For office use only

T1 ________

T2 ________

T3 ________

T4 ________

Team Control Number

1926093

Problem Chosen

C

For office use only

F1 ________

F2 ________

F3 ________

F4 ________

2019

MCM/ICM

Summary Sheet

A Systematic Analysis of the Epidemic Opioid Crisis Based on Data Profile

Desire, depravity and sin are dark sides of human beings and with the fast development of economy and technology, these dark sides are devouring our society. Abusing drug is the worst of all sins. In order to solve the devastating drug crisis in the society of United State, our team analyze the data of reported drug provided by NFLIS and build a model based on the pattern we find. After optimizing the model with the socio-economic data provided, we propose some recommendations for the government regarding how to alleviate the situation based on our model.

For part I, our team first plot the geographical distribution of the total drug report in 2010 and 2016 to better visualize the data. Then we identify that the secondary derivative of reported drug over time is related to the among of existing drug reported. Based on this characteristic and the nature of the drug crisis, we decide to use SIRS to model this crisis. After fitting the derivative of drug report over time, we can plot the overall pattern of drug reported over time. With the pattern, we classify the counties into three categories and identify sources, threshold, and prediction based on the category of each county.

For part II, we first perform Principle Component Analysis of the socio-economic data and get a matrix consisting of the projections of the original factors to new axis. Then we use Lasso Regression to fit the derivative of drug report in each county to identify which entries in the PCA matrix correlate to the change of drug report. After back tracing these entries to their original factors, we can identify the responsible factors in the change of drug report.

Finally, we use least square regression to optimize our SIRS model in part I using the result of Lasso Regression and derived new parameters for the model. The resulting R square of the fitting shows the accuracy of our model. Based on this new optimized final model our team can show the correlation between the factor we find in Part II and the changing rate of drug report k. Thus, we can provide strategy of how to solve the crisis by making changes to these factors.

Contents

Memo

From: MCM 2019 Team 1926093
To: The group of Governors
Date: 28, January 2019
Subjects: A heuristic analysis of opioid crisis

Dear governors, it is our honor to show you the result of our data analysis and modeling based on the data provided by NFLIS. With the data, we make several conclusions based on the situation for these counties in the five states and also provide some recommendations regarding how to stop the existing trend of abusing drug. The details are listed as follows:

Current situation:

- There are several counties that are classified as the source for several types of opioid: Philadelphia in Pennsylvania, Hamilton in Ohio, Jefferson in Kentucky, Fairfax in Virginia and Berkeley in West Virginia. In these counties, government should spend more expenditure on drug crime investigation.
- There are several counties that are in the stage of threshold. So we warn the governments of these counties to use our strategy as soon as possible so that these counties will not become a source of drug like the Five counties above. These counties are: Marshall, Knox, York, etc. The detailed table is in the paper.

Recommending Strategy:

In the second part of our paper, we give a list of factors that can either positively or negatively influence the changing rate of drug report. Our strategy is to encourage the negative factors and eliminate or at least decrease the positive factors. We give several examples:

- Since college-level education is negatively related to the changing rate of drug report, the government should spend more funds on education, especially on college level and over.
- Immigration population for both entering before and after 2000 is positively related to changing rate of drug, so the should probably limit the number of immigrants. (This result doesn't represent our own opinion. It is just the result of data analyzing. Please discuss this issue with Mr. Trump.)
- The population of people who doesn't possess housing is also positively related to the changing rate. So the governments should probably establish some policy to limit the number of homeless people.

If you have any questions, please contact us.

1 Introduction

1.1 Problem background

Addiction to drug can cause a series of social problem if people like high-educated people or important figures in industries are devoured by this monster. This crisis cannot be countered by simply enforcing existing laws. As a result, the DEA/National Forensic Laboratory Information System (NFLIS) is calling actions to analyze the data from crime laboratory in order to find pattern among the data so that we can prevent this epidemic crisis from polluting our society further. The data from five states of U.S are provided (Ohio, Kentucky, West Virginia, Virginia, and Pennsylvania).

1.2 Our Goals

Based on our analysis of the problem, we set the following goals:

- Use the given data of Drug Report from each county to find the pattern (i.e. a development model) behind the crisis.
- With the model, find the county in which each different types of opioid might have started for each of the five states.
- Identify a threshold to warn government once a county has passed it.
- Analyze the socio-economic data and optimize the model based on the factors it provides.
- Finally identify a possible strategy to counter the crisis.

2 Assumption and Notation

2.1 Assumption

In order to have a clearer understanding of the provided data and perform modeling with our knowledge, we make several assumptions before the analysis. All of our subsequent models and analysis are based on these assumptions.

- The number of drug reported is positively related to the real number of people addicted to opioid and heroin.
- The drug-related law in these states will not be changed in the following years so that there will be no unexpected significant change in drug report.
- Even without any strategy, the number of people addicted to drug will not grow infinitely to the entire population; instead, it will converge to a number.
- Once a county is suspected to be the earliest to develop in the drug crisis, it is more likely to be the source for all kinds of drugs.

2.2 Notation

- S - susceptible people that can be addicted to drug in each county.
- I - people that are addicted to drug in each county.
- R - people that have detox in each county.
- N - maximum number of people that can be addicted to drug even without any strategy for each state.
- k - changing rate of drug reports.
- k' - derivative of k.
- β - the rate of people addicting to drug of a county.
- γ - the rate of detoxing.
- ϵ - the rate which detoxed individuals return to the susceptible state.

3 Data Processing

First, we classify the data in the worksheet into three categories: redundant data, missing data, and normal data. Then, as for redundant data, we remove them to make it more convenient for us to search the required data. As for missing data, we fill them with either 0 or mean of nearby values. As for normal data, we sort them and study the factors. Also, we derive some additional information from existing data. The works we have done in data processing are as follows.

3.1 Delete the redundant data

There is data redundancy in the worksheet "MCM NFLIS Data". As we can see in Table 1, "State" and "FIPS_State" can be derived from "FIPS_Combined". So, we deleted the redundant data "State" and "FIPS_State".

Table 1: Example of redundant data

FIPS_State	**FIPS_County**	**FIPS_Combined**	**...**
51	1	51001	...
39	1	39001	...

3.2 Fill the missing values

Again, in the worksheet “NFLIS Data”, reports of drugs are not always available in each county in each year. In order to solve this problem, we filled the missing values with means of nearest two years. For the first year and last year, values of nearest one year is used to fill missing values. For other data sets, which are containing a large amount of missing values, we simply removed them from the data sets for later analysis.

4 Part I Analysis: Model Construction, Data Fitting, and Prediction

4.1 Model construction: SIRS model

To begin with, we plot the total drug reports in each county to better visualize and identify the pattern in the development of this drug crisis.

