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

Team Control Number
1923122

Problem Chosen
C

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

2019
MCM/ICM
Summary Sheet

As the opioid crisis gets increasingly severe in the U.S. in the 21st Century, it has become necessary and urgent to find out the cause, characteristics, and driving force of the crisis. In this paper, our team has built several models to interpret the opioid crisis in many aspects:

Enlightened by the SIR model, we created the Drug User Estimation Method which reveals a concerning trend for the US Government describing the relationship between heroin and other opioids.

In order to better help the government track back where the opioid drugs are originated from, we created a static model, County Ranking Model, and another dynamic model, Inverse Markov Chain Model, to track back the source of every specific drug. Though the two models view the spreading of drugs in different aspects, they come up with surprisingly similar results, which help us discover and verify the accuracy of identifying "drug origins" per state.

Combining the previous data with that of the U.S. Census Bureau, we implemented machine learning methods to find out the driving force of the crisis:
First, we created a Support Vector Machine for the U.S. Government. As far as some relevant data of a specific county is input, the machine will come up with the prediction of whether the county has high frequency of drug usage or not. After numerous testing, the prediction of SVM is highly accurate. Therefore, it might be a good auxiliary tool for policy making of government.

Then, using the method of Decision Tree and Random Forest, we found the characteristics of people that are highly related with this opioid crisis. Because of the robustness of the model, if more related data could be collected, the algorithm would have perform even better.

Based on all the models above, we picked three strategies for U.S. Government to counter this opioid crisis. Furthermore, we tested that if those strategies were implemented, the opioid crisis will be significantly relieved.
Finally, we composed a memo directed to the Chief Administrator of DEA/NFLIS about different possibilities to alleviate the prevalence of opioids in the states. We hope it may help with solving this opioid crisis.

Opioid Crisis on the Rise: Analysis of US Drug Trends Using SIR Models, Stochastic Process, and Machine Learning

By Team #1923122

Content

1. Introduction

In the past few decades, a new epidemic has grown to become a significant threat to the US people: The Opioid Crisis. Starting from the early 1990s, the rapid spread and abuse of opioids came about largely through their over-prescription by doctors as painkillers. Because many of these substances have side effects that lead to high chances of addiction, this has led to their popularity for abuse and illegal trading. Furthermore, addiction to these painkillers have shown to be gateways to more intense drugs such as heroin and fentanyl. (National Institute on Drug Abuse, 2019; U.S. Department of Health and Human Service, 2019; Clair F., 2019)

The similarities between opioids and heroin may cause this Opioid Crisis to become more severe. They are both fast-acting and create a short but intense rush, and they are also both highly addictive. (American Addiction Center, 2019) Therefore, heroin can serve as a substitution of medical opioid. Because heroin is more addictive than medical opioid, the abuse of opioid will only lead to a more severe abuse of heroin. The data shows that 1.6 percent of Americans used heroin at some point, while about 950,000 used last year. (U.S. Department of Health and Human Service, 2019)

In order to find out the characteristics and sources of this opioid crisis and reason for the abuse of heroin, we focus on the data provided by DEA/National Forensic Laboratory Information System (NFLIS) and the U.S. Census Bureau centered around five states (Virginia, Ohio, Kentucky, Pennsylvania, and West Virginia). The model based on the data provides an interpretation and overview of this Opioid Crisis as well as analyzed the feature of residents where opioid and heroin are abused. From the model, specific suggestions are derived for the U.S. government to deal with this opioid crisis.

2. Assumptions & Justification

Assumption 1: Assume that the area created by the 5 states are flat plane.
Justification 1: Since the area constructed by the 5 states is comparatively much smaller than the surface area of the earth, we can regard it as flat.
Assumption 2: Assume that people are equally likely to receive an opioid prescription and take Heroin as well.
Justification 2: We do not have the age distribution of opioid or heroin takers, so it is hard to estimate the age of reporters. In addition, the large number theorem can counteract bias.
Assumption 3: Drug usage will randomly flow from county to county.
Justification 3: Though the strength of drug flow may vary due to population, distance, and other factors, there is no evidence that the drug will flow by a certain direction.

3. Variables Declaration

Global Variables	
Variable Name	Definition
t	Time
Dis_{ij}	Distance between county/state i, j

$p_{d,c,t}$	Number of drug reports about drug d in county c during the t th year after such drug coming out
Variables of the Drug User Estimation Model	
$O_{i,t}$	Number of opioid reports for county/state i at time t
$H_{i,t}$	Number of heroin reports for county/state i at time t
$R_{i,t}$	Number of people that are removed from heroin using group for county/state i at time t
$F(t), G(t)$	Fourier Functions
α_{ij}	Rate of heroin-addicted people moving from i to j
β_i	Rate of people who used opioid changing to use heroin
γ_i	Rate of people that are removed from heroin using group
$lat_i, long_i$	Latitude and longitude of county/state i
k, q, m	Parameters
Variables of the County Ranking Model	
δ	Degradation Factor
$Score_{d,c}$	Score of ranking of drug d in county c
$Space_{d,c}$	Score of spacing of drug d in county c
$Time_{d,c}$	Score of timing of drug d in county c
$Prob_{d,c}$	Probability of county c becoming the source of drug d
Variables of the Inverse Markov Chain Model	
$v_{c,l}$	The value of the node (county) c at iteration l
T_l	The transition matrix of Markov Chain at iteration l
$tr_{ij,l}$	Probability of a unit of drug flux flowing from county i to j in iteration l
$in_{i,l}, out_{i,l}$	The number of units of drug flux moving in/out to/from county i in iteration l

4. Drug User Estimation Model Based on SIR Model

4.1 Fourier Function within the States

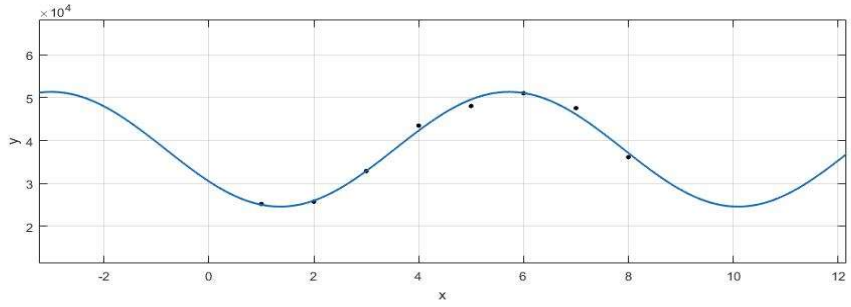
If we add up the number of people who use heroin and other opioids in the five states, we can get the following table:

Year	NO. of people using Opioid	NO. of people using Heroin	% of people using Heroin	Total No. of Drug Users
2010	215,466	25,232	10.48%	240,698
2011	198,853	25,786	11.48%	224,639
2012	200,251	32,923	14.12%	233,174
2013	205,906	43,513	17.45%	249,419
2014	196,955	48,054	19.61%	245,009
2015	192,402	51,074	20.98%	243,476
2016	205,545	47,581	18.80%	253,126
2017	221,484	36,152	14.03%	257,636

Table 1: Overview of People Using Drugs in 5 States

From the data above, it is shown that the number of people using heroin and the percentage of people using heroin in the 5 states over 2010 to 2017 follows a Fourier Function as following:

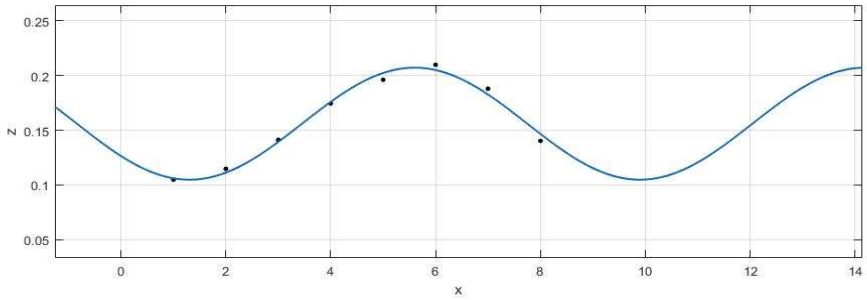
(For convenience, we denote 2010 as $x = 1$, so as following.)



