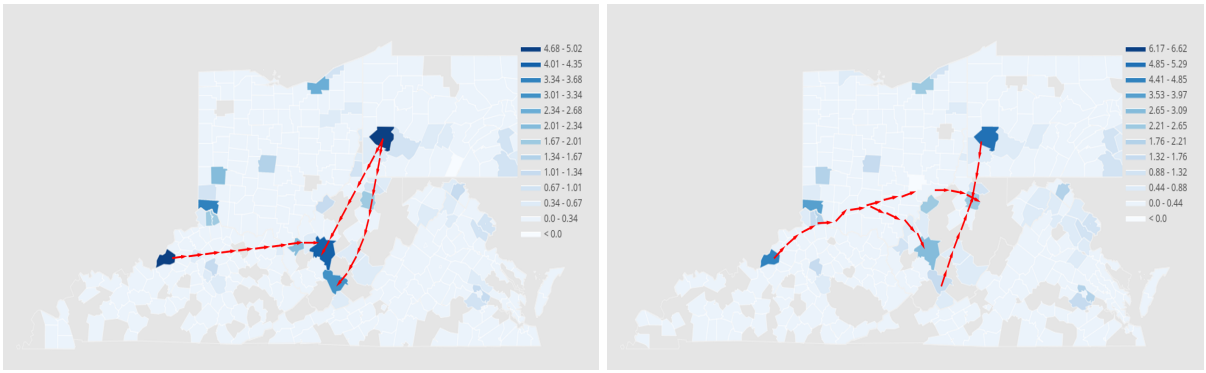
Figure 3: The Spread of Heroin Counties of VA



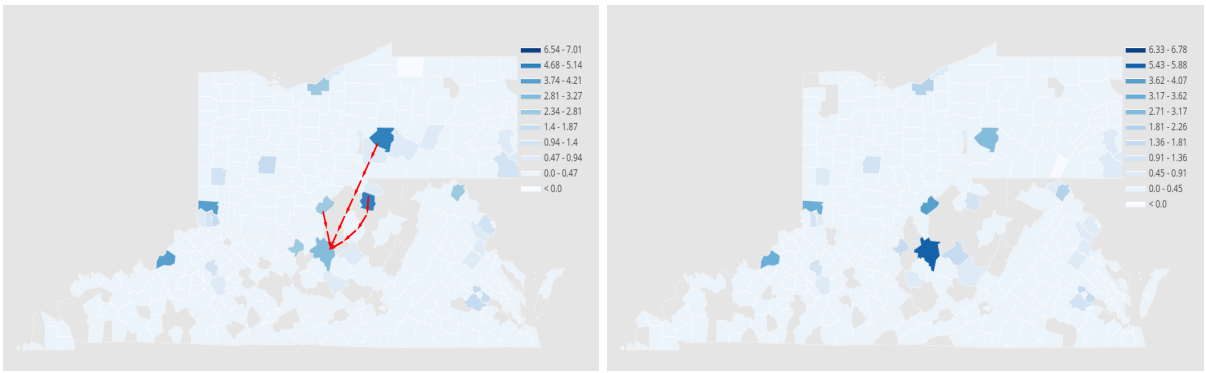
(a) 2010 (b) 2011



(c) 2012 (d) 2013



(e) 2014 (f) 2015



(g) 2016 (h) 2017

Figure 4: The Spread of Heroin Between Five States

4.2 Identify Specific Opioid Abuse Places

We use PCA to find the Principal Component of some indexes, they are the heat of heroin $HDIC(heroin)$ in a state $HDIC_{state}\,(herion)$,the $HDIC$ of heroin $HDIC_{heroin}$ and DIC of herion $DIC\,(herion)$. Then we rank PCA score($PCAS$) and find the threshold of PCAS.($TPCAS$)

Base on $TPCAS$ we identified any possible locations where specific opioid use might have started in each of the five states.

4.2.1 Compute the PCA Score

First, we calculated $HDIC_{state}\,(herion)$ by equation (4). $HDIC_{county}(heroin)$ is calculated by equation (3) and $DIC(herion)$ is given.

$$HDIC_{state}\,(herion) = \frac{DIC_{state}\,(herion)}{DIC_{state}} \tag{4}$$

Second, we used euqation (5) to normalize the data. $\hat{a}_{ik}$ is the data after being normalize and A_{ik} is the initial data.

$$\begin{aligned} \hat{a}_{ik} &= \frac{a_{ik} - \mu_k}{s_k} \\ \mu_k &= \frac{1}{n}\sum_{i=1}^{n} a_{ik} \\ s_k &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(a_{ik} - \mu_k\right)^2} \end{aligned} \tag{5}$$

Then we computed c_{ij}(coefficient of association matrix) by equation (6)

$$c_{ij} = \frac{\sum_{k=1}^{n} \hat{a}_{ik} \cdot \hat{a}_{kj}}{n-1} \tag{6}$$

Next we calculated the eigenvalue w_i and eigenvector $\vec{t_i}$ of coefficient of association matrix by equation (7).

$$y_i = c_{ij}\begin{bmatrix} \hat{a}_{11} & \cdots & \hat{a}_{1n} \\ \vdots & \ddots & \vdots \\ \hat{a}_{n1} & \cdots & \hat{a}_{nn} \end{bmatrix} \tag{7}$$

Finally, the $PCAS$(PCA score) is calculated by equation (8). Meanwhile, we ranked the $PCAS$ of each counties over time, the result is showed Appendix A.

$$\begin{aligned} b_j &= \frac{w_i}{\sum_{t=1}^{n} w_t} \\ PCAS &= \sum_{j=1}^{n} b_j y_j \end{aligned} \tag{8}$$

4.2.2 Find the PCA Score Threshold

First, the situation of each state is different, such as area,population etc., so we executed the PCA algorithm on one state after another. Then we get $\Delta PCAS$ between each two adjacent $PCAS$. Next we found the threshold of $PCAS$.($TPCAS$) by equation (9).The threshold of each state is showed in table 2.

$$\begin{aligned} TPCAS &= \frac{1}{2}\left(PCAS_t + PCAS_{t+1}\right) \\ s.t. &\Delta PCAS_{t,t+1} \rightarrow \max \end{aligned} \tag{9}$$

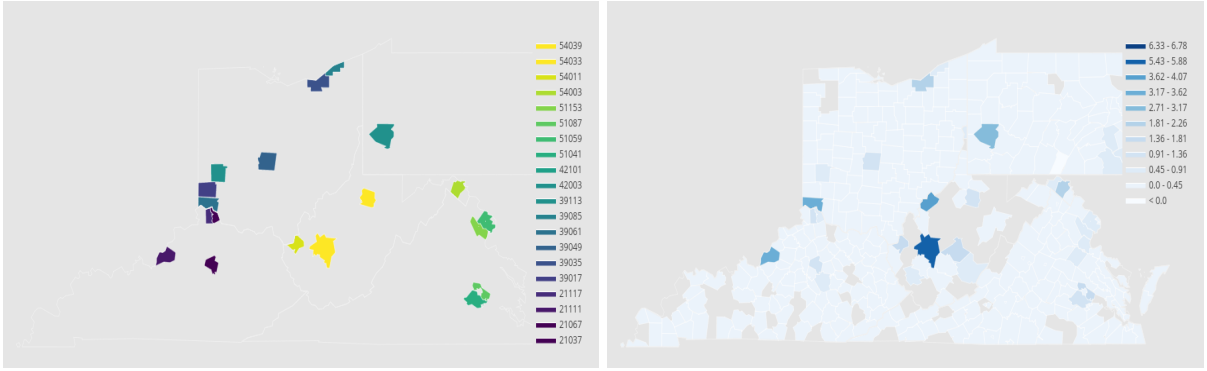
4.2.3 Identify Possible Opioid Abuse Place

Base on the $PCAS$ and $TPCAS$ of each counties, we identified *any possible locations where specific opioid use might have started in each of the five states*. The result is also showed in table 2 and visualized in figure 5. The counties on the map are the possible opioid abuse locations in the five states.