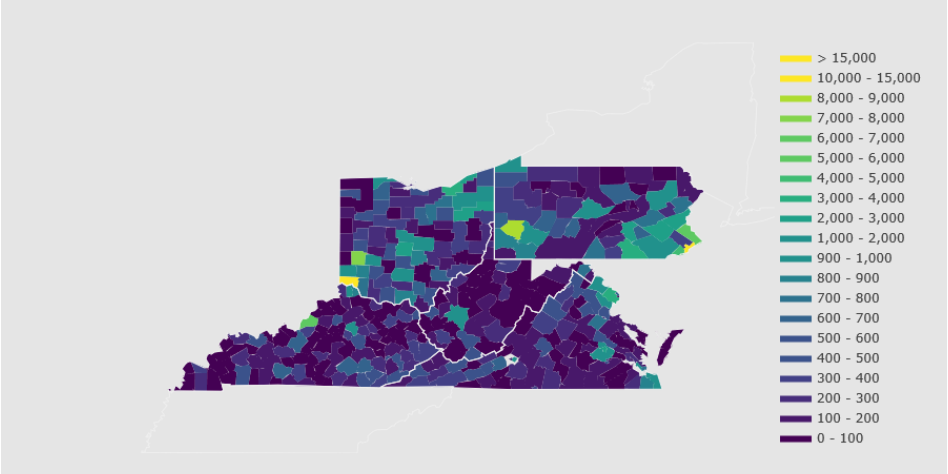


Figure 1: Drug Reports Number Distribution 2010

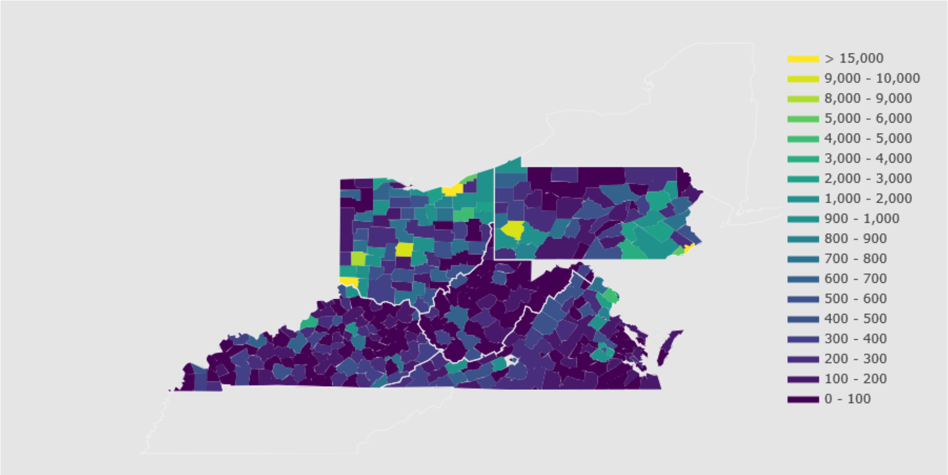


Figure 2: Drug Reports Number Distribution 2016

After processing the data from the five state, we can find that from 2010 to 2016, the drug report distribution of Ohio state and Pennsylvania state increase significantly. Also, there is an overall trend of shifting of drug from east coast to inland states. Besides, as the figure below, we find that the derivative of drug reports changing rate, k', is not 0 for each of the county and it is also related to the existing number of drug reports. **These k' are obtained by performing linear regression on the k of each county.** Combing this fact with our assumptions:

- Addicted people can detox.

- The number of people addicted to drug will converge to a number.

We find out that SIRS model, the model for infectious diseases, fits our context extremely well because this model has infectious rate, recover rate and return-to-susceptible rate, which correspond to our rate of addiction to drug, detoxication rate, and re-addiction rate. So we decide to use SIRS model to simulate the change of drug report. The model is as follow:

$$\frac{d_S}{d_t} = -\beta\frac{SI}{N}$$

$$\frac{d_I}{d_t} = \frac{\beta SI}{N} - \gamma I$$

$$\frac{d_R}{d_t} = \gamma I - \epsilon R$$

$$S + I + R = N$$

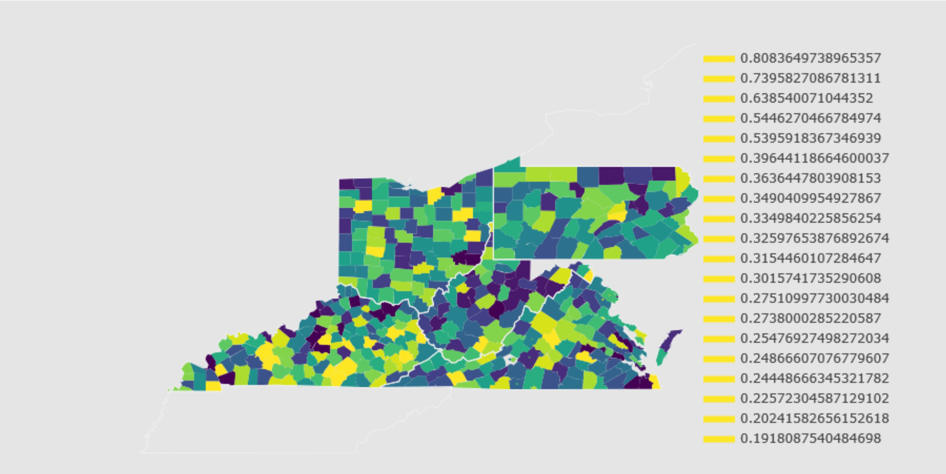


Figure 3: k' Heat Map

In the figure shown above, dark colors denote that the corresponding areas have negative k'; green colors denotes the corresponding areas having k' being close to zero. And yellow means corresponding areas have positive k'. This means drug reports in these areas are not only increasing, but also increasing faster and faster.

4.2 Data fitting with least square

$d_I/d_t = k$ because both of k and d_I/d_t represent the change of drug report over time. Since we have assumed that "Although the probability is very low, it is still possible for a person to detox", which means γ is very small, in this part we treat it as 0 for the simulation of β. Then we can derive the following equation:

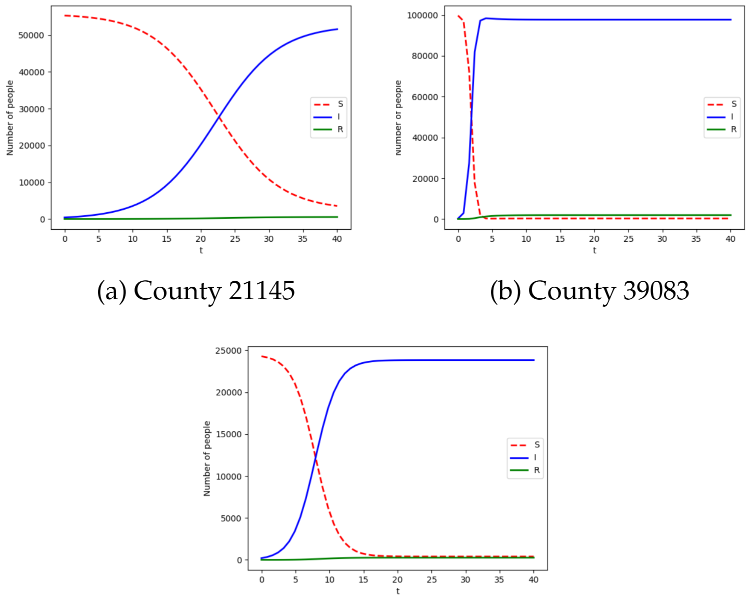
$$k = \frac{d_I}{d_t} = \beta I - \frac{\beta I^2}{N} + c$$

Then we use least square as the loss function to curve fit the β and N for our SIRS model. With the six k for each county in these 7 years, we can have a table of β, N and corresponding R^2 for each county.