Graph 1: The Curve of Number of Heroin Users vs. Time

$$Fit\ of\ Heroin:\ f(t) = 38000 - 7503\cos(0.7192t) - 11070\sin(0.7182t)$$

$$R^2 = 0.9906\ \ R^2_{adj} = 0.9835$$



Graph 2: The Curve of Percentage of Heroin Users vs. Time

$$Fit\ of\ Percentage:\ g(t) = 0.156 - 0.0296\cos(0.7314t) - 0.0418\sin(0.7314t)$$

$$R^2 = 0.986\ \ R^2_{adj} = 0.9754$$

The high R squared and adjusted R squared values indicate that the curves fit the data well and do not overfit. In order to check the generality, we fit the number and percent of people using heroin in the 5 states individually. We get the following result:

Number of People Using Heroin in States						
$f(t) = a_0 + a_1\cos(wt) + a_2\sin(wt)$						
State	a_0	a_1	a_2	w	R^2	$R^2\ adj.$
VA	2976	-692.4	-948	0.6168	0.6432	0.3775
OH	16400	-4706	-4537	0.6858	0.9757	0.9575
KY	2310	-2149	-52.4	0.5557	0.9434	0.9009
PA	15110	250.9	-3858	0.8332	0.9656	0.9398
WV	1253	-152.9	-374.9	0.9638	0.8037	0.6564
Percentage of People Using Heroin in States						
$g(t) = a_0 + a_1\cos(wt) + a_2\sin(wt)$						
State	a_0	a_1	a_2	w	R^2	$R^2\ adj.$
VA	0.0863	-0.0089	-0.041	0.6975	0.9627	0.9347

OH	0.1698	-0.0354	-0.0226	0.743	0.8957	0.8174
KY	0.0895	-0.0779	-0.0165	0.5935	0.9592	0.9285
PA	0.1922	-0.0306	-0.0527	0.7091	0.9945	0.9903
WV	0.1652	-0.0562	-0.0344	0.5944	0.9267	0.8717

Table 2: The Fitting Result of the Fourier Function

From the table, it is shown that for most of the fitting results, the R square is greater than 0.85, which means the Fourier Function fits the data well.

Based on the result of fitting, we can draw the conclusion that the number and percentage of people using heroin in the 5 states follows Fourier Function.

Since the total drug takers can be presented as:

$$Total\ Drug\ Takers = \frac{Heroin\ Takers}{\%\ of\ Heroin\ Takers}$$

Therefore, it can be modelled by the division of Fourier functions, and the number of other opioid drug takers can also be deduced. In conclusion, based on the overview of data in states. We can conclude that for state i:

$$H_i(t) = F_i(t)$$

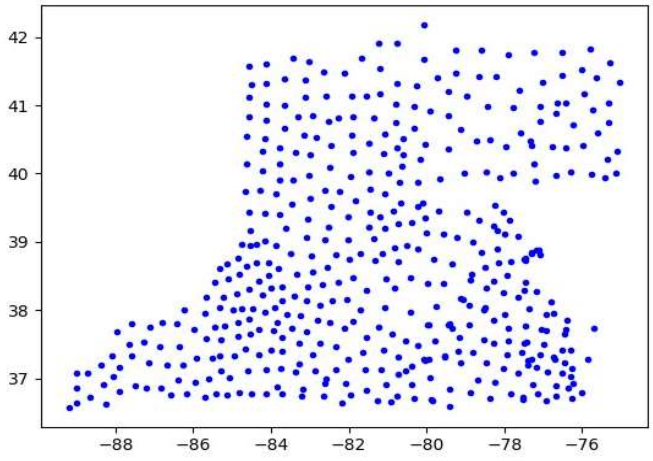
$$O_i(t) = \frac{F_i(t) - F_i(t)G_i(t)}{G_i(t)}$$

$$F_i(t), G_i(t)\ are\ Fourier\ functions$$

4.2 Coordinate System of the Five States

In order to further explore the characteristic of spreading between states and counties, we need a dynamic model that can simulate the flow of drug from counties to counties. It requires the geographical position of counties. Therefore, we obtained the latitudes and longitudes of counties and create a coordinate system of the five states.

We obtained the latitudes and longitudes of counties from Google Maps and plotted the counties in a coordinate system. The south-most county is Fulton County, KY with latitude of 36. 57518 °N; the north-most county is Erie County, PA with latitude of 42.18275 °N. The west-most county is Fulton County, KY with 89. 2031 °W; the east-most county is Pike County, PA with 75.0241 °W. The plot is shown as following:



Graph 3: The Plot of Counties in the Coordinate System

When calculating the distance between counties, we will only use two counties'

latitudes and longitudes. However, we realized that since the earth is a sphere, 1-degree latitude is not equal to 1-degree longitude. Therefore, based on our assumption that we regard the 5 states as a flat plane and the transition between latitude and longitude. We define the distance in the Coordinate System of Five States:

$$Dis_{ij} = \sqrt{\left(lat_i - lat_j\right)^2 + \left(long_i - long_j\right)^2 \cos^2\left(\frac{lat_i + lat_j}{2}\right)}$$

4.3 Modified SIR Model for the Drug User Estimation

The SIR model is a math model that can simulate the spreading of disease. The core idea is to divide the whole population into a Susceptible Group (S Group), an Infected Group (I Group), and a Removed (or Recovered Group) (R Group). In each time period, some people will get infected and transfer from the S Group to I Group, and another group of people will get recovered (or pass away) and transfer from I Group to R Group. The model can simulate the process of a kind of disease from spreading to being wiped out.

Drug Usage's similarity to the SIR model

According to the background research, the process of people using opioid and Heroin can be simulated as an SIR model theoretically. When people receive a prescription of some opioid drugs, there is a probability that they get addicted to opioid drug and try to find Heroin as substitution when they are running out of opioid drug. That is, they are transferred from Opioid Group (O Group or S Group) to Heroin Group (H Group or I Group). Also, after a period of time, people who are taking Heroin may either get rid of it or die because of it. Therefore, they are transferred from Heroin Group to Removed Group (R Group).

Drug Usage's difference from the SIR model

There are mainly three differences: 1. There are new patients taking opioids, so we can't ignore the number of people joining O Group, instead of ignoring the birth rate in SIR model; 2. In the state and county aspect, there are people who are moving from state to state (or county to county); 3. According to our assumption that the drug usage flow is totally random, there should be white noise over time. (That is, to simulate real situation, the model shouldn't be deterministic).

4.4 Drug User Estimation Model

According to the analysis in 4.3, we define the following differential equations:

$$\frac{dO_i}{dt} = \left[\left(1 - \frac{1}{G_i(t)}\right)F_i(t)\right]' - \frac{\beta O_i^2}{O_i + H} + N_1$$

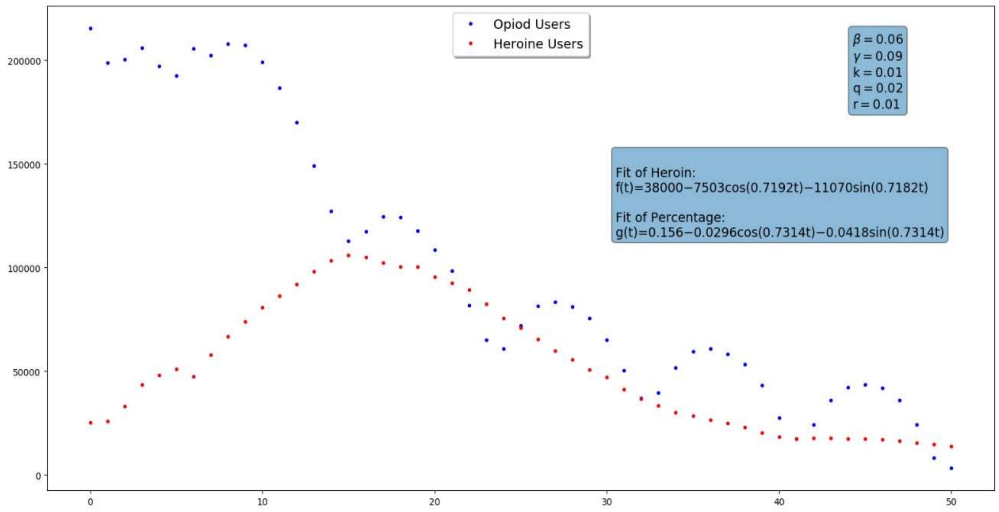
$$\frac{dH_i}{dt} = \frac{\beta O_i^2}{O_i + H_i} + \sum_j \alpha_{ji}\left(\gamma H_i\right) - \gamma H_i + N_2$$

$$\frac{dR_i}{dt} = \gamma H_i - \sum_j \alpha_{ij}\left(\gamma H_i\right) + N_3$$

$$R_{i,0} = 0 \quad N_1 \sim Normal\left(0, k * O_i(t)\right) \quad N_2, N_3 \sim Normal\left(0, q * H_i(t)\right)$$

From 4.1, we obtained that the function of users of opioid drugs is roughly composed by Fourier Functions. Therefore, $F_i(t)\&G_i(t)$ are Fourier Functions.