Table 2: Threshold of PCA and Possible Opioid Abuse Counties

Satate	PCA_KY	PCA_OH	PCA_PA	PCA_VA	PCA_WV
Threshold	865.04477	809.9552937	3306.060233	859.1580341	408.0049
Counties	*JEFFERSON* *KENTON* *CAMPBELL* *FAYETTE*	*HAMILTON* *CUYAHOGA* *MONTGOMERY* *FRANKLIN* *LAKE* *BUTLER*	*PHILADELPHIA* *ALLEGHENY*	*HENRICO* *FAIRFAX* *CHESTERFIELD* *PRINCE WILLIAM*	*KANAWHA* *HARRISON* *BERKELEY* *CABELL*

Any possible locations where specific opioid use might have started in each of the five states are associated with each condition of a year. We compared our identify result with the 2017 opioid abuse of each counties as a sample to test and verify our PCA predict model.



(a) Possible counties (b) The opioid use in 2017

Figure 5: Possible opioids abuse place

4.3 Predict Drug Identification Counts

We first established a Regression Tree Model. Using this model, we performed regression prediction on DIC. The MSE is calculated 0.096. Since the amount of data given is small and the fluctuation is large, we establish a quadratic exponential smoothing model to make the prediction again, and get the MSE of 0.065, which is smaller than the prediction result of the regression tree. We established a SES(Second Exponential Smoothing) model to predict DIC.

By analyzing the historical observation data and the data given, we find that these data have these characteristics, the volatility is relatively obvious, and the number is relatively small. Therefore, it is suitable to use the second exponential smoothing(SES) model to predict Drug Identification Counts in each state.

Based on the heat of some kind of drug in each counties and the state prediction drug identification counts, we can predict the county Drug Identification Counts in each state.

We have applied the predict model on heroin data of KY.

4.3.1 In A State

The SES model is described by equation (10), the $S_t^{(2)}$ and $S_{t-1}^{(2)}$ is the SES values in t and $t1$ periods.

$$S_t^{(2)} = aS_t^{(1)} + (1-a)\, S_{t-1}^{(2)} \tag{10}$$

If the $S_t^{(2)}$ and $S_{t-1}^{(2)}$ is known, the SES model is described by equation (11).

$$\begin{aligned} \hat{Y}_{t+T} &= a_t + b_t \cdot T \\ S_t^{(2)} &= aS_t^{(1)} + (1-a)\, S_{t-1}^{(2)} \\ a_t &= 2S_t^{(1)} - S_t^{(2)} \\ b_t &= \frac{a}{1-a}\left(S_t^{(1)} - S_t^{(2)}\right) \end{aligned} \tag{11}$$

T is the predicted number of advance periods, and a is a constant.

First we calculated the initial $a = 0.6$, $S_0^{(2)} = S_0^{(1)} = y_1 = 629$. Then we used equation (11) to calculate the exponential smoothing value of each period.The parameter α of the second exponential smoothing model is 1, β is 0.1.

The prediction of Kentucky State's Heroin identification counts are shown in table 3 and visualized in figure 6.

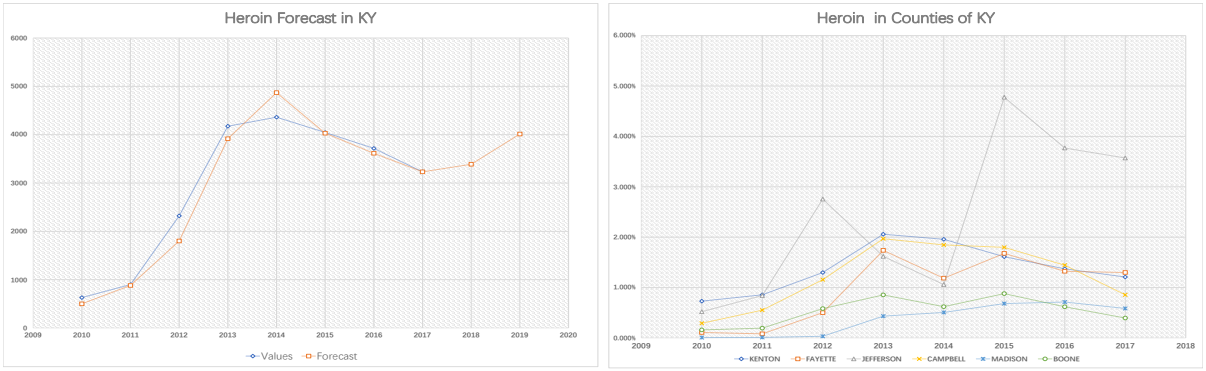
Finally, we calculate the sum of variance(SSE) and root mean square(MSE) of the predicted and observed values to estimate the accuracy of the model:(12).The SSE is 0.95 and MSE is 0.22. The closer MSE and $RMSE$ are to 0, the better the prediction will be, so we can use the second exponential smoothing model to predict the

Table 3: Prediction Kentucky State's Heroin identification counts

Timeline	Values	Forecast
2010	629	500
2011	899	880
2012	2320	1800
2013	4175	3916
2014	4362	4870
2015	4045	4030
2016	3716	3615
2017	3231	3231
2018		3387.63
2019		4012.89

$HDIC_{state}(heroin)$ showed in figure 6.

$$
\begin{aligned}
SSE &= \sum_{i=1}^{m} w_i \left(y_i - \hat{y}_i\right)^2 \\
RMSE &= \sqrt{MSE} = \sqrt{\frac{SSE}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} w_i \left(y_i - \hat{y}_i\right)^2}
\end{aligned}
\tag{12}
$$



(a) DIC(hreoin) Forecast of KY (b) Top DIC(heroin) counties of KY

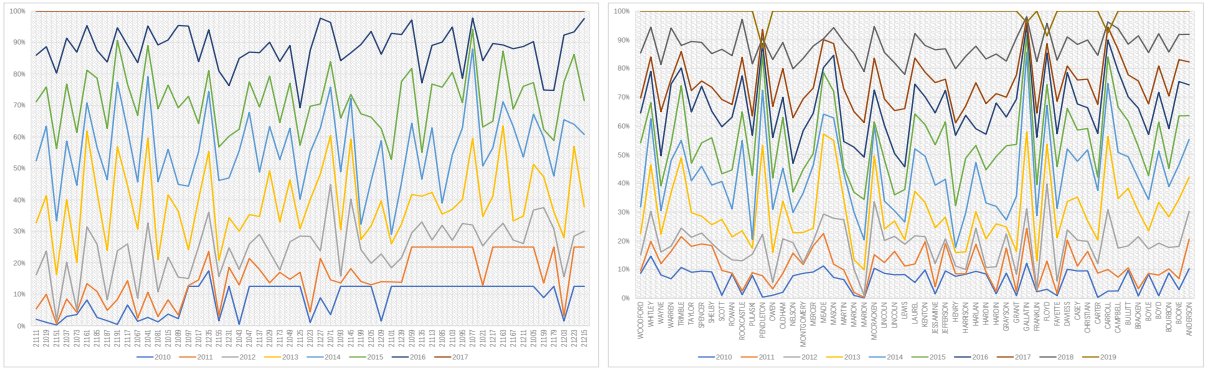
Figure 6: DIC(heroin) of KY

4.3.2 In A County

First, we use the SES model to predict the $HDIC_{county}$ of the counties in a state. The predict result is showed in Appendix A table6 and visualized in figure 7.

After predict the DIC_{state} and $HDIC_{county}$, base on the equation (3) we calculate the Drug Identification Counts by a variation of equation(3), namely the equation (13).

$$
DIC_{county}\left(herion\right) = HDIC_{county}\left(herion\right) \cdot DIC_{state} \tag{13}
$$



(a) HDIC(Heroin) of each counties in KY (b) Predict HDIC(Heroin) in each counties of KY

Figure 7: DIC(heroin) of counties in KY

5 Part II

5.1 The Impact Factor of Opioids Abuse

We chose the DIC_{KY} as an example to apply our model. We first removed the missing values from the provided U.S. Census socio-economic data and filtered out the index associated with DIC_{KY} from census data.

5.1.1 Filter Out The Index From Census Data

We treated DIC_{KY} as observation values and calculated the correlation coefficient between the DIC_{KY} and each index of census data.

First we calculated the correlation coefficients between each index of census data and DIC_{KY}.$X = [x_1, x_2, \cdots x_n]andY = [y_1, y_2, \cdots y_n]$ are two random variables in statistics. The $Cov(X, Y)$ is the covariance of X and Y.ρ_{XY} is correlation coefficient or standard covariance of random variables X and Y.