Table 2: Example of parameter table

FIPS_Combined	β	N	R^2
21001	7.095	123974	0.872
21003	1.824	50000	0.371
	...	...	
54107	0.949	210416	0.589

With the β and N simulated from k we can derive a function graph for the SIRS model for each county. Below are graph of simulations with three randomly chosen counties. In this graph, we set ϵ, the rate of detoxed people readdicted to drug, is 0.5; set γ, the rate of detoxication, to 0.01 (only 1% percent of people who are addicted to drug can recover).



(a) County 21145

(b) County 39083

(c) County 42009

Figure 4: Red: Non-addicted; Blue: Addicted; Green: Recovery

With the graph, we identify a clear pattern: the rate of change k follows a function like a logistic function. When drug report is either very small or very big, k is small; when drug report is in the development state, k should be big. This pattern fits more than 90% percent of the county. Thus we can use this pattern to identify the stage to which each county belong in this drug crisis. Most of the counties fall into groups differed by characteristics of being at each of these three stages. In turn, we can group the counties by analyzing which stage they are currently at.

4.3 Source and threshold Identification

From the conclusion we derived above, there are three groups that will have k' close to zero:

- Source of drugs. This is because numbers of people being addicted in these counties are nearly saturated.
- Counties having increasing rate at warning threshold. These counties have highest increasing rate and the numbers of addicted people have not yet saturated, which means not much suppression.
- Counties being relatively safe. They have both k and k' at low level.

Thus, our strategy is to first select counties with k' close to zero. This is achieved by calculating the frequency distribution of k' and then selecting the counties having k' that fall in the middle 90% interval. After that, we divide these counties into two groups: one with k at high level and the other with k close to or even less than zero. We set this threshold to be 10% from above.

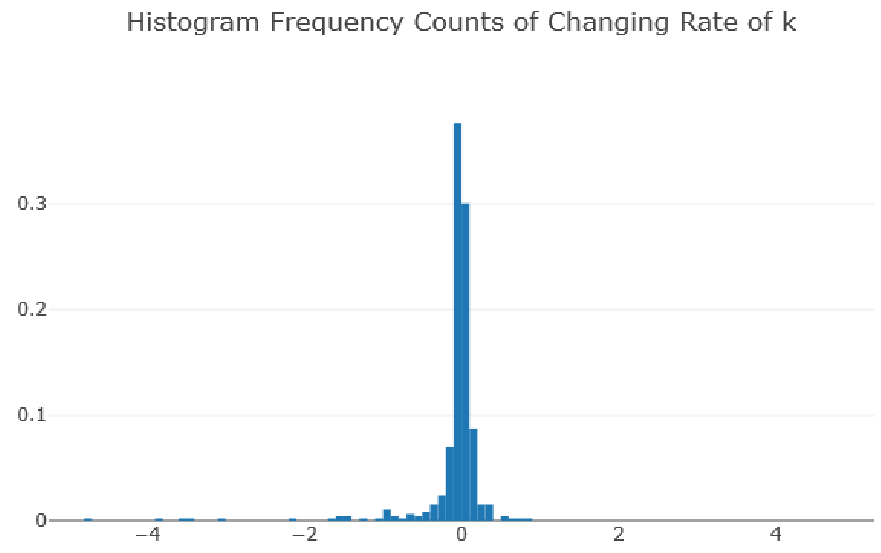
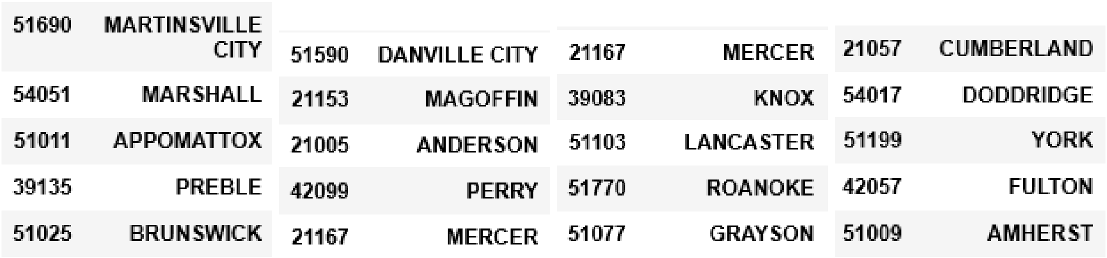


Figure 5: k' Distribution 2016

In order to perform prediction and select threshold, we choose the values of k from data of year 2016. And the lower bound of top 10% is $k = 0.375$. As for selecting source from the lower 90% interval, we sorted the group according to the drug reports with data from year 2010.

After that, we calculated the mean of report in the group, which is $mean = 686$. Counties with number of drug reports higher than the mean will be selected into the source candidate group.



51690	MARTINSVILLE CITY	51590	DANVILLE CITY	21167	MERCER	21057	CUMBERLAND
54051	MARSHALL	21153	MAGOFFIN	39083	KNOX	54017	DODDRIDGE
51011	APPOMATTOX	21005	ANDERSON	51103	LANCASTER	51199	YORK
39135	PREBLE	42099	PERRY	51770	ROANOKE	42057	FULTON
51025	BRUNSWICK	21167	MERCER	51077	GRAYSON	51009	AMHERST