The model can be verified by the total population of people using opioid drugs and Heroin. Since we are considering the total population of 5 states, we don't need to consider the move between states/counties. That is, for all i,j $\alpha_{ij} = 0$. When we set the initial value of O_0, H_0 to the number of total opioid drug reports and Heroin cases in 2010, respectively. The future trends of opioid drug reports and Heroin cases are shown as following:



Graph 4: The Estimation of Total Opioid and Heroin Usage

Therefore, from the result, it is shown that following this trend, there will be a serious concern for the U.S. government. In around 2024, the cases of people using Heroin will become over 100,000 (currently it's 30,000 to 50,000). This will continue for around 5~6 years before it come back to less than 100,000, and it will take around 15~20 more years to gradually come back to the current level.

4.4 Discussion about the Flow between States/Counties

In reality, instead of each state/county has its own Drug User Estimation Model, people actually move from state/county to state/county. Therefore, it is important to add the α_{ij} term in the model, which means the rate of people moving from state/county i to j. Since the further two state/county are from each other, fewer people tend to move between them, so the rate is negatively correlated with the distance between state/county i and j.

We set the function α_{ij} as

$$\begin{cases} \alpha_{ij} = 0 & if\ i = j \\ \alpha_{ij} = \dfrac{1}{mDis_{ij}} & if\ i \neq j \end{cases}$$

The distance between two states/counties is calculated using the coordinate system of the five states (section 4.2). Considering the flow between states/counties, we estimate the number of heroin usage cases in five states in future, and the result is shown

in Section 7.

5. County Ranking Model

5.1 Model Construction

The Drug User Estimation Model provides with the method of estimating number of Heroin and non-Heroin drug users in future years and reveals the characteristics of spreading. However, we need another model to find out the possible locations where specific opioid use might have started.

Considering the possible features of the drug, we assume that drug sources tend to have two features:

- As the source of a specific drug, it is more likely to become a "center" of that specific drug. That is, the counties around them tends to have more users of that drug;
- As the source of a specific drug, it is more likely to have more users, and the users in that county would increase over years.

Since the target of the model is to find out the possible location of drug source, we only need to focus on the counties using a specific drug d when this drug is initially (the first year) being used (**called candidate counties**). The rest of the counties will not be the candidates of the drug source. If there are more than one county use the drug in the first year, it's hard to determine which one is the drug source, but by analyzing the two features above, the County Ranking Model is supposed to give out a list of probabilities of the certain county becomes the source of the drug.

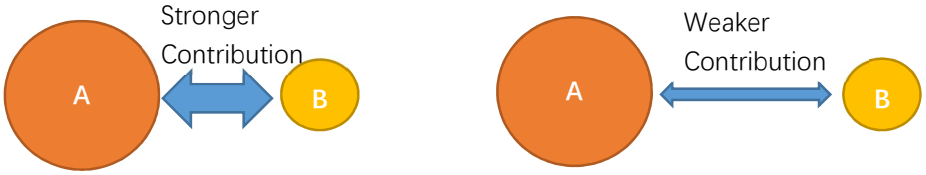
5.2 Score of Ranking and County Ranking Model

In order to calculate the probability of a county becomes the source of the drug, we define a score of ranking for each state:

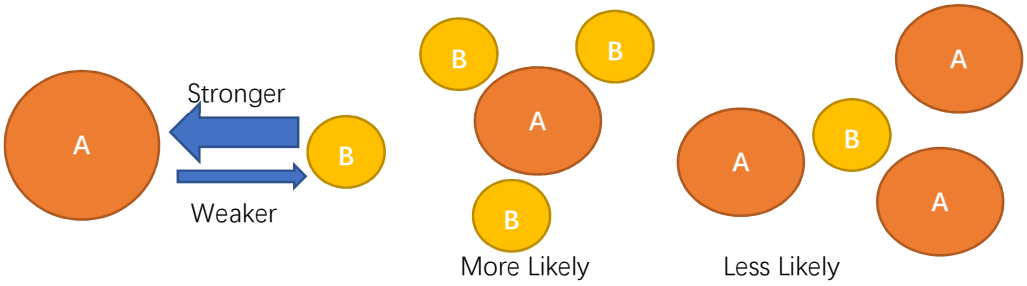
$$Score_{d,c} = Score\ of\ Spacing + Score\ of\ Timing$$

The score of spacing follows the following principles:

1. (main principle) The likelihood of a county (**called center county**) being as the drug source will increase if counties around it tend to use that drug (**called following counties**);
2. Further the following county is, the less it contributes to the likelihood of the center county being a drug source.
3. With the same distance, the smaller county contributes more to the larger county's likelihood than that of the larger county contributes to the smaller county. (By the consensus that larger county is more likely to be the source.



Graph 5: Illustration of Principle 2



Graph 6: Illustration of Principle 3

The score of timing follows the feature of a time series. That is, the closer to the first year of drug being used, the higher the number drug reports will contribute to the likelihood. A time degradation factor will be added to the number of drug reports of following years.

Based on the analysis above, the model is shown as following, for center county c:

$$Space_{d,c} = \sum_{i \in C} \left(\ln \left(\frac{p_{d,c,0} + p_{d,i,0}}{p_{d,i,0}} \right) \right)^{\frac{1}{25 * Dis_{ci}}}$$

$$Time_{d,c} = \sum_{t=0, t \in T} \delta^t p_{d,c,t}$$

$$C = \{All\ Candidate\ Counties\ except\ the\ Center\ County\}$$
$$T = \{0, 1, \ldots, 2017 - Year\ Released\ for\ Drug\ d\}$$

For example, if a drug d is released in 2013, and in 2013, Philadelphia, Jefferson, Wood, and Hamilton counties are using it. Then if the center county is Philadelphia, we can set

$$C = \{Jefferson, Wood, Hamilton\} \quad T = \{0,1,2,3,4\}$$

After the score of all candidates are calculated, we can normalize the score to the final probability of a specific center county c becoming the source of drug d:

$$Prob_{d,c} = \frac{Score_{d,c}^2}{\sum_{i \in C} Score_{d,i}^2}$$

6. Inverse Markov Chain Model

6.1 Model Construction

In section 6, the County Ranking Model tries to track back the source of drug by a static model. In order to improve the accuracy, another dynamic model is needed so that the drug's dissemination from counties to counties can be simulated.

Consider the drug spreading between counties, we can regard it as a drug flow: Initially, there is only the drug source county has drug reports. As time goes by, the drug may be spread to other counties so that there are reports from other counties as well. We can regard this process as "a number of drug flows from the source county to other counties". To quantify the effect of flow, the effectiveness can be shown by number of

reports in different counties.

Since we assume that the flow of drug is totally random, then the spreading process can be simulated by Markov Chain. However, we only know the number of drug users at the end of the first year when the drug came out. In order to track back the source of drug, we need to do the inverse of Markov Chain. The source of the drug can be tracked by Monte Carlo Method.

6.2 Algorithm of the Inverse Markov Chain

The algorithm of the inverse Markov Chain is composed by the following steps.