Calculations step of correlation coefficients:

- Calculate $Cov(X, Y)$:

$$\operatorname{Cov}(X, Y) = E\left\{[X - E(X)] \cdot [Y - E(Y)]\right\} \tag{14}$$

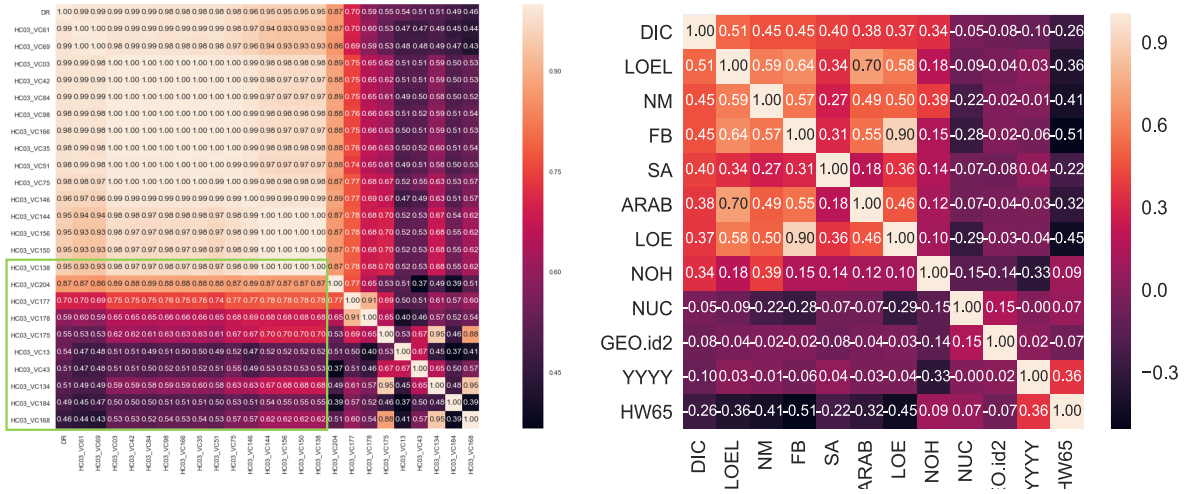
- Calculate ρ_{XY}:

$$\rho_{XY} = \frac{\operatorname{Cov}(X, Y)}{\sqrt{D(X) \cdot D(Y)}} \tag{15}$$

- Determining the type of relevance

$$\rho_{XY} = \begin{cases} 1 & \text{Positive correlation} \\ -1 & \text{Negative correlation} \\ 0 & \text{Independently irrelevant} \end{cases} \tag{16}$$

The correlation coefficient matrix is shown in figure 8. DR is DIC_{KY}, $HCXX_VCXX$ is the ID of each county in KY state, which only acts as an identifier.



(a) Correlation Coefficient Matrix of all index (b) Correlation Coefficient Matrix of filtered index

Figure 8: Correlation Coefficient Matrix

We found that in the census data, the population-related statistics were positively correlated with the DIC_{KY} via correlation analysis. In theory, we should think that census data has a strong correlation with DIC_{KY}, people are the user of drugs, simply considering the population is not a smart way to deal with the abuse of opioids. We select the indicators according to the following rules:

- The census data with correlation greater than 0.4 and less than or equal to 0.95 are extracted.
- The census data are combined according to the theme.

Finally the table 4 was obtained, which gives the non-population index which is correlated with DIC_{KY} of the census data. The correlation coefficient matrix is showed in figure 8.

Table 4: Index Census Data correlated to DIC_{KY}

Dimension	Abbreviation	Meaning
ANCESTRY	SA	Subsaharan African
	ARAB	Arab
HOUSEHOLDS BY TYPE	HW65	Households with one or more people 65 years and over
LANGUAGE SPOKEN AT HOME	LOEL	Language other than English, Other languages
PLACE OF BIRTH	FB	Foreign born
U.S. CITIZENSHIP STATUS	NUC	Naturalized U.S. citizen
LANGUAGE SPOKEN AT HOME	LOE	Language other than English
HOUSEHOLDS BY TYPE	NOH	Nonfamily households
MARITAL STATUS	NM	Never married

5.1.2 Modify The Model in Part I

We modified the regression tree model in Part I (see in 4.3). We trained a random forest model composed by regression trees,

The random forest is made up of several CART(Classification And Regression Tree). For each tree, the training set they use is back-sampled from the total training set, which means that some samples in the total training set may appear multiple times in a tree's training set, or they may never Appeared in the training set of a tree. When training the nodes of each tree, the features used are randomly extracted from all the features in a certain proportion without random return.

The process of building a CART regression tree:

- The training set is $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ D refers to the training data set, which represents the impact factors and observations of the sample.
- The output Y is a continuous variable, and the input is divided into M regions, which are respectively $R_1, R_2, \cdots, R_m$. The output values of each region are $c_1, c_2, \cdots, c_m$, then the regression tree model can be expressed as equation 17

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{17}$$

- Then the square error is calculated by equation 18

$$\sum_{x_i \in R_m} (y - f(x_i))^2 \tag{18}$$

- Use s which is the value of the feature j to divide the input space into two regions respectively, they are $R_1(j, s) = \{x | x^{(j)} \leqslant s\}$ and $R_2(j, s) = \{x | x^{(j)} > s\}$
- We need to minimize the loss function. Lose function is the description of error of training. This step is conveyed by equation 19

$$\min_{j,s} \left[\sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \sum_{x_i \in R_2(j,s)} (y_i - c_1)^2 \right] \tag{19}$$

  Where c_1 and c_2 are the average values of the outputs in the R_1 and R_2 intervals, respectively.

A random forest is composed by CART regression trees. The steps of building a random forest are follows:

- **Step 1.** D(*Determine training set*), T(*test set*), F(*feature dimension*). In our model, training set is the index in table 4 and DIC_{KY}. We use $K - Ford$ method to test our model. F of our model is the index in table 4.

  Determine the parameters of our model, the t(*count*) of CART regression trees is 103, the d(depth) of each CART regression tree is random between 1 and 20 and f(*the count of features*) is random between 4 and 11.

  When the sample used in nodes is minimum, one node of tree is generated.

- **Step 2.** The $D(i)$(*training set*) with the same extraction scale as D from D, as the sample of the root node, starts training from the root node.

- **Step 3.** If the termination condition is reached on the current node, the current node is set as a leaf node, and the predicted output is the average value of each sample value of the current node sample set. Then continue to train other nodes. If the current node does not reach the termination condition, the f-dimensional feature is randomly selected from F(*the feature dimension*) without being put back. Using this f-dimensional feature, k (*the best-dimensional one-dimensional feature*) and its b(*boundary*) are searched. The sample with the $k - b$ dimension of the sample on the current node is less than b and is divided into the left node, and the rest is divided into the right node. Continue to train other nodes.
- **Step 4.** Repeat step 2 and 3 until all nodes have been trained or marked as leaf nodes.
- **Step 5.** Repeat step 2,3 and 4 until all CART have been trained.

We built a random forest model from the sample from 2010 to 2015 based on the rules above. Then we evaluated the model via predicting the DIC_{KY} in 2016, the root-mean-square error $RMSE$ calculated 52 by equation 12.

6 Part III

6.1 Some possible strategy

From some background knowledge, we gave some possible strategies as follows:

- **Strategy 1.** The correlation coefficients of these three parameters NUC, EO.id2, YYYY are -0.05,-0.08,-0.10 which are relatively small, so we did not consider them as factors affecting the use of opioids.
- **Strategy 2.** The living conditions and status of citizens of different ethnic groups in the United States are not the same, so the proportion of drug abuse is also different. This is one of the important influencing factors identified by our model. The proportion of Sub Saharan Africa and Arab are more relevant to the abuse of opioids, which is a long-standing historical cause, and we should also work to address these issues. First, we should provide better health care for Sub Saharan Africa and Arab, which can reduce the use of offending prescriptions. Second, there are often leaders with higher levels of education within the community. We can invite them to promote the harm of opioids, which is often more acceptable.
- **Strategy 3.** In our model, the immigration is the most relevant factor. We should strengthen the review and knowledge dissemination of immigrants, including checking social accounts and work experience. At the same time, we should figure out whether immigrants have a history of family drug abuse and criminal history. Reject immigrants with greater potential for drug use and risk of drug trafficking. For immigrants and their families who already live in the state, education is very necessary, we can provide them with handbooks about the risks of opioids.