Figure 6: List of Counties in Warning

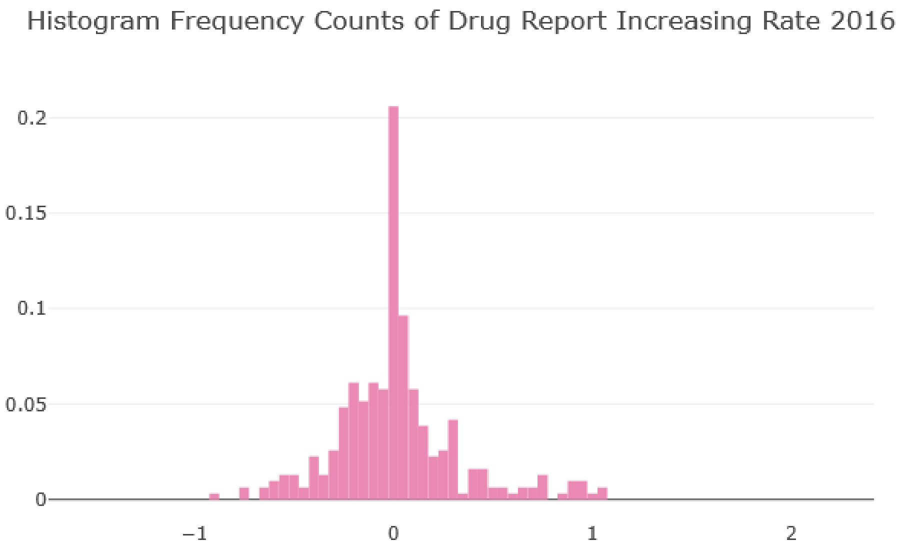


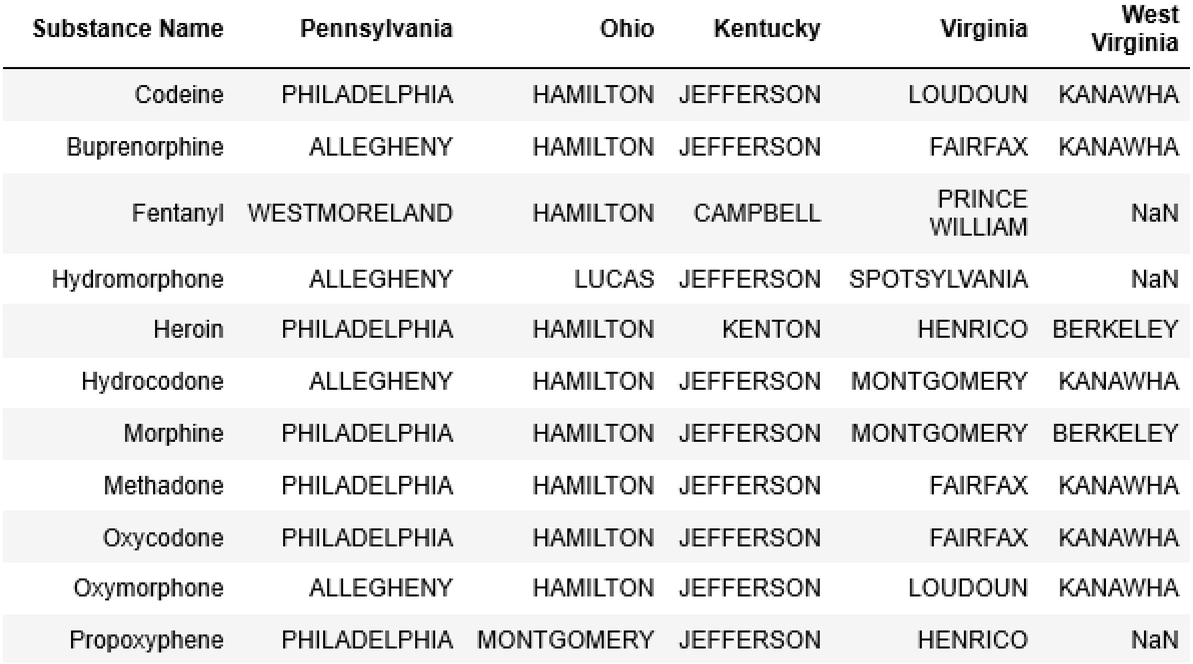
Figure 7: k Distribution 2016

At this point, only 51 counties left. Among these counties, sorting is performed again based on the drug reports of each specific kind of drugs. This strategy is based on the assumption that the sources of all kinds of drugs are more likely to be the source of each specific kind of drugs. Results are shown below.

4.4 Prediction of Safe Counties

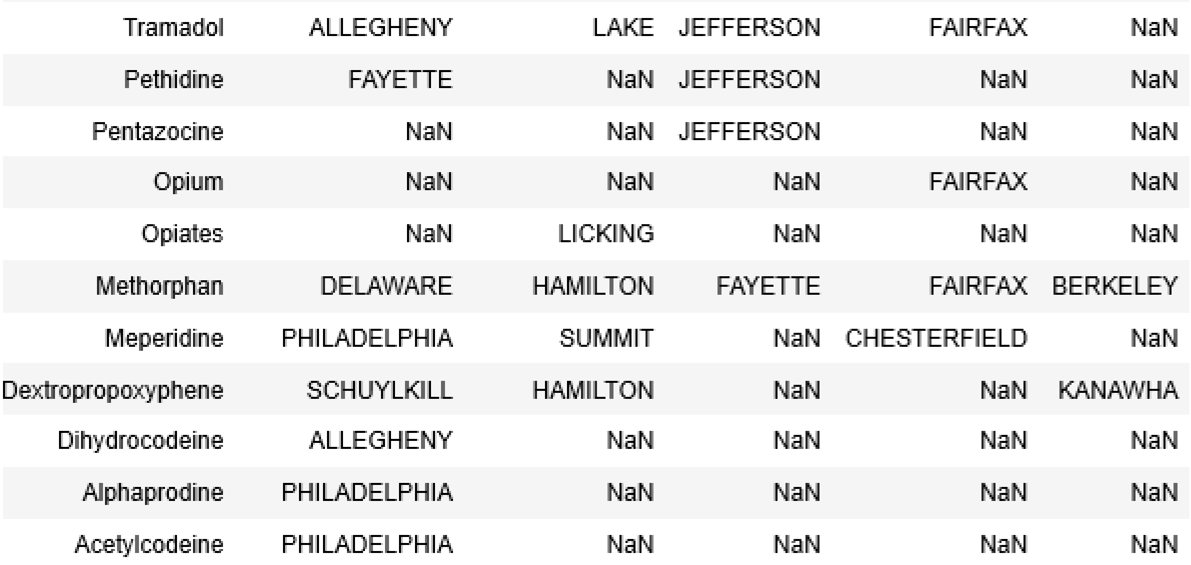
For further prediction, we select the counties in the third group after filtering, which are those currently relatively safe counties. Our strategy for prediction is simple. We divide the difference between current k and threshold with corresponding k' of that county. Then we can get the time it will cost for those counties to reach the threshold. This is because the time interval of six years is relatively small. Thus k' can be treated as a constant. So, we can use the k' for linear approximation. Below are part of the results in ascending order of time interval. The unit of time is year.

Although figures shown are not the full results, which containing more than 200



Substance Name	Pennsylvania	Ohio	Kentucky	Virginia	West Virginia
Codeine	PHILADELPHIA	HAMILTON	JEFFERSON	LOUDOUN	KANAWHA
Buprenorphine	ALLEGHENY	HAMILTON	JEFFERSON	FAIRFAX	KANAWHA
Fentanyl	WESTMORELAND	HAMILTON	CAMPBELL	PRINCE WILLIAM	NaN
Hydromorphone	ALLEGHENY	LUCAS	JEFFERSON	SPOTSYLVANIA	NaN
Heroin	PHILADELPHIA	HAMILTON	KENTON	HENRICO	BERKELEY
Hydrocodone	ALLEGHENY	HAMILTON	JEFFERSON	MONTGOMERY	KANAWHA
Morphine	PHILADELPHIA	HAMILTON	JEFFERSON	MONTGOMERY	BERKELEY
Methadone	PHILADELPHIA	HAMILTON	JEFFERSON	FAIRFAX	KANAWHA
Oxycodone	PHILADELPHIA	HAMILTON	JEFFERSON	FAIRFAX	KANAWHA
Oxymorphone	ALLEGHENY	HAMILTON	JEFFERSON	LOUDOUN	KANAWHA
Propoxyphene	PHILADELPHIA	MONTGOMERY	JEFFERSON	HENRICO	NaN