Step1: Construct the Graph

For specific drug d, find out the earliest year that the drug is used, and find out which county used it. Add them into the candidate set C. For example, for the drug Opiates, it was first used in 2010, and set C contains Delaware, Fairfield, Licking, and Perry.

Step2: Set the Initial Value of the Nodes

For each node, its initial value, which is the initial drug flux, is set to be the number of drug reports in the first year that the drug was released, all the nodes are connected.

$$v_{i,0} = p_{d,i,0}$$

Step3: Set the Transition Matrix of the Chain

The transition matrix T_l records the probability of stage transition, in which

$$tr_{ij,l} = T_{ij,l}$$

It stands for the unadjusted probability of one unit of drug flux flow from county i to county j. It is determined by the following equation:

$$tr_{ij,l} = \frac{v_{j,l}}{\left(\sum_{k\in C} v_{k,l}\right)\left(1 + Dis_{ij}\right)}$$
$$Note: Dis_{ii} = 0$$

According to the rule of Markov Chain that the sum of probabilities departing from a node should add up to 1, normalize tr.

$$\overline{tr_{ij,l}} = \frac{tr_{ij,l}}{\sum_{k\in C} tr_{ik,l}}$$

Step4: Tracking back for One Iteration and Update the Parameters

In each iteration, every unit of drug flux will decide its status in the next stage, and the decision will be made by random. For example, if there is a node A that has 50 unit of drug flux at the beginning of the simulation, and during the Markov Process, 30 units in total move to other counties (B & C), while 20 units "decide to stay" in A. Also, there are 10 units that decide to flow into county A. Therefore, at the beginning of the next iteration, there are 30 units of drug flux remaining in A. In formula,

$$v_{i,l+1} = v_{i,l} + in_{i,l} - out_{i,l}$$

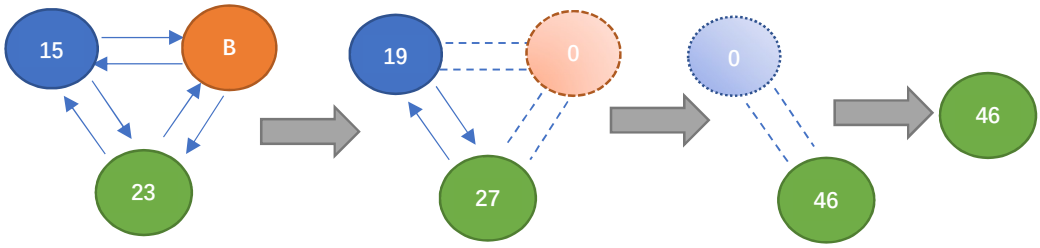
Step5: Examine the Status of Nodes

Examine whether there are nodes that have 0 unit of drug flux. If so, remove that unit from the graph. The rest of the nodes would form a new graph.

Step6: Repeat until the algorithm terminates

Repeat step 2 to step 5 until the there is only one node in the graph. Then this node will become the source of the drug d. Record the last-eliminated node of each state,

the corresponding county will become the location where drug is originated from.

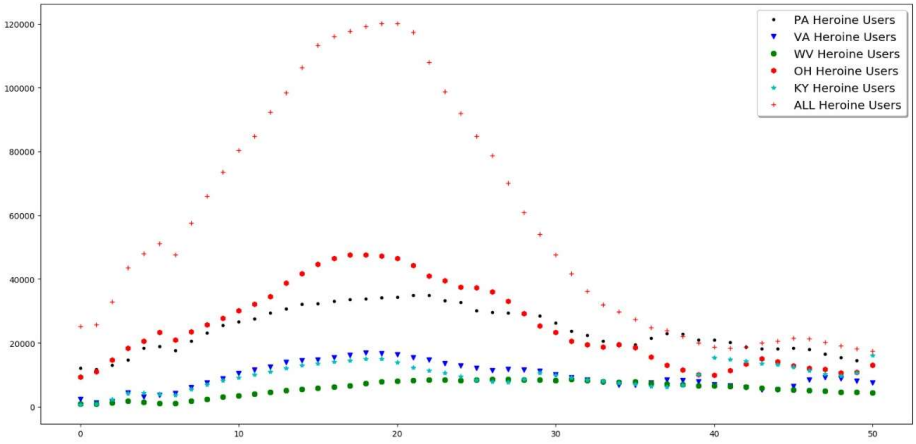


Graph 7: Illustration of Inverse Markov Chain

7. Verification of the Models

7.1 Result of the Drug User Estimation Model

Except for the concern in section 4.3, we tested the model by estimating the future Heroin cases for the 5 states. The result is shown as following:



Graph 8: Estimation of Heroin Usage of 5 states

It is clear that every state shows the peak in at around 2024~2025 more or less, which validates the concern in section 4.3. Also, it proves that the concern is not a one-state problem, but instead, it may become a problem of overall America.

7.2 Comparison of the County Ranking & Inverse Markov Chain Model

In order to find out which is the most likely drug source for each state for each drug, we use both the County Ranking and Inverse Markov Chain Model to test the data.

For the County Ranking Model, we pick the county with the highest probability (and score of ranking) within each state as the “most probable drug source” of that specific drug.

For the Inverse Markov Chain Model, we test the chain 100 times for each drug and pick the county that “survived” the most times of each state as the “most probable drug source” of that specific drug.

State	No. of time being Drug Source (By County Ranking)	No. of time being Drug Source (By Inverse Markov Model)
Virginia	28	28
Ohio	40	40
Kentucky	25	25
Pennsylvania	43	43
West Virginia	25	25

Table 3: Result of County Ranking & Inverse Markov Chain Model

In the County level, using the County Ranking Model, Fairfax, VA (11 times); Hamilton, OH (16 times); Jefferson, KY (10 times); Allegheny, PA (20 times); Ohio, WV (5 times) become the most probable drug source for each state, and Philadelphia, PA (11 times); Cuyahoga, OH (6 times); Kanawha, WV (4 times) are also counties that originates drugs relatively frequently in their states.

In the County level, using the Inverse Markov Chain Model, Fairfax, VA (7 times); Hamilton, OH (14 times); Jefferson, KY (9 times); Allegheny, PA (13 times); Ohio, WV (4 times) become the most probable drug source for each state, and Philadelphia, PA (12 times); Cuyahoga, OH (6 times); Kanawha, WV (4 times) are counties that are close to the counties above in each states.

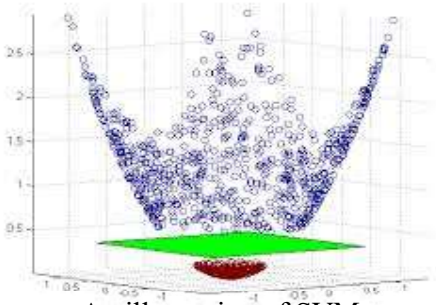
The two models provide similar result of the model. The similarity of the two models (for all drug in all states) is 91.12%. Results are attached to the appendix.

8. Support Vector Machine in Drug Usage Distinction

8.1 Support Vector Machine Method (SVM) Introduction

SVM is a supervised learning method that is used widely in the classification process. After obtaining the input of a set of training samples and the categories they belong to (usually two categories in total), SVM can create a hyperplane to separate the training examples that belong to one category from those belonging to the other.

Using the trained SVM, when a new testing example is input, SVM can output the category that it belongs to, which realize the classification of data.



An illustration of SVM
(From Google)

8.2 SVM Design for Drug Usage Distinction

From the data set of the U.S. Census Bureau, we notice that though there are around 465 counties in total, only about 358 has opioid usage report or Heroin usage cases. **Therefore, we first have to find out the difference between counties that use drugs frequently and those who use drugs less frequently using SVM learning method.**

Output Categorization:

In order to distinguish whether a county is using drug frequently, we set the

threshold to 1,000. We mark those counties that have more than 1,000 drug reports as Class Positive (denoted by 1), while we mark those counties that have less than 1,000 drug reports as Class Negative (denoted by 0).