- **Strategy 4.** People who do not live with their families are more likely to become "addicts." In our model, the factor of nonfamily households was positively correlated with the abuse of opioids. And we are pleasantly surprised to see in the RFRT model that families with older people over the age of 65 are significantly less likely to have opioid addicts, which can be the solution of the nonfamily households. Therefore, we should encourage and promote people to take care of the elderly, at the very least, to increase the frequency of communication.
- **Strategy 5.** Unmarried people are more likely to take drugs. We have confirmed in the model that the "prejudice" in their lives is not groundless. In fact, lack of family support and involvement will increase the risk of drug use. The work of reducing this factor seems to have a long way to go. In fact, we can influence it with some simple work. Holding festive parties in an area where singles gather is a good choice. More activities and communication can not only ease the pressure of life, but also a good opportunity for people to end their celibacy.

6.2 Identify Possible Strategies

Then we gave some biases on each index, according to the strategies above. We put the biased indexes according to the parameter of the model built in part I and part II. We calculated the predict results increase or decrease relative to training outcomes. The result is showed in table 5.

Table 5: Strategy test

Parameter	Bais	Total drug of counties	Margin of Error
HW65	+50%	19220	-23.18%
HW65	-50%	32554	30.12%
NM	-50%	16112	-35.60%
NM	+20%	40983	63.81%
NOH	+50%	33173	32.60%
NOH	-50%	23808	-4.84%
LOE	+50%	26526	6.03%
LOE	-50%	24542	-1.90%
FB	+50%	25727	2.83%
FB	-50%	24066	-3.81%
SA	+50%	27238	8.87%
SA	-50%	20945	-16.28%

The training value of the total drug of KY is 25018, the real value is 24379. The gap between them is 2.6%.

From table5, we saw that the change in the total amount of drugs in Kentucky is in line with our estimates. Changing the parameters will reduce the amount of drugs to an acceptable range. Therefore, our policies and models match each other, which is feasible and practical.

7 Model evaluation and analysis

7.1 Strengths and Weakness

For strengths, our model combined the Second Exponential Smoothing Model and random forest. The accuracy of Random Forest is deeply associated with the quantity of sample. If the quantity of sample is small we can use Exponential Smoothing Model while if the quantity is large we can use the Random Forest model. In this way we avoided the errors caused by chance.

For weakness, Random forests have been proven to be over-specified on certain noisy classifications or regression problems.[6]. For data with different values, attributes with too many indexes will have a greater impact on random forests, so the attribute weights generated by random forests on such data are not credible.

7.2 Future Work

In the future,we intend to optimize our model in two ways.

- Collect lower-status statistics such as age, gender, and family status of drug users. With this statistics, we can do some more targeted work and give a more specific advice to the government.
- Analysis of factors related to economic development and opioid crisis. For example, During our modeling in this problem, we came to know a concept of Rust Belt[1], with high unemployment rate of the manufacturing industry and heavy industry composed places in the Midwest and Great Lakes.

References

[1] En.wikipedia.org. (2019). Rust Belt. [online] Available at: https://en.wikipedia.org/wiki/Rust Belt

[2] CDC Reports Highest Drug Overdose Death Rates in Record. (2018). PR Newswire, p. PR Newswire, Aug 21, 2018.

[3] Mohamadi, A. et al., 2018. Risk Factors and Pooled Rate of Prolonged Opioid Use Following Trauma or Surgery: A Systematic Review and Meta-(Regression) Analysis. The Journal of bone and joint surgery. American volume, 100(15), pp.13321340.

[4] Centers for Disease Control and Prevention. (2019). Confronting Opioids. [online] Available at: https://www.cdc.gov/features/confronting-opioids/index.html

[5] En.wikipedia.org.(2019).Opioid.[online]Available.at: https://en.wikipedia.org/wiki/Opioid.cite-note-unodc-223.

[6] Salles et al., 2018. Improving random forests by neighborhood projection for effective text classification. Information Systems, 77, pp.121.

[7] Walker, G. (2018). The opioid crisis: A 21st century pain. Drugs of Today (Barcelona, Spain : 1998), 54(4), 283-286.

[8] Florence, C. S., Zhou, C., Luo, F., & Xu, L. (2016). The Economic Burden of Prescription Opioid Overdose, Abuse, and Dependence in the United States, 2013. Medical Care, 54(10), 901-906.

[9] "Opioid danger." Chain Drug Review 27 Apr. 2015: 227. Business Insights: Global. Web. 25 Jan. 2019.

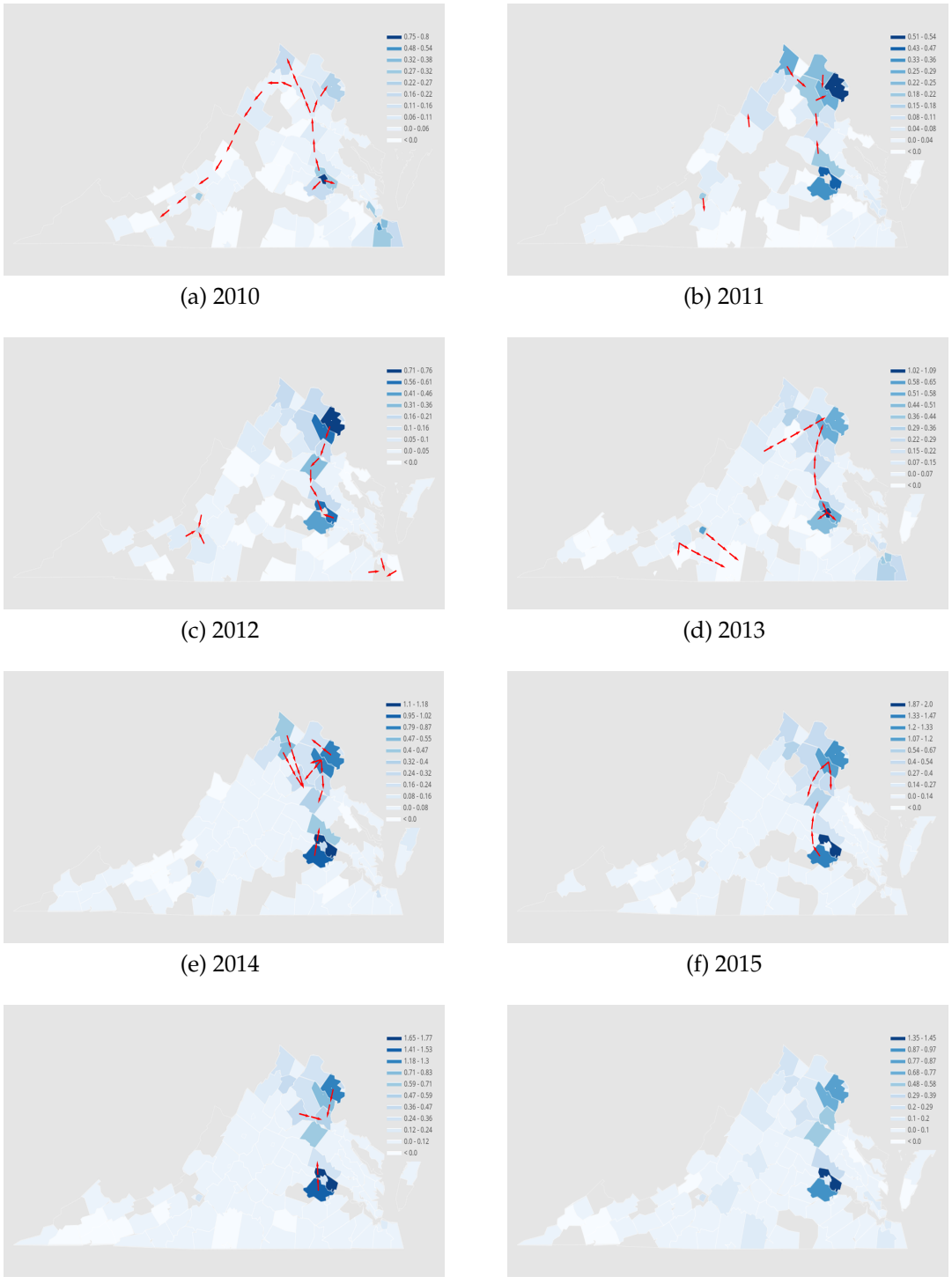
[10] Anon, (2019). [online] Available at: https://www.usnews.com/opinion/blogs/policy-dose/articles/2016-02-01/10-ways-to-combat-americas-drug-abuse-problem [Accessed 28 Jan. 2019].