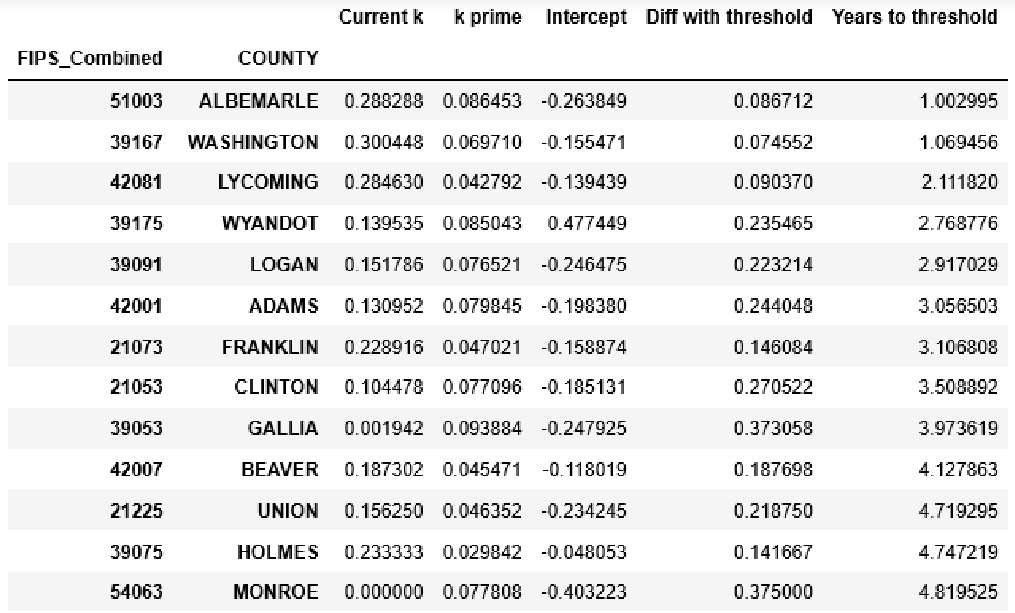
Figure 8: Possible source by drugs part1



Tramadol	ALLEGHENY	LAKE	JEFFERSON	FAIRFAX	NaN
Pethidine	FAYETTE	NaN	JEFFERSON	NaN	NaN
Pentazocine	NaN	NaN	JEFFERSON	NaN	NaN
Opium	NaN	NaN	NaN	FAIRFAX	NaN
Opiates	NaN	LICKING	NaN	NaN	NaN
Methorphan	DELAWARE	HAMILTON	FAYETTE	FAIRFAX	BERKELEY
Meperidine	PHILADELPHIA	SUMMIT	NaN	CHESTERFIELD	NaN
Dextropropoxyphene	SCHUYLKILL	HAMILTON	NaN	NaN	KANAWHA
Dihydrocodeine	ALLEGHENY	NaN	NaN	NaN	NaN
Alphaprodine	PHILADELPHIA	NaN	NaN	NaN	NaN
Acetylcodeine	PHILADELPHIA	NaN	NaN	NaN	NaN

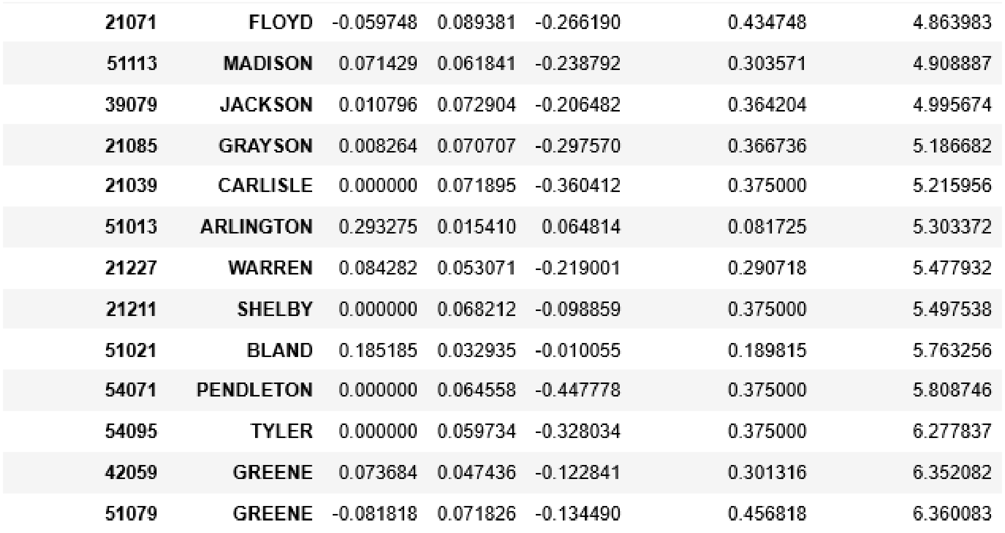
Figure 9: Possible source by drugs part2

counties, they revealed the counties that need to be concerned and the urgency of current situation. We can see that in no more than 6 years, nearly 20 counties will reach the threshold. Though the number of drug reports may not be significant in general, it can still be devastating for those counties themselves because low drug reports are correlated with low local population.



		Current k	k prime	Intercept	Diff with threshold	Years to threshold
FIPS_Combined	COUNTY					
51003	ALBEMARLE	0.288288	0.086453	-0.263849	0.086712	1.002995
39167	WASHINGTON	0.300448	0.069710	-0.155471	0.074552	1.069456
42081	LYCOMING	0.284630	0.042792	-0.139439	0.090370	2.111820
39175	WYANDOT	0.139535	0.085043	0.477449	0.235465	2.768776
39091	LOGAN	0.151786	0.076521	-0.246475	0.223214	2.917029
42001	ADAMS	0.130952	0.079845	-0.198380	0.244048	3.056503
21073	FRANKLIN	0.228916	0.047021	-0.158874	0.146084	3.106808
21053	CLINTON	0.104478	0.077096	-0.185131	0.270522	3.508892
39053	GALLIA	0.001942	0.093884	-0.247925	0.373058	3.973619
42007	BEAVER	0.187302	0.045471	-0.118019	0.187698	4.127863
21225	UNION	0.156250	0.046352	-0.234245	0.218750	4.719295
39075	HOLMES	0.233333	0.029842	-0.048053	0.141667	4.747219
54063	MONROE	0.000000	0.077808	-0.403223	0.375000	4.819525