Input Variable Choosing:

From the background research about drug usage, we conclude that the amount of drug usage of a community is related to its age distribution, ethnicity composition, gender distribution, education level, mental healthiness and so on. (Cari N., 2019; Nadia K., 2019; Phillips JK., Ford MA., Bonnie RJ., 2017; Berenson A., 2005)

Based on the research and the data we had from the U.S. Census Bureau, we define the input of each example (county) has the following property (all in No. of people):

- Population of the county (No. of households & Average household size);
- Marital Status (Never Married, Now Married, Separated and so on);
- Fertility (General Birth Rate, Unmarried Women Birth Rate);
- Education Level (Primary School, High School, Bachelor and so on);
- Ancestry (Place where people are originated from);
- Disability of People;

Within those inputs, we make slight adjustment to some categories. For example, we conclude all European Ancestry into one group, and we merge the group with Bachelor degree & Master of Professional Degree. That is, we merge some similar groups in order to prevent over fitting of the SVM.

8.3 Training & Testing Result

For each year, we include the following variables in the input:

- Estimated Total Household;
- Estimated Average Household Size;
- Number of People that Never Married;
- Number of People that are Married but not Separated;
- Number of People that are Separated, Widowed, or Divorces;
- Number of People with College Degree;
- Number of People without College Degree;
- Number of People whose Ancestry is Arab;
- Number of People whose Ancestry is Sub-Saharan African;
- Number of People whose Ancestry is Europe (including all Europeans);
- Number of People whose Ancestry is Slavic;
- Number of People who are from North America and England;

Therefore, our input is a vector with length 12:

$$x = [x_1, x_2, \dots, x_{12}]$$

According to the soft margin SVM classifier, we need to

$$minimize \ \frac{1}{2} w^T \cdot w + C \sum_{i=1}^{m} \zeta^i$$

We are using polynomial kernel function, so the objective function is subject to

$$t^i(\gamma w^T \cdot x + r)^d \geq 1 - \zeta^i$$

For each year, we randomly assign part of the counties as training set, and the rest of them are the testing set. We use the training set to train the SVM and use the testing

set to test whether the SVM is robust.

The training report is as following, and the code of SVM is attached to the appendix.

Hyper Parameters							
$C = 5, r = 1, d = 3,$ $Training\ Set = 0.7$ $Testing\ Set = 0.3$							
Year	2010	2011	2012	2013	2014	2015	2016
Testing Accuracy	94.96%	96.40%	95.68%	94.78%	95.52%	96.74%	95.65%

(From 2013, the category Arab is removed, so one instance changes to 11 inputs; From 2014, the county "Bedford City" is removed which is not included in the set)

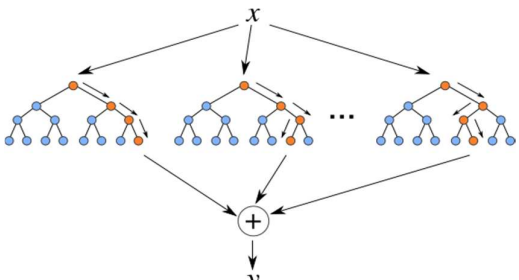
Table 4: Result of the SVM

Therefore, from the SVM, given the feature of the county, we can estimate whether the county is probable to have more frequent drug usage. Since the SVM performs well on the testing set, it is possible to generalize the model to provide a approximate drug usage prediction of the county.

9. Random Forest in Predicting Drug Usage

9.1 Introduction to Decision Tree and Random Forest

Decision Tree is a tree-like model that uses a tree-like model of decisions to realize classification and regression. An input is classified or regressed by checking at the decision tree nodes and selecting its future branches. It will end up in leaf nodes, which represent the final result of classification and regression. The core method of creating a decision tree is to decrease the general information impurity of all the data by choosing different feature of the input as decision criterion.



An illustration of Random Forest
(From Google)

Random Forest is a kind of ensemble learning for classification and regression. The main algorithm is construct multiple decision trees based on different groups of training data, and take the mode of class (classification) or mean of regression (regression) of the trees. Random Forest can provide with a better result than a single decision tree because of the randomness in the training will decrease the probability of overfitting.

9.2 Random Forest model of Drug Report Classification

Since the number of drug reports in each state varies significantly, the Random Forest with the Decision Tree Regressor may not be able to estimate the precise value. Therefore, we opted to use Random Forest Classifier to classify the number of drug reports.

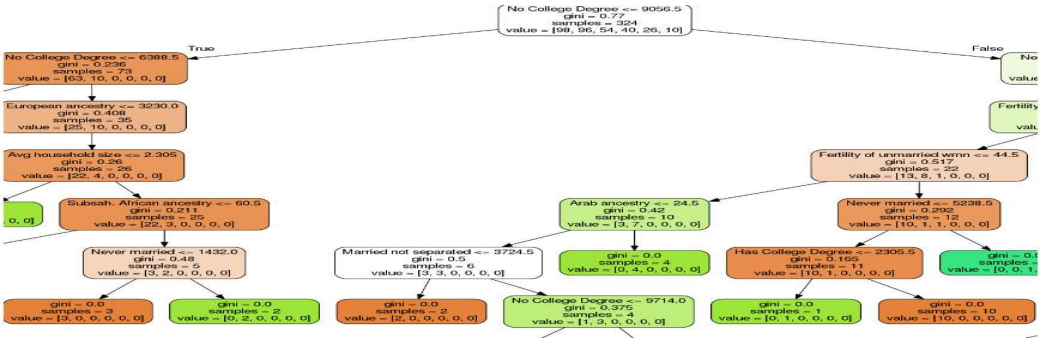
In order to simplify the classification, we divide the counties into 6 categories by the number of drug reports by each year: 0~100 (denote class 0), 101~250 (denote class 1), 251~500 (denote class 2), 501~1000 (denote class 3), 1001~3000 (denote class 4),

3000~ (denote class 5). Then we try to minimize the "Gini Impurity".

$$Gini\ Impurity_i = 1 - \sum_{k=1}^{n} p_{i,k}^2$$

In the equation, $p_{i,k}$ is the ratio of class k instances among the training instances in the i th decision node. For example, if a node can perfectly distinguish whether an instance is from class A or B, then under its branches, all units under one branch should be either A or B; then it's Gini Impurity is 0. The goal of the Decision Tree method is to find the criterion that can create the lowest Gini Impurity.

In our model, we use the same 12 inputs as those in section 9, the outputs are number of the categories (from 0 to 5). Using scikit-learn in Python, we create the random forest classifier. A part of one decision tree in the forest is shown below:



Graph 9: A Part of the Decision Tree

The entire decision tree is attached to the appendix.

Using the Random Forest Model, the out-of-bag index (an index that is used to check the performance of the forest) is averaged at 0.6 for the average of four years. Though the result is good, it is still not satisfying. The reason why the random forest cannot make high accuracy prediction is because the input variables from the original data set are still limited. There are likely still other reasons that some counties have higher drug usage that are not accounted by the data (or variables) we currently have, so they are not able to be added to the decision tree.

Despite these shortcomings, the random forest still provides some useful information: the root decision node. As the first decision node, it is chosen because it can decease the Gini Impurity the most. In other words, it is the most significant feature to distinguish counties from which category.

From 2010 to 2016, the root decision nodes are the following:

Year	2010	2011	2012	2013	2014	2015	2016
Node Variable	No. of People without College Degree	No. of People without College Degree	No. of House-holds	No. of People Get Divorced	No. of People Get Divorced	No. of People Get Divorced	No. of People Get Divorced

Table 5: The Root Decision Nodes for Each Year

Therefore, from the result above, **we can see that No. of People without college degree and No. of People who get divorced become important factors to decide**

whether a county may have more drug users. It is reasonable since low-educated people are comparatively more likely to use drugs, and the people that are divorced may have higher probability of getting mental illness such as depression, making them prone to entering/staying in a cycle of drug addiction.

10. Strategy of Countering Opioid Crisis

10.1 Regulate the Opioid Usage in each County

The regulation includes stop the lab from inventing new opioid and prevent the clinics or hospitals use opioid prescription too frequently.