[11] BBC News. (2019). Five ways to tackle the US drug epidemic. [online] Available at: https://www.bbc.com/news/world-us-canada-40479686 [Accessed 28 Jan. 2019].

A Appendix

A.1 Appendix A

The spread of heroin of each counties in each state.



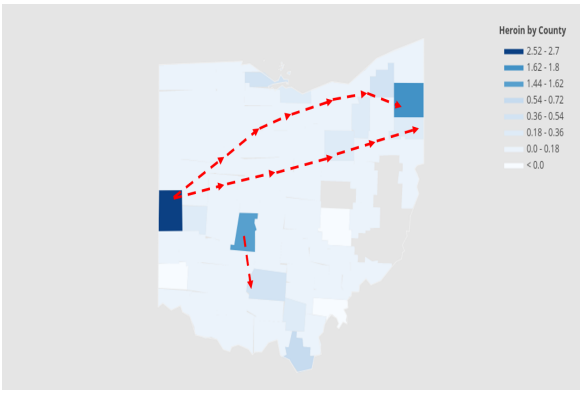
(a) 2010 (b) 2011

(c) 2012 (d) 2013

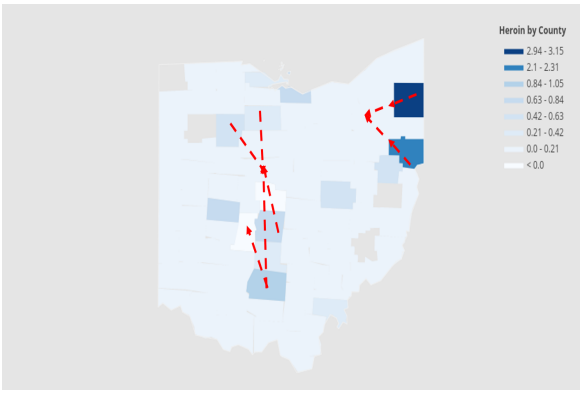
(e) 2014 (f) 2015

(g) 2016 (h) 2017

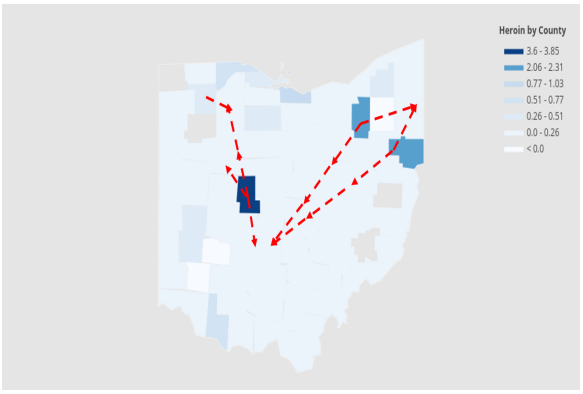
Figure 9: The Spread of Heroin in the VA State Counties



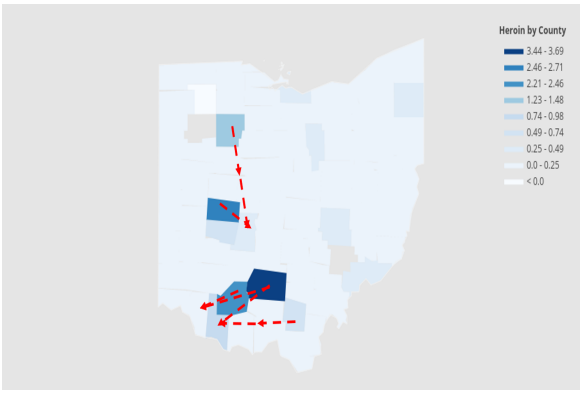
(a) 2010



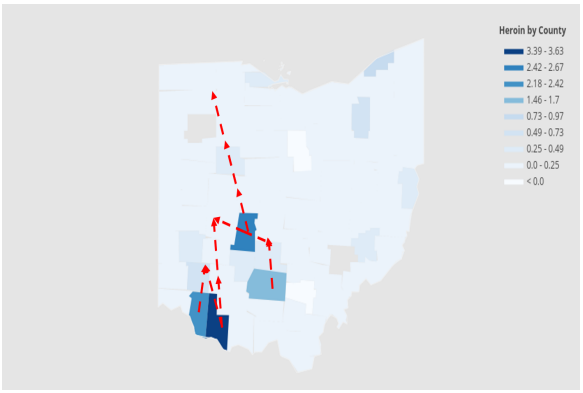
(b) 2011



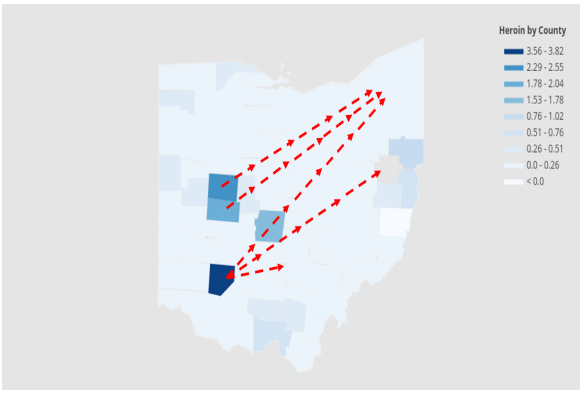
(c) 2012



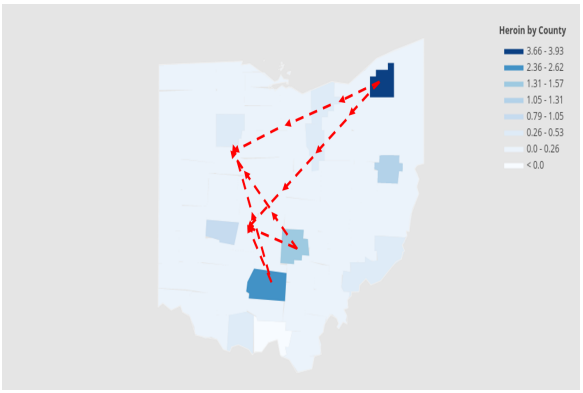
(d) 2013



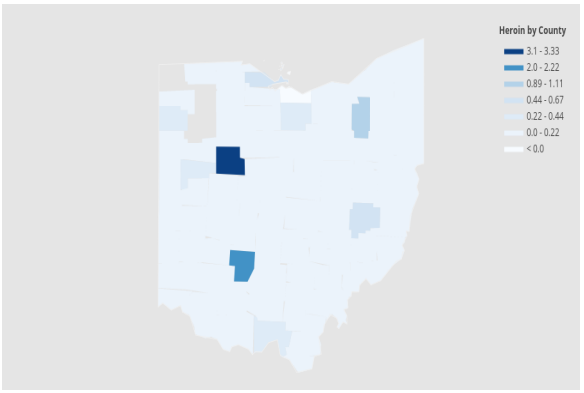
(e) 2014



(f) 2015

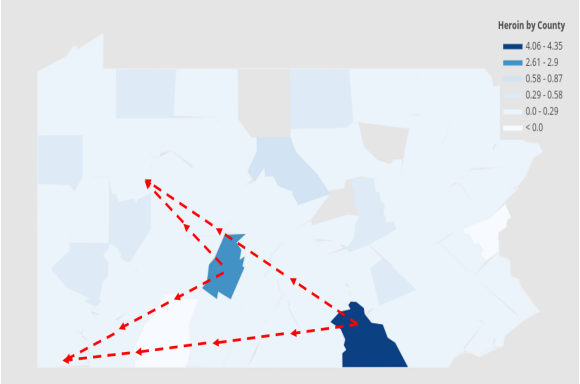


(g) 2016

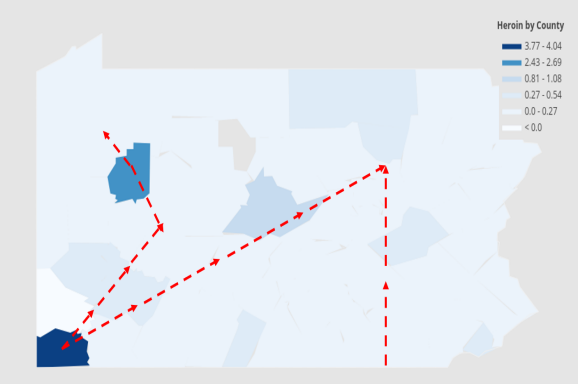


(h) 2017

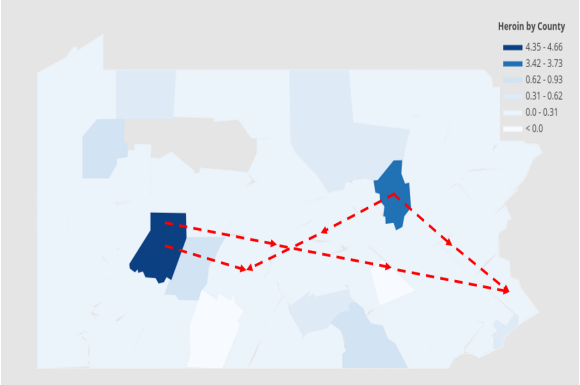
Figure 10: The Spread of Heroin in the OH Counties