Figure 10: Prediction of Crisis Development of Safe Counties Part1



21071	FLOYD	-0.059748	0.089381	-0.266190	0.434748	4.863983
51113	MADISON	0.071429	0.061841	-0.238792	0.303571	4.908887
39079	JACKSON	0.010796	0.072904	-0.206482	0.364204	4.995674
21085	GRAYSON	0.008264	0.070707	-0.297570	0.366736	5.186682
21039	CARLISLE	0.000000	0.071895	-0.360412	0.375000	5.215956
51013	ARLINGTON	0.293275	0.015410	0.064814	0.081725	5.303372
21227	WARREN	0.084282	0.053071	-0.219001	0.290718	5.477932
21211	SHELBY	0.000000	0.068212	-0.098859	0.375000	5.497538
51021	BLAND	0.185185	0.032935	-0.010055	0.189815	5.763256
54071	PENDLETON	0.000000	0.064558	-0.447778	0.375000	5.808746
54095	TYLER	0.000000	0.059734	-0.328034	0.375000	6.277837
42059	GREENE	0.073684	0.047436	-0.122841	0.301316	6.352082
51079	GREENE	-0.081818	0.071826	-0.134490	0.456818	6.360083

Figure 11: Prediction of Crisis Development of Safe Counties Part2

4.5 Sensitivity Analysis

In our model of SIRS, we used γ = 0.01 and ϵ = 0.50 for finding pattern. However, the real value of them is unknown due to the limitation of data and resources. So in this part, we should analyze the real influence of ϵ and γ to our SIRS model. Here we pick the county 21145, one of the county from our randomly chosen counties set, to analyze the influence of ϵ and γ to the prediction value.

Table 3: Estimate drug reports of 2017

γ	$\epsilon = 0.50$	$\epsilon = 0.90$
0.01	519.109	519.107
2γ	514.887	514.889

Table 4: Estimate drug reports of 2026

γ	$\epsilon = 0.50$	$\epsilon = 0.90$
0.01	1988.390	1988.700
2γ	1849.34	1849.83

From the table we can find that the value of ϵ has little influence on our model due to the small value of R. γ, on the other hand, have a greater impact on our model. However, even if we double the estimated value of γ, the change of our estimated value of 2026 changes only 7%, which is an acceptable variation. Note that even though the value varies, the overall pattern of the graph never changes. Thus our method of classifying counties to different categories always stand.

5 Part II: Data dimension reduction and Analysis of factors

In order to optimize our SIRS model, we need to identify more factors related to the changing rate of drug report other than merely taking existing drug report into account. After observing the socio-economic data, we found that many factors are related to each other and all the factors can be simplified as several direct features. As a result, we decide to use principal components analysis (PCA) to perform the data dimension reduction. The PCA algorithm makes an orthogonal transformation on a covariance matrix and convert the matrix into a set of values of linearly uncorrelated variables which are called principal components. Each principal component with a high variance possible is under the constraint that it is orthogonal to the preceding components[1].

5.1 Introduction to Principle Component Analysis

This is achieved by calculating the eigen values of covariance matrix. And then use the corresponding eigen vectors as the new base. Since the eigen vectors are orthogonal to each other, the new representation of the data will have very low covariance with each other.

$$Cov(a,b) = \frac{1}{m}\sum_{i=1}^{m} a_i b_i$$

Assume that we have a data set X with m samples and only two original factors a and b, the corresponding covariance matrix will be $\frac{1}{m}XX^T$. Each entry of its diagnal is the variance of a and b. Each of the other entries are the covariance of a and b.

$$\frac{1}{m}XX^T = \begin{pmatrix} Var(a) & Cov(a,b) \\ Cov(a,b) & Var(b) \end{pmatrix}$$

Let the new axis after transformation be P, then coordinates of original data on new axis will be $Y = PX$. The covariance matrix of transformed data D would be

$$\frac{1}{m}YY^T = \frac{1}{m}(PX)(PX)^T = P(\frac{1}{m}XX^T)P^T$$

If we choose eigen vectors as the axis, the corresponding D will be nearly a diagnal matrix because eigen vectors are orthogonal with each other. Also, the corresponding eigen values are the metric of "importance", or the information it retained from original data representation, of this new axis. By selecting smaller number of eigen vectors than the number of original factors, we, in turn, can derive a matrix consisting of the projections of the original factors to new axis.

5.2 Principle Component Analysis with socio-economic data

Based on the theory of PCA, we later conduct dimension reduction on the given socio-economic data sets. First of all, we aggregated the data of the six years into a single data set. This is because **we assume that the same abstraction of raw data into the higher order representation is suitable and proper in any of the seven years.** For example, if the factor of school enrollment and the one of educational experience are highly correlated in year 2010, then they will still be highly correlated in the following year and so on. Thus, we first created a matrix of shape **3257 by 132**, where 3257 are the sample size, which is data of 466 counties in seven years combined. Then a covariance matrix of shape **132 by 132** is created to perform PCA.

After that, eigen values and eigen vectors of covariance matrix is solved and we use the following formula to determine the number of dimension that we will reduce the original data to:

$$Fraction\ of\ Information\ Preserved = \sum_{j=1}^{m} \lambda_j / \sum_{j=1}^{n} \lambda_j$$

where m is the number of eigen vectors or new axis that we wanted to preserve and n is the same as number of original factors. This formula is based on that eigen values reveals the amount of covariance that will be projected to the corresponding axis, which is just the original information that will be preserved. If we choose n eigen vectors then all of the covariance will be preserved. We set the threshold to be 0.99 and there are 19 factors left for later analysis. The coordinates after data transformation will be called **score on the axis.** It is hard to determine the meanings from sheer values of scores before analyzing the values of eigen vectors on each of the 132 entries and corresponding original factor names. Further analysis and back trace will be discussed later.

5.3 Identify factors with Lasso regression

Now we have converted all the data into more predictive models, we can identify the relationship between those principal components and those parameters. In order to figure out the clearer connection between those factors and the changes of number of reports, we decide to use Lasso regression to reduce the dimension of factors further. Based on our higher order representation of factors and rate of change of reports' number, we pick up all the factors that directly contribute to the rate of change of number of reports. Lasso regression performs both variable selection and regularization to improve the accuracy of prediction and interpretability of the statistical model.

$$k = w_1 f_1 + w_2 f_2 + ... + w_{19} f_{19}$$
$$Loss = \sum (k - k_{predicted})^2 + \sum |W|$$

As the formula shown above, the summation of the absolute values of weights are also part of the error function. Thus, the value of weights will be restricted to a small value. In this case, those factors that have bigger of non-zero absolute weights will be selected as more related factors.

After optimizing with Lasso regression model, we can derive a vector of coefficient of each principal factors. Theses coefficients can represent the level of correlation between principal factors and the changing rate of drug report k. Mean values of weights across all counties are used for estimation.

After excluding those with coefficient approximating to zero, we have a list of principal factors that have strong correlation to the changing rate of drug report k, which is as follows.