For the former one, if a new opioid is introduced, then it will experience a procedure that more and more people start to use it until the frequency of usage comes to a peak. Forbidding from inventing new opioid can decrease the life circle of a specific drug, which can further control the drug use.

For the latter one, controlling clinics and hospitals from using opioid prescription can decrease the possibility that people intaking opioid, which also further decrease the probability of people getting addicted to drugs.

10.2 Restrict Heroin Usage and Heroin Transaction

As a prescription medicine, opioid drugs have their medical usage. However, it will become harmful when the medical usage turns to addiction in Heroin. Therefore, another way to deal with the opioid crisis is to restrict Heroin Transaction. If there is no channel for people to get Heroin as substitution, then the negative effect of opioid usage can be solved.

10.3 Take Care of Specific Group of People

From section 9 and 10, we acknowledge that some specific group of people (people with low education level and who are divorced) may be positively related to drug usage. Local governments can take more care of those group of people. On one hand, the government should put an eye on the drug usage of people in these groups and the people related to them. More importantly, the local government should put more funding toward resources and programs that can remedy and counsel the mental health of the people, such as for people undergoing divorce, or people with low education level. Since opioid is mostly used to deal with pain and depression, if these people's mental health is taken care of by alternatives resources, less drugs will be used and have the opportunity to be abused.

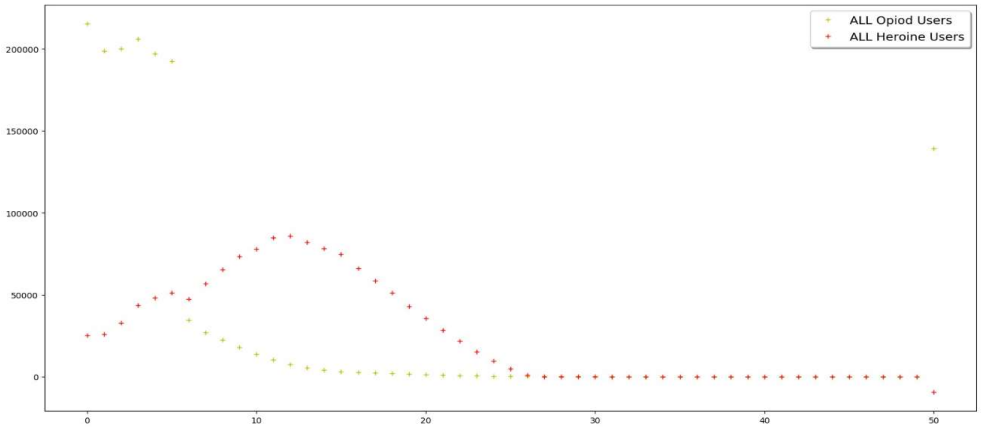
11. Estimation of Effectiveness & Sensitivity Analysis

In 4.3, it is found that the U.S. Government would have a potential concern that the cases of people using Heroin in the 5 States will reach 100,000 and being kept for 5~6 years. Also, it needs 15~20 more years to get back to the current level. Therefore, we are going to analyze the effectiveness of the model by whether the strategy can

eliminate the concern of the U.S. Government.

Effectiveness of Strategy 1:

In order to test the Effectiveness of Strategy 1, we need to simulate the case where opioid usage is restraint, so we scale the term $\left[\left(1-\frac{1}{G_i(t)}\right)F_i(t)\right]'$ in the first differential equation in section 4.4 by factor 0.8, and plot the estimation of opioid and heroin Usage:

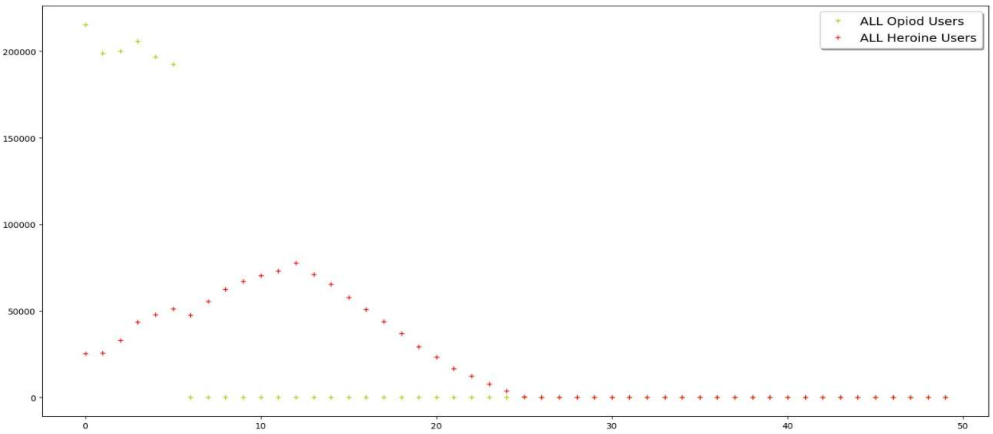


Graph 10: Opioid and Heroin Usage Estimation using Strategy 1

We can see that though the Heroin cases of 5 states will still reach a peak at around 2024~2025, the value of the peak drops to around 80,000, and the opioid usage report will decrease dramatically in a few years. **So, it is proved that this strategy is useful.**

Effectiveness of Strategy 2:

In order to simulate the case where the channel of getting Heroin is cut or regulated by government, we change the β value (the rate of people changing from opioid user to Heroin user) from 0.055 to 0.05, and we get the following result:



Graph 11: Opioid and Heroin Usage Estimation using Strategy 2

We can see that the peak of Heroin usage comes earlier to around 2021~2022, while the value of Heroin usage cases decreases from 100,000 to around 75,000. Also, the peak will no longer continue for 5~6 years, but the Heroin usage cases will decrease linearly at a faster speed. **So, it is proved that this strategy is useful.**

Effectiveness of Strategy 3:

For the Random Forest Model that is trained using the data of 2016, since our strategy is to take care of specific group of people, we factor the number of people who are categorized in "People without College Degree" and "People that get divorced" by 0.8 to simulate that those people become unlikely to take drugs.

After we plug in the adjusted data to the Random Forest Model of 2016 (latest one), the model prints out the drug usage level of each county. Compared with the original result with the data unadjusted, 20 counties' drug usage levels decrease for at least 1 level, while only 3 counties' drug usage levels increase. **Therefore, it is proved that this strategy is useful.**

Except for the strategies, we also test the sensitivity of the SVM. We test when the frequency threshold is equal to 200, 500, and 800.

Hyper Parameters							
$C = 5, r = 1, d = 3,$ $\quad Training\ Set = 0.7 \quad Testing\ Set = 0.3$							
Year	2010	2011	2012	2013	2014	2015	2016
Testing Accuracy (200)	82.80%	77.42%	78.36%	78.49%	74.19%	76.48%	75.04%
Testing Accuracy (500)	94.62%	88.17%	86.02%	84.95%	89.25%	93.78%	87.10%
Testing Accuracy (800)	96.77%	94.62%	91.40%	92.47%	96.28%	95.83%	93.27%

Table 6: Sensitivity Analysis of SVM

From the table, we see that in general cases, the SVM method remains a high accuracy, which is increasing as the threshold value increases. It is reasonable since when the threshold level is low, more pieces of data are clustered and overlapping, which provides difficulty for the SVM to classify those data.

12. Model's Strengths and Weaknesses

Strengths of the Model:

For the models that estimate the opioid and heroin usage (section 4, 5, and 6), we analyze the data in a comprehensive way. The County Ranking Model provides a static view of tracing the source of drugs, while the Inverse Markov Chain Model dynamically and stochastically track back the source of drugs. The Drug User Estimation Model is modified from the SIR model but adds white noise so that it's closer to real life.

Also, we focus more on the details of the model. For example, we consider the "drug flow" between counties (section 4 & 6), the time degradation factor (section 5), and the difference of inter-effect between big county and small county (section 5). The consideration of those details will make the model more accurate.

For the models that try to find out the related factors with drug usage (section 9 & 10). Using machine learning can provide an easier way to deal with large amount of data with low time and space complexity. Furthermore, the model using machine learning method is more robust. That is, it can be easily adjusted according to different input variables, and it can also be used to explore the relationship between drug and other factors that are not in the data set (such as races, gender, etc.) with a high efficiency and good performance.