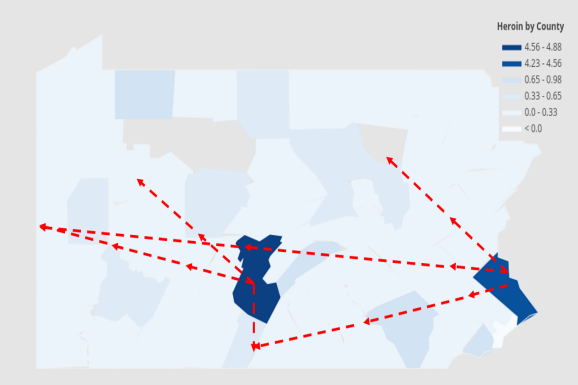
(a) 2010



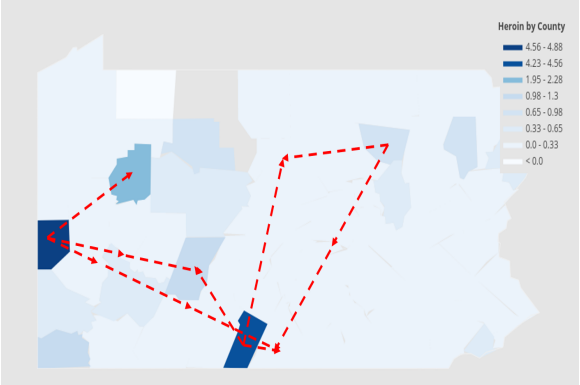
(b) 2011



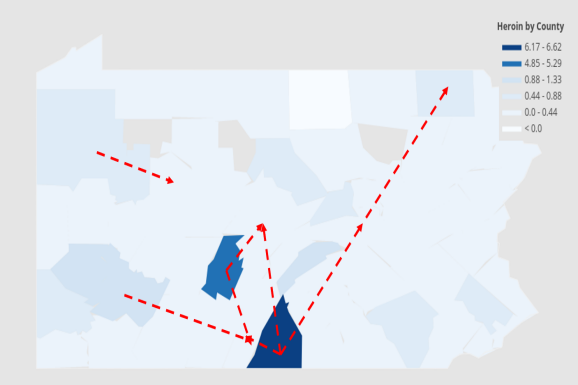
(c) 2012



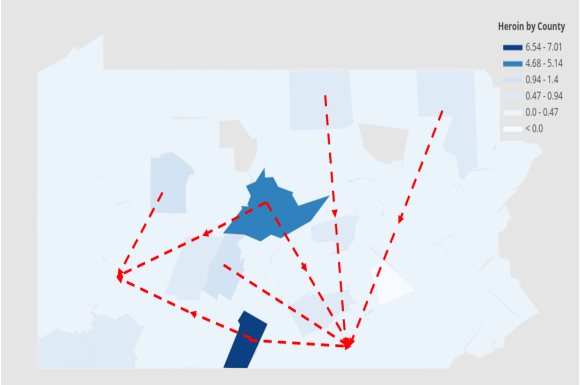
(d) 2013



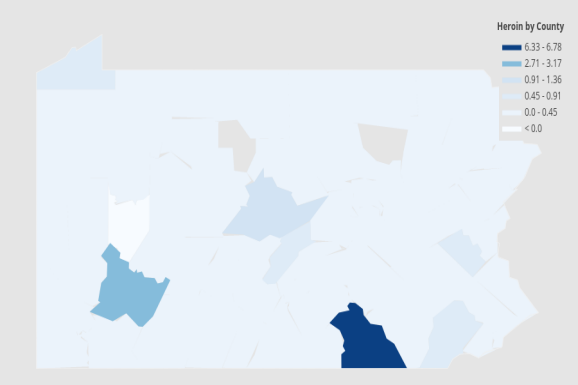
(e) 2014



(f) 2015

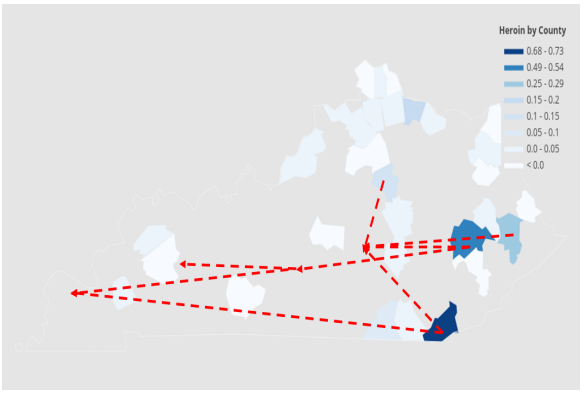


(g) 2016

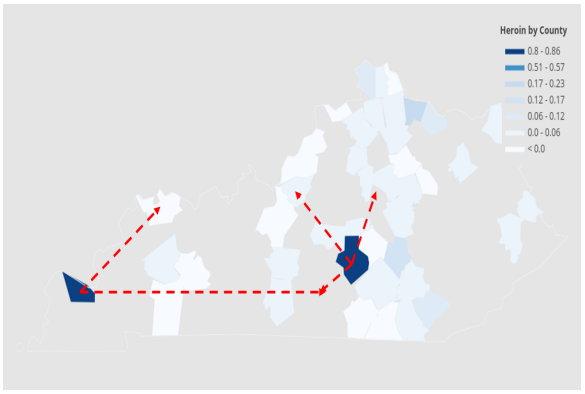


(h) 2017

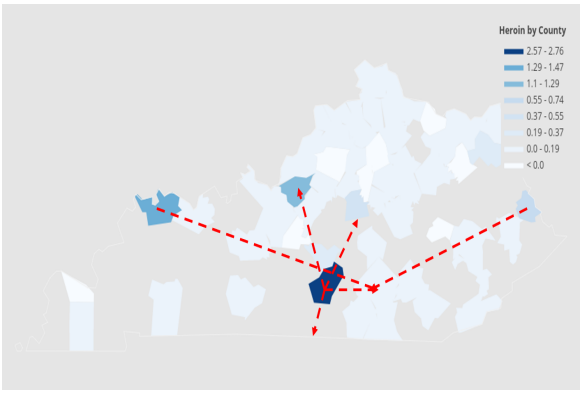
Figure 11: The Spread of Heroin in the PA Counties