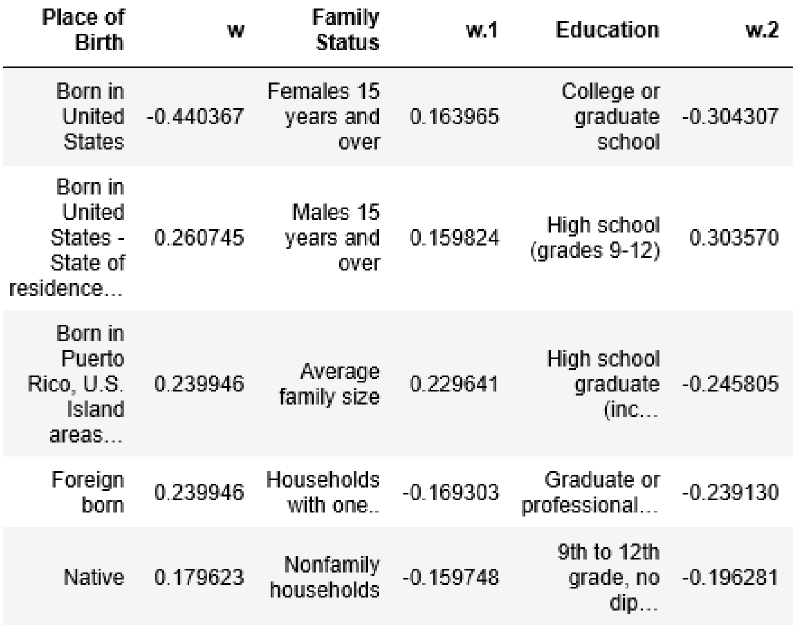
Table 5: None zero weights of Lasso regression

w_1	w_2	w_3	w_4	w_5	w_6
0.00011	0.00011	-0.00013	0.00011	-0.00026	0.00013

5.4 Factors Back-tracing and analysis

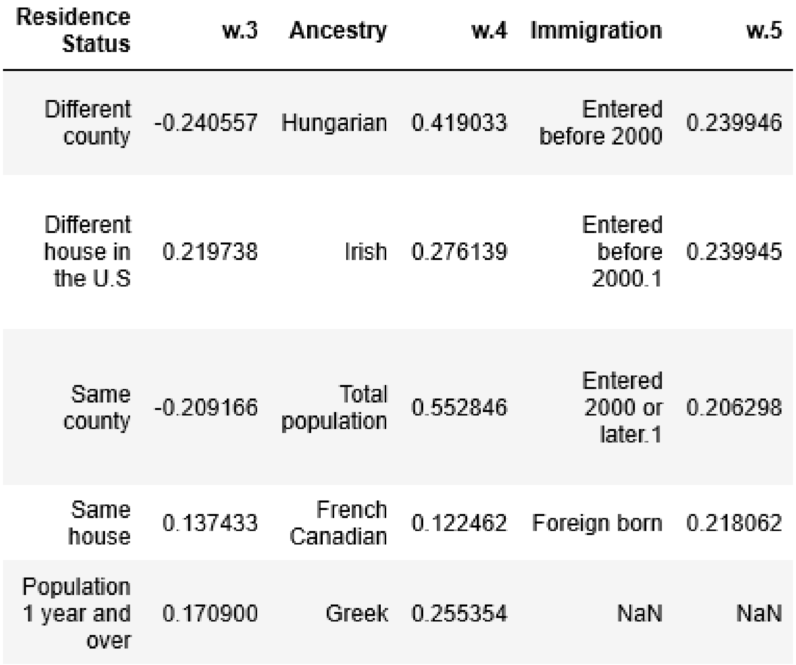
With the six principal factors with the strongest correlation to k, we can back trace to the responsible original remaining 132 factors. Below are some of the corresponding original factors that has relatively large values on the entries of each eigen vectors, or the factors that contribute a lot to the scoring of higher order feature.

This table of factors can tell us the correlation and the level of correlation between original factors and k. For example, based on our model, we find that educational level is negatively related to the changing rate of report drug. This means that if a county mainly consists of people with college level education or higher, it will have a small k. Another example is family situation. Based on our result of Lasso regression, those people who married in a young age or in a unstable family situation are more likely to be addicted to drugs.



Place of Birth	w	Family Status	w.1	Education	w.2
Born in United States	-0.440367	Females 15 years and over	0.163965	College or graduate school	-0.304307
Born in United States - State of residence...	0.260745	Males 15 years and over	0.159824	High school (grades 9-12)	0.303570
Born in Puerto Rico, U.S. Island areas...	0.239946	Average family size	0.229641	High school graduate (inc...	-0.245805
Foreign born	0.239946	Households with one..	-0.169303	Graduate or professional...	-0.239130
Native	0.179623	Nonfamily households	-0.159748	9th to 12th grade, no dip...	-0.196281

Figure 12: Higher Order Feature Part1



Residence Status	w.3	Ancestry	w.4	Immigration	w.5
Different county	-0.240557	Hungarian	0.419033	Entered before 2000	0.239946
Different house in the U.S	0.219738	Irish	0.276139	Entered before 2000.1	0.239945
Same county	-0.209166	Total population	0.552846	Entered 2000 or later.1	0.206298
Same house	0.137433	French Canadian	0.122462	Foreign born	0.218062
Population 1 year and over	0.170900	Greek	0.255354	NaN	NaN

Figure 13: Higher Order Feature Part2

6 Part III Analysis: Model Optimization and Strategy

6.1 Strategy for countering the crisis

Based on the set of factors that are related to k, we can give some advice to the government regarding these factors. We still pick the example of education level. We now know that the changing rate of drug report k is negatively related to average education level in a county, so we suggest the government to increase the expenditure of education. Also the number of veterans is also negatively related to k so government can call for more soldiers if necessary because it can also alleviate this crisis in some extend.

For those factors that are positively related to k, government can establish some policies to erase or at least to ease them. For example, the number of early marriage for both male and female are positively related to k. To erase this factor, governments can increase the penalty for early marriage or increase the earliest marriage age.

There are a lot more factors than these few examples I gave in this part. They are all in the Table we derived from the back trace of our principal component matrix. Governments can regulate the policy based on those factors.

6.2 Model Optimization with least square regression

In the SIRS model we simulate in part 4.2, both β and N are number driven, which means that it is possible for the machine learning algorithm to over fit the data. If it is the case, β and N don't really have any meaning. N is suppose to be all people who can be addicted to drug in a county. However, in some simulation, the N we derive from the least square regression is much larger than the reasonable. For example, N can be as high as several million in some county, while the real population is only several thousand. In the context of the problem, it means that the number of people that can be addicted to drug in a county is a lot more than the real population of this county.

To solve this problem, we use the result of our Lasso regression: six principal factors. They are projections of the number of people of several factor and we have shown their correlations to k with Lasso regression. Instead of using the least square regression to find N, we use these principal factors and their coefficient as N. Now we have the following equation:

$$k = \frac{d_I}{d_t} = \beta I - \frac{\beta I^2}{\alpha N} + c$$

With this equation, we can derive a new set of β and corresponding R^2.