Weaknesses of the Model:

For the models that estimates the opioid and heroin usage, The Drug Usage Estimation Model has too many parameters so that it becomes hard to control. Some fluctuation in the original data may cause a big change in the parameters of the model. As for the County Ranking Model and Inverse Markov Chain Model, the concepts of the model (such as score of spacing, drug flux) are concept so that sometimes it may be hard for others to understand.

For the models that try to find out the related factors with drug usage. Because we use machine learning techniques, some models, like the random forest classifier, become black box models. That is, using computer to train the data, we may not know the complete inside structure of the model. In addition, when testing the model, the input is directly matched to an output (category or value), making it hard for others to find out the intermediate steps that occurred.

Memo to the Chief Administer, DEA/NFLIS Database

Dear Sir,

We are team # 1923122 from MCM, and we studied your data based combining with the socio-economic factor database from the U.S. Census Bureau in the past few days. We discovered some important trends and are going to share the conclusions we reached with you.

Though the numbers of drug identification cases of opioid drug usage and heroin usage cases fluctuate from 2010 to 2017, we found statistical evidence that these changes are related. From our model, we found that consistent with all the news and rumors, people who take the prescription of opioid drugs will often end up taking Heroin as substitution when they are addicted to the opioid drugs. This is the reason why we want to call your concern:

If the opioid drug is not regulated, according to the trend, there will be more than 100,000 Heroin usage reports per year within the five states (VA, OH, KY, PA, and WV) around 2024 to 2025, and this phenomenon will continue for 5 to 6 years until the year 2030. As for the recovery, it may take 15 more years to let the Heroin usage level drop to the current level. Therefore, the regulation of opioid drugs is necessary and urgent.

Here, we have some evidence of where those drugs are originated from. Based on the estimation of our models, the following counties have the highest probability of being origins of those new opioid drugs:

- Fairfax, VA
- Hamilton, OH
- Cuyahoga, OH
- Jefferson, KY
- Allegheny, PA
- Philadelphia, PA
- Ohio, WV
- Kanawha, WV

Therefore, please pay more attention to the regulation of these counties, since they tend to be the center of opioid drugs.

Talking about the opioid drugs, we discover that in recent years (especially 2016 and 2017), there are many new opioid drugs being created and used. It will make our current situation even worse, since every new drug has its life circle, so that when it is spreading, more people may use this opioid drug and eventually get addicted to heroin.

Using a machine learning method, we also found out that for people who don't have a college degree or people who are divorced, they are positively related with opioid or heroin usage. It may partially be because they are more easily to get addicted, or they may be more vulnerable to neglecting mental health that could lead to the use of opioid drugs.

Also, we create a program called support vector machine (SVM), so as far as you input some key feature of the community, such as population, and set a drug usage

threshold (total drug report you estimated), the program will come out with the result that you are overestimating or underestimating. The program's accuracy is mostly greater than 85%, especially for those big counties.

The very last thing I want to mention is that we discuss the strategy of countering the current opioid crisis in the U.S. There are three ways that we proved useful based on our model:

- Regulate the opioid usage in counties. It includes development of new opioid drugs and giving opioid prescriptions. In this way can the total amount of opioid use be reduced;
- Cut the channel of heroin. People usually switch from opioid to Heroin, because heroin is easy to get. Therefore, if the government can make it harder for citizens to get heroin by reducing gateway drugs that can be prescribed, then the cases of heroin abuse will be decreased;
- Look out for certain groups of people. Since there groups of people in which opioid drugs are more popular in, we need to take more care of their mental health by offering alternative programs and resources to help assist any struggles they may encounter. This is a strong way to counteract the opioid crisis by attacking it from its source: why people need/desire opioid drugs in the first place. If this problem can be addressed, then their use will gradually decrease.

In conclusion, based on the data, we believe that the opioid crisis is not something that can not be handled. People let opioid become popular by overdraft their happiness because of their will, but now it is time for us to pay the bill.

This memo covers all my discoveries and suggestions, thanks for reading!

Best wishes,
Team # 1923122

References:

American Addiction Center (Jan. 25th, 2019). Fentanyl vs. Heroin: The Similarities and Differences between Two Powerful Opioids. Retrieved from: https://americanaddictioncenters.org/fentanyl-treatment/similarities

Berenson A. (Jan. 28th, 2019) Shifting Patterns of Prescription Opioid and Heroin Abuse in the United State. New York Times, December 4, 2005.

Cari N. (Jan. 27th, 2019). Who Uses Heroin? Not Who You Might Think. Retrieved from: https://www.livescience.com/45969-who-uses-heroin.html

Clair F. (Jan. 25th, 2019). The U.S. Opioid Epidemic. Retrieved from: https://www.cfr.org/backgrounder/us-opioid-epidemic

Google (Jan. 28th, 2019) Harp Documentation. Retrieved from: https://www.google.com/search?q=random+forest&source=lnms&tbm=isch&sa=X&ved=0ahUKEwi-0byto5DgAhWKDnwKHXAnDuUQ_AUIDygC&biw=2133&bih=1020#imgrc=83JSc_QyJnVyRM:

Google (Jan. 28th, 2019). SVM Versus a Monkey. Retrieved from: https://www.google.com/search?q=support+vector+machine&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjnkezyzY_gAhUHGnwKHfaSCqkQ_AUIECgD&biw=2133&bih=1074#imgrc=RqBTnesgBtL5zM:

Nadia K. (Jan.27th, 2019). Greatest Rise in Heroin Use was among White People, Study Says. Retrieved from: https://www.cnn.com/2017/03/29/health/heroin-abuse-increase-study/index.html

National Institute on Drug Abuse (Jan. 25th, 2019). Opioid Overdose Crisis. Retrieved from: https://www.drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis

Phillips JK., Ford MA., Bonnie RJ. (Jan. 27th, 2019). Trends in Opioid Use, Harm, and Treatment. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK458661/

Sean M., Michele M., James C., Jorge D., Melnee M., & Carol B. (Jan. 28th, 2019) Race/Ethnicity and Gender Differences in Drug Use and Abuse Among College Students. J Ethn Subst Abuse. 2007; 6(2): 75-95.

U.S. Department of Health and Human Service (Jan. 25th, 2019). What is the U.S. Opioid Epidemic? Retrieved from: https://www.hhs.gov/opioids/about-the-epidemic/index.html

Appendix I: Result of County Ranking Model

Name of County	Times	Name of County	Times
Virginia		**Ohio**	
Fairfax	11	Hamilton	16
Chesterfield	3	Cuyahoga	6
Stafford	2	Montgomery	3
Wise	2	Franklin	2
Fauquier	1	Greene	2
Franklin	1	Allen	1
Hanover	1	Athens	1
Pulaski	1	Clark	1
Roanoke	1	Huron	1
Scott	1	Jackson	1
Shenandoah	1	Lake	1
Spotsylvania	1	Licking	1
Warren	1	Lorain	1
Wythe	1	Lucas	1
Kentucky		Warren	1
Jefferson	10	Wood	1
Fayette	4	**Pennsylvania**	
Kenton	2	Allegheny	20
Warren	2	Philadelphia	11
Barren	1	Dauphin	2
Campbell	1	Delaware	2
Grayson	1	Armstrong	1
Harlan	1	Crawford	1
Hopkins	1	Fayette	1
Perry	1	Lycoming	1
Taylor	1	Monroe	1
West Virginia		Northampton	1
Ohio	5	Northumberland	1
Kanawha	4	York	1
Mercer	3	**West Virginia**	
Raleigh	3	Marion	1
Berkeley	2	Mingo	1
Harrison	2	Nicholas	1
Mineral	2	Wood	1