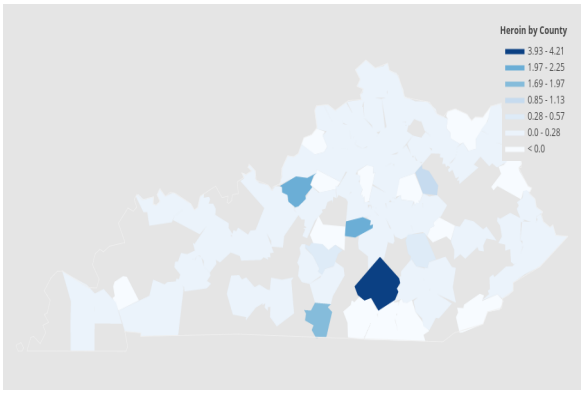
(a) 2010



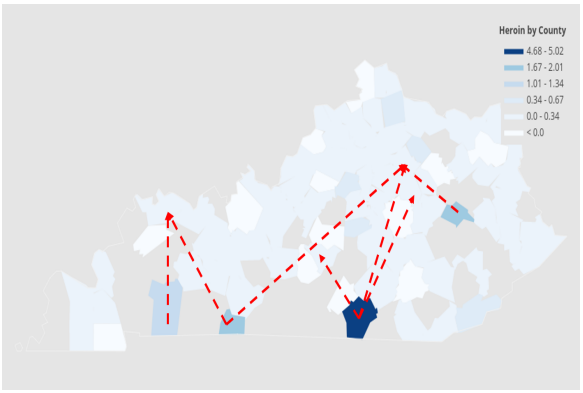
(b) 2011



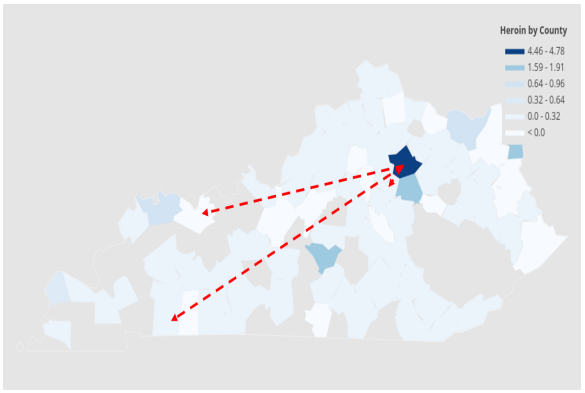
(c) 2012



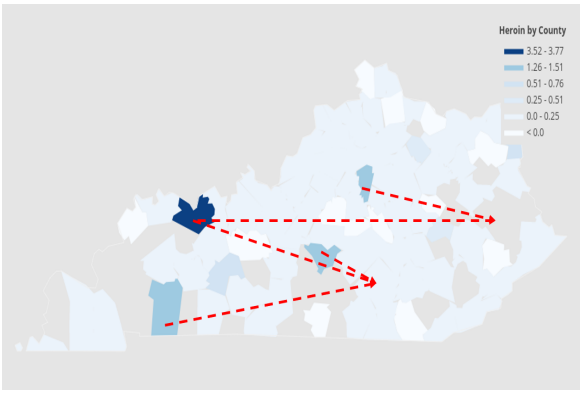
(d) 2013



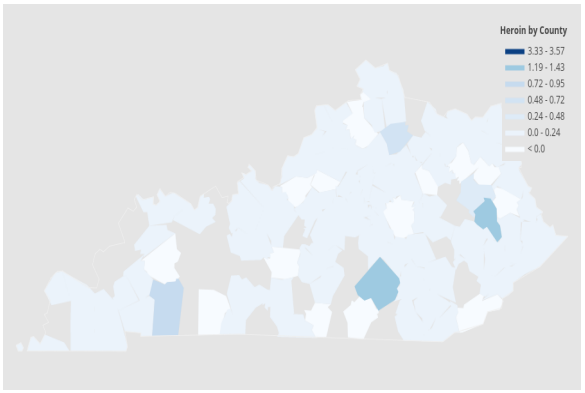
(e) 2014



(f) 2015

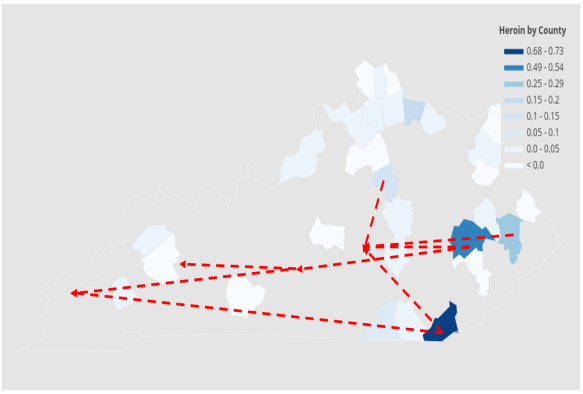


(g) 2016

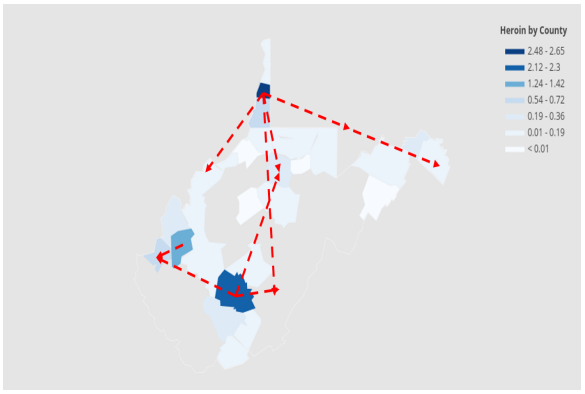


(h) 2017

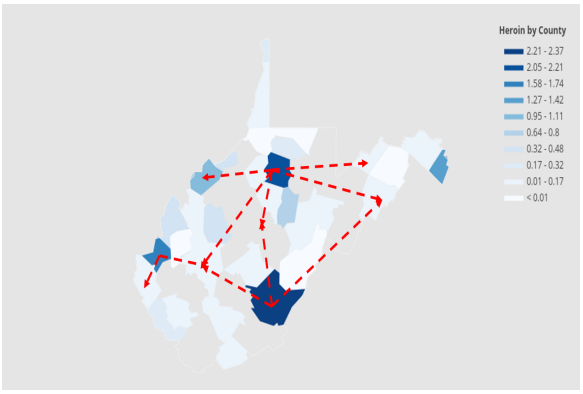
Figure 12: The Spread of Heroin in the KY Counties