Appendix II: Result of Inverse Markov Chain

Name of County	Times	Name of County	Times
Virginia		**Ohio**	
Fairfax	7	Hamilton	14
Chesterfield	3	Cuyahoga	6
Roanoke	3	Franklin	2
Richmond	2	Greene	2
Stafford	2	Lake	2
Wise	2	Lucas	2
Wythe	2	Montgomery	2
Franklin	1	Allen	1
Hanover	1	Athens	1
Pulaski	1	Huron	1
Scott	1	Licking	1
Shenandoah	1	Lorain	1
Spotsylvania	1	Madison	1
Warren	1	Stark	1
Kentucky		Summit	1
Jefferson	9	Warren	1
Fayette	3	Wood	1
Warren	3	**Pennsylvania**	
Kenton	2	Allegheny	13
Barren	1	Philadelphia	12
Campbell	1	Bucks	2
Grayson	1	Dauphin	2
Harlan	1	Delaware	2
Hopkins	1	Luzerne	2
Muhlenburg	1	York	2
Perry	1	Beaver	1
Taylor	1	Crawford	1
West Virginia		Fayette	1
Ohio	4	Lycoming	1
Kanawha	3	Monroe	1
Mercer	3	Northhampton	1
Mineral	3	Northumberland	1
Raleigh	3	Schuylkill	1
Harrison	2	**West Virginia**	
Marion	2	Hardy	1
Berkeley	1	Nicholas	1
Upshur	1	Wood	1

Appendix III: (Partial) Code for SIR Model (In Python)

Remark: Since we create a class for county & drugs, the code can't directly run, but the code is for reference.

```
#movement function used to model population transfer between states
def movementFunction(state1,state2):
    return 1/(10*distance(state1,state2)) if state1 != state2 else 0

#A general template of a harmonic functionn
def generalHarmonic(a0,a1,a2,freq):
    return (lambda x : a0 + a1 * math.cos(freq*x) + a2 * math.sin(freq*x))

#various functions describing fourier functions for each state
def PAfit(x):
    Fa0 = 15110
    Fa1 = 250.9
    Fa2 = -3858
    Ffreq = 0.8332
    Ga0  = 0.1922
    Ga1 = -0.0306
    Ga2 = -0.0527
    Gfreq = 0.7091
    return            (generalHarmonic(Fa0,Fa1,Fa2,Ffreq)(x)                    -
generalHarmonic(Fa0,Fa1,Fa2,Ffreq)(x)                                           *
generalHarmonic(Ga0,Ga1,Ga2,Gfreq)(x))/generalHarmonic(Ga0,Ga1,Ga2,Gfreq)(x)

def VAfit(x):
    Fa0 = 2976
    Fa1 = -692.4
    Fa2 = -948
    Ffreq = 0.6168
    Ga0 = 0.0863
    Ga1 = -0.0089
    Ga2 = -0.041
    Gfreq = 0.6975
    return            (generalHarmonic(Fa0,Fa1,Fa2,Ffreq)(x)                    -
generalHarmonic(Fa0,Fa1,Fa2,Ffreq)(x)                                           *
generalHarmonic(Ga0,Ga1,Ga2,Gfreq)(x))/generalHarmonic(Ga0,Ga1,Ga2,Gfreq)(x)

def WVfit(x):
    Fa0 = 1253
    Fa1 =  -152.9
    Fa2 = -374.9
    Ffreq = 0.9638
```

```
    Ga0 = 0.1652
    Ga1 = -0.0562
    Ga2 = -0.0344
    Gfreq = 0.5944
    return            (generalHarmonic(Fa0,Fa1,Fa2,Ffreq)(x)                        -
generalHarmonic(Fa0,Fa1,Fa2,Ffreq)(x)                                               *
generalHarmonic(Ga0,Ga1,Ga2,Gfreq)(x))/generalHarmonic(Ga0,Ga1,Ga2,Gfreq)(x)

def OHfit(x):
    Fa0 = 16400
    Fa1 = -4706
    Fa2 = -4537
    Ffreq = 0.6858

    Ga0 = 0.1698
    Ga1 = -0.0354
    Ga2 = -0.0226
    Gfreq = 0.743
    return            (generalHarmonic(Fa0,Fa1,Fa2,Ffreq)(x)                        -
generalHarmonic(Fa0,Fa1,Fa2,Ffreq)(x)                                               *
generalHarmonic(Ga0,Ga1,Ga2,Gfreq)(x))/generalHarmonic(Ga0,Ga1,Ga2,Gfreq)(x)

def KYfit(x):
    Fa0 = 2310
    Fa1 = -2149
    Fa2 = -52.4
    Ffreq = 0.5557

    Ga0 = 0.0895
    Ga1 = -0.0779
    Ga2  = -0.0165
    Gfreq = 0.5935
    return            (generalHarmonic(Fa0,Fa1,Fa2,Ffreq)(x)                        -
generalHarmonic(Fa0,Fa1,Fa2,Ffreq)(x)                                               *
generalHarmonic(Ga0,Ga1,Ga2,Gfreq)(x))/generalHarmonic(Ga0,Ga1,Ga2,Gfreq)(x)

def totalfit(x):
    Fa0 = 4000#38000
    Fa1 = 4000#7503
    Fa2 = 2000#11070
    Ffreq = 0.7192

    Ga0 = 0.156
    Ga1 = 0.0296
```

```
    Ga2 = 0.0418
    Gfreq = 0.7314
    return            (generalHarmonic(Fa0,Fa1,Fa2,Ffreq)(x)                    -
generalHarmonic(Fa0,Fa1,Fa2,Ffreq)(x)                                           *
generalHarmonic(Ga0,Ga1,Ga2,Gfreq)(x))/generalHarmonic(Ga0,Ga1,Ga2,Gfreq)(x)

#function takes state and returns a fourier fuction cooresponding to the state
def FitData(st):
    switcher = {
        PA : PAfit,
        VA : VAfit,
        WV : WVfit,
        OH : OHfit,
        KY : KYfit,
        total : totalfit
    }
    return (lambda i : (switcher[st](i) - switcher[st](i-1)) )
```

Appendix IV: (Partial) Code for Inverse Markov (In Python)

```
while(len(G) > 1):
    # go through each county
    for i, countyCase in enumerate(countyListCases):
        [county_i, dCases] = countyCase
        for k in range(0, dCases): # go through each "case"
            for j, county_j in enumerate(countyList):
                if random() < G[county_i][county_j]['weight']: # if "case" decides to
move
                    countyListCases[i][1] -= 1 # decrease drug case amount in host
county
                    countyListCases[j][1] += 1 # increase drug case amount in neighbor
county

            # remove any dead nodes from
            i = 0
            while i < len(countyListCases):
                county_i, dCases = countyListCases[i]
                if dCases <= 0:
                    stateCount[county_i.m_state] -= 1
                    if stateCount[county_i.m_state] <= 0:
                        stateRecords[county_i.m_state][county_i.m_name] += 1
                    G.remove_node(county_i)
                    countyListCases.pop(i)
                    countyList.pop(i)
                else:
                    i += 1

            # refresh node weights
            totalCases = County.sumPop(lambda x: x.drugCases(drug, BEGIN_YR), drug,
BEGIN_YR, countyList)
            for county_i in countyList:
                for county_j, dCaseJ in countyListCases:
                    dist_ij = 1 + county_i.distanceTo(county_j)
                    G[county_i][county_j]['weight'] = calcEdge(dCaseJ, totalCases,
dist_ij)
            normalizeEdges(G)
```

Appendix V: (Partial) Code for SVM (In Python)

Prepare Dataset

```
cleanedData = pd.concat([
    household,
    neverMarried,
    marriedSeparate,
    divorced,
    notCollegeGrad,
    collegeGrad,
    arab,
    africa,
    europe,
    slavonic,
    na
], axis=1)

partition = int(TEST_RATIO*uscbYr.shape[0])
uscbTest = cleanedData[:partition]
expectedTest = expected[:partition]
uscbVal = cleanedData[partition:]
expectedVal = expected[partition:]
```

Scale data, run SVM

```
poly_kernel_svm_clf = Pipeline([
    ("scaler", StandardScaler()),
    ("svm_clf", SVC(kernel="poly", coef0=1, C=5))
])
poly_kernel_svm_clf.fit(uscbTest, expectedTest)

pred = poly_kernel_svm_clf.predict(uscbVal)
accuracy_score = np.mean(pred == expectedVal)
```