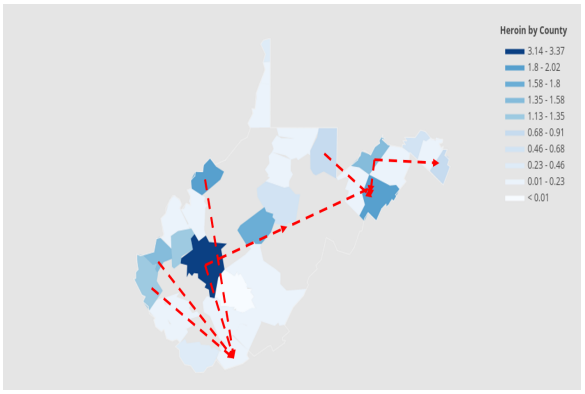
(a) 2010



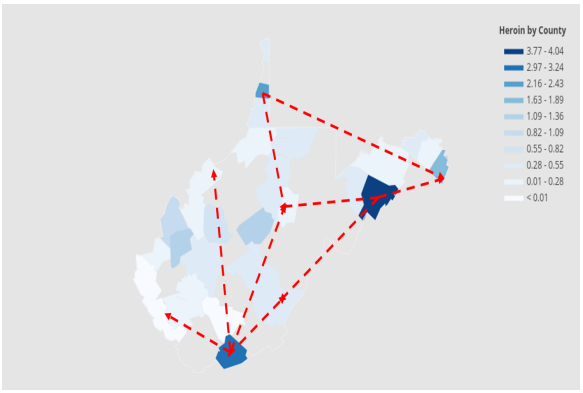
(b) 2011



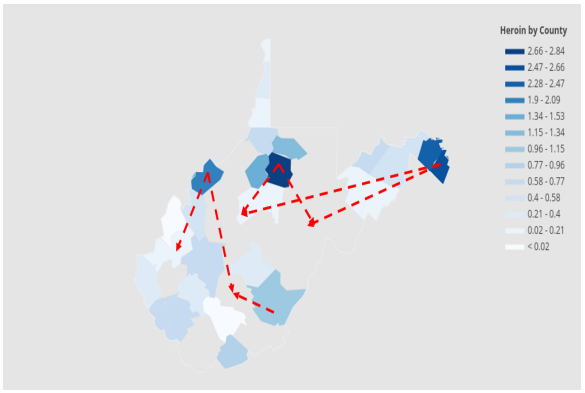
(c) 2012



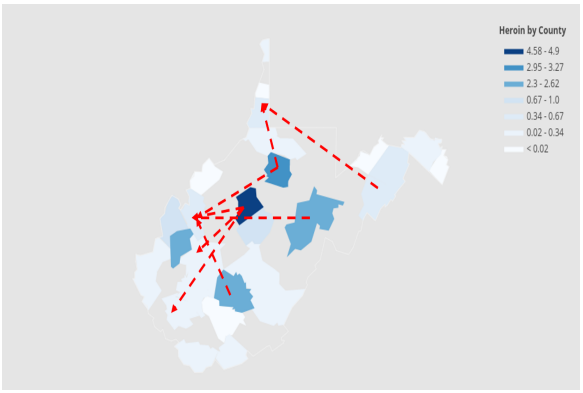
(d) 2013



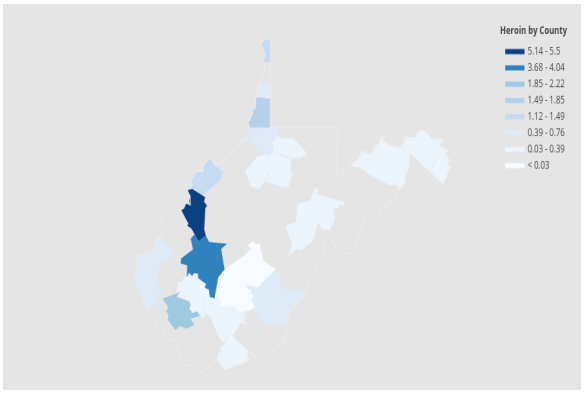
(e) 2014



(f) 2015



(g) 2016



(h) 2017

Figure 13: The Spread of Heroin in the WV Counties

Table 6: The HDIC of each counties in KY

Pre.proportion(%)	2010	2011	2012	2013	2014	2015	2016	2017	**2018**	**2019**
WOODFORD	0.035	0.004	0.022	0.030	0.037	0.089	0.041	0.021	*0.063*	*0.058*
WHITLEY	0.010	0.004	0.007	0.011	0.011	0.004	0.008	0.003	*0.007*	*0.004*
WAYNE	0.015	0.007	0.007	0.011	0.015	0.015	0.019	0.028	*0.030*	*0.033*
WARREN	0.014	0.019	0.004	0.037	0.022	0.012	0.041	0.003	*0.036*	*0.012*
TRIMBLE	0.013	0.013	0.004	0.030	0.007	0.023	0.008	0.007	*0.003*	*0.015*
TAYLOR	0.063	0.063	0.022	0.060	0.078	0.043	0.124	0.052	*0.119*	*0.074*
SPENCER	0.018	0.018	0.007	0.011	0.033	0.015	0.038	0.003	*0.026*	*0.021*
SHELBY	0.123	0.123	0.011	0.086	0.185	0.221	0.124	0.111	*0.157*	*0.198*
SCOTT	0.014	0.113	0.080	0.153	0.174	0.035	0.215	0.125	*0.229*	*0.175*
ROWAN	0.087	0.004	0.047	0.082	0.100	0.139	0.188	0.045	*0.175*	*0.159*
ROCKCASTLE	0.003	0.004	0.028	0.028	0.085	0.027	0.034	0.017	*0.036*	*0.008*
PULASKI	0.086	0.011	0.069	0.022	0.033	0.240	0.151	0.073	*0.195*	*0.198*
PENDLETON	0.003	0.067	0.131	0.280	0.174	0.112	0.034	0.045	*0.000*	*0.000*
OWEN	0.003	0.007	0.007	0.034	0.048	0.035	0.045	0.035	*0.052*	*0.054*
OLDHAM	0.010	0.032	0.058	0.063	0.055	0.085	0.034	0.048	*0.044*	*0.053*
NELSON	0.086	0.086	0.040	0.037	0.078	0.077	0.109	0.173	*0.186*	*0.220*
MONTGOMERY	0.119	0.042	0.007	0.145	0.188	0.112	0.185	0.152	*0.193*	*0.226*
MERCER	0.061	0.061	0.011	0.030	0.148	0.027	0.094	0.059	*0.098*	*0.082*
MEADE	0.024	0.024	0.015	0.060	0.015	0.031	0.004	0.021	*0.001*	*0.021*
MASON	0.061	0.039	0.135	0.227	0.066	0.077	0.106	0.035	*0.047*	*0.048*
MARTIN	0.014	0.007	0.036	0.015	0.019	0.004	0.019	0.038	*0.034*	*0.022*
MARION	0.003	0.004	0.027	0.011	0.055	0.023	0.053	0.042	*0.068*	*0.050*
MARION	0.007	0.011	0.033	0.433	0.506	0.682	0.712	0.585	*0.861*	*1.021*
MCCRACKEN	0.039	0.018	0.069	0.060	0.041	0.004	0.041	0.042	*0.041*	*0.020*
LINCOLN	0.016	0.007	0.015	0.007	0.018	0.027	0.023	0.017	*0.031*	*0.027*
LINCOLN	0.017	0.017	0.011	0.011	0.007	0.012	0.030	0.031	*0.035*	*0.038*
LEWIS	0.020	0.007	0.018	0.004	0.015	0.027	0.019	0.048	*0.029*	*0.053*
LAUREL	0.730	0.856	1.298	2.058	1.957	1.612	1.372	1.209	*1.132*	*1.038*
KENTON	0.222	0.222	0.040	0.268	0.366	0.248	0.219	0.194	*0.211*	*0.271*
JESSAMINE	0.520	0.841	2.756	4.213	5.015	4.777	3.769	3.571	*3.874*	*4.573*
JEFFERSON	0.043	0.043	0.007	0.034	0.059	0.089	0.049	0.017	*0.047*	*0.059*
HENRY	0.031	0.004	0.011	0.019	0.007	0.058	0.098	0.017	*0.075*	*0.080*
HARRISON	0.163	0.004	0.029	0.119	0.262	0.372	0.290	0.062	*0.337*	*0.307*
HARLAN	0.006	0.006	0.004	0.004	0.011	0.004	0.004	0.010	*0.008*	*0.008*
HARDIN	0.082	0.007	0.015	0.097	0.122	0.112	0.121	0.104	*0.151*	*0.162*
HARDIN	0.007	0.004	0.036	0.063	0.026	0.074	0.079	0.014	*0.059*	*0.063*
GRAYSON	0.013	0.013	0.007	0.004	0.004	0.039	0.015	0.010	*0.019*	*0.026*
GRANT	0.010	0.014	0.065	0.086	0.207	0.194	0.170	0.090	*0.139*	*0.100*
GALLATIN	0.039	0.039	0.022	0.086	0.089	0.019	0.011	0.007	*0.002*	*0.000*
FRANKLIN	0.051	0.011	0.011	0.216	0.351	0.240	0.366	0.187	*0.398*	*0.391*
FLOYD	0.003	0.011	0.029	0.015	0.015	0.008	0.012	0.003	*0.008*	*0.000*
FAYETTE	0.108	0.081	0.502	1.738	1.186	1.678	1.323	1.299	*1.650*	*1.976*
DAVIESS	0.031	0.031	0.011	0.030	0.055	0.043	0.038	0.007	*0.031*	*0.027*
CASEY	0.118	0.021	0.113	0.186	0.155	0.136	0.113	0.104	*0.154*	*0.145*
CHRISTIAN	0.010	0.007	0.004	0.007	0.026	0.008	0.008	0.010	*0.014*	*0.011*
CARTER	0.003	0.092	0.036	0.089	0.188	0.050	0.166	0.111	*0.186*	*0.168*
CARROLL	0.027	0.081	0.225	0.276	0.199	0.101	0.064	0.048	*0.019*	*0.000*
CAMPBELL	0.291	0.552	1.156	1.969	1.847	1.798	1.444	0.856	*0.834*	*0.698*
BULLITT	0.157	0.014	0.124	0.324	0.177	0.201	0.136	0.125	*0.172*	*0.187*
BRACKEN	0.003	0.011	0.076	0.037	0.048	0.046	0.057	0.040	*0.066*	*0.036*
BOYLE	0.129	0.004	0.129	0.097	0.166	0.128	0.219	0.163	*0.274*	*0.221*
BOYD	0.017	0.141	0.218	0.280	0.351	0.198	0.204	0.180	*0.220*	*0.155*
BOURBON	0.022	0.004	0.018	0.026	0.026	0.015	0.034	0.028	*0.037*	*0.035*
BOONE	0.159	0.194	0.582	0.854	0.620	0.879	0.618	0.395	*0.454*	*0.422*
ANDERSON	0.058	0.058	0.055	0.067	0.074	0.046	0.060	0.045	*0.054*	*0.046*