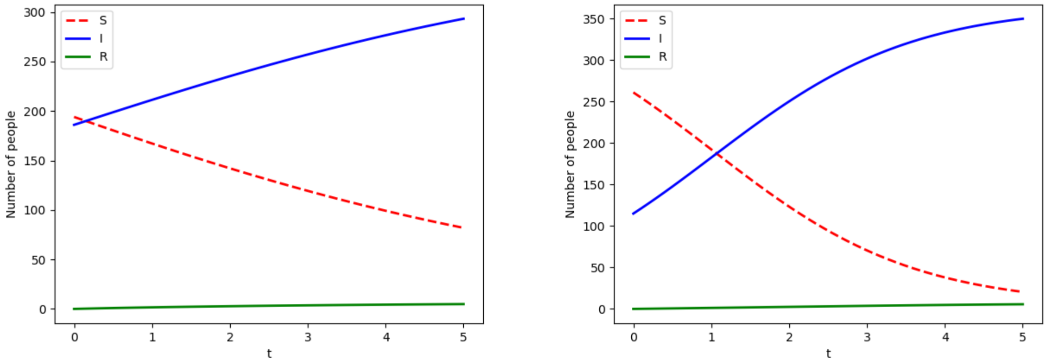
Table 6: Example of new parameter table

FIPS_Combined	β	α	R^2
21001	0.783	-5.731	0.892
21003	0.096	-1.920	0.730
	...	...	
54097	0.440	-0.312	0.701

The αN should be the possible number of people who can be addicted to drug. This αN, different from our N in the first part, is a coefficient times a population score, which has a solid base in reality. Also, the responding R^2 for our model has increased significantly, which proves the effectiveness of this optimization.

6.3 Effectiveness test of strategy

Now we have an optimized model for SIRS. We can test the effectiveness of our strategy based on our model. We again randomly choose two counties, 21001 and 51163. Before any strategy and changing of policy, the predicted number of drug report in five years from 2016 is as follow:

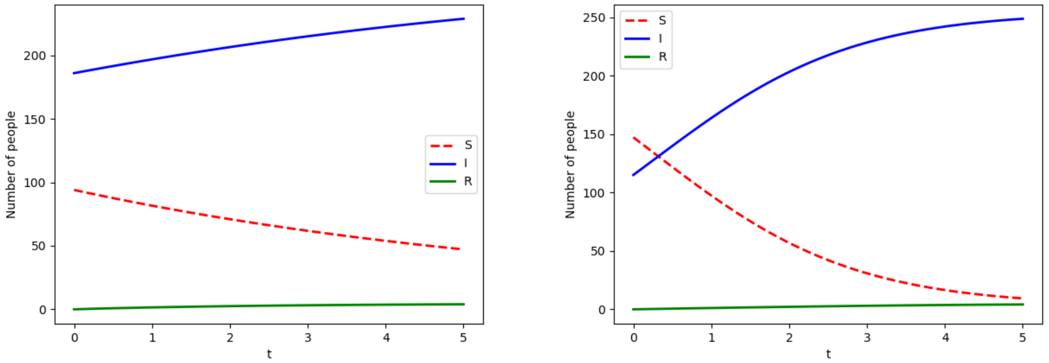


(a) County 51163

(b) County 21001

Figure 14: Simulation of Countering Strategy, Before

After making changes of policies and successfully reducing the correlated factors, assuming N reduces by 30%, we can derive a new set of graphs of these counties, which is shown below.



(a) County 51163

(b) County 21001

Figure 15: Simulation of Countering Strategy, After

From the graph, we can clearly identify the difference before and after the changes. In theory, if we increase the factors that is negatively to k and decrease the factor that is positively to k, the total possible number of people have the potential to be addicted to drug will decrease and the saturation will increase. Thus, the total number of drug report will converge to a smaller value.

7 Strength and Possible Optimization

7.1 Strength

Data processing: When facing big data problem, the accuracy and clearness of data is one of the most important procedure. In this part, we scientifically clean the original data, which consists a lot of missing values, so that we can process the entire problem more accurately and more fluently.

Combination of a lot of factors: In the first part, we use SIRS model to quantify the influence of saturation on the drug. In the third part, we further optimize our model with all possible factors including educational level, household situation, etc.

Accuracy of Model with sample data: After the the optimization in part III, the result of our model has a significantly high R square value when fitting the data, which shows the accuracy of our model.

7.2 Possible Optimization

Missing factor: We identify another factor, distance to source county, also should have a great influence of the changing rate of drug report k. However, with the limitation of time, we are not able to quantify this factor in our model.

Lack of necessary data: The sample data of drug report only has six years, which is a relatively small data set for data analysis. If we can have a larger data set, we can fit our parameter more accurately. Also, if we are allow to access the data for the population of each county, we can better analyze the interaction between the possible addicted group and our principal component matrix.

Simplifying assumption: we assume that once a county is classified as possible source, its possibility of being source of specific opioid is higher. However, it is not always the case. Each specific opioid might have started from different counties so the source cannot be simply identified by comparing the value.

8 Conclusion

In this paper, we perform data analysis based on both the existing drug report and possible related factors in the four hundred counties in the 5 states of United State and was able to draw some conclusions and predictions and prediction based on our model.

Firstly, based on the pattern the data, we decide to use SIRS model and thus able to classify all counties to three categories: starting state, threshold state and saturated state. The saturated states are candidate for all types of opioid and we have a list of possible sources of each types of in five states. For counties in threshold state, we warn the government about the situation and for counties in starting state, we make prediction on when they will turn into a threshold state.

Secondly, we analyze the socio-economic data and perform PCA on these factors. With the principal matrix, we perform Lasso regression to identify the correlation between these factors and changing rate of drug. Then we get a set of coefficients representing the level of correlation and we list sample of possible factors that can affect the number of drug report.

Finally, we give a strategy regarding how to solve this drug crisis and further optimize our SIRS model based on new factors. With the optimized SIRS model, we show the effectiveness of our strategy.

References

[1] Schölkopf B., Smola A., Müller KR. (1997) Kernel principal component analysis. In: Gerstner W., Germond A., Hasler M., Nicoud JD. (eds) Artificial Neural Networks - ICANN'97. ICANN 1997. Lecture Notes in Computer Science, vol 1327. Springer, Berlin, Heidelberg https://doi.org/10.1007/BFb0020217

[2] Bayesian lasso regression. Biometrika, Volume 96, Issue 4, 1 December 2009, Pages 835-845, https://doi.org/10.1093/biomet/asp047

[3] Beretta, E. & Takeuchi, Y. J. Math. Biol. (1995) 33: 250. https://doi.org/10.1007/BF00169563

[4] Shulgin, B., Stone, L. & Agur, Z. Bull. Math. Biol. (1998) 60: 1123 https://doi.org/10.1006/S0092-8240(98)90005-2

[5] Hubbard, R. L., Marsden, M. E., Rachal, J. V., Harwood, H. J., Cavanaugh, E. R., & Ginzburg, H. M. (1989). Drug abuse treatment: A national study of effectiveness. Chapel Hill, NC, US: University of North Carolina Press.

[6] SMART, Reginald G., et al. "United Nations Office on Drugs and Crime." Integrity in the Criminal Justice System www.unodc.org/unodc/en/data-and-analysis/bulletin/bulletin_1971-01-01_2_page004.html

[7] US Census Bureau. "Data." Census Bureau QuickFacts, United States Census Bureau, 4 Sept. 2018 www.census.gov/data/datasets/2017/demo/popest/counties-total.html

[8] NBER, NBER, www.nber.org/data/county-distance-database.html

[9] Nora D. Volkow, MD; Yu-Shin Ding, PhD; et al; Arch Gen Psychiatry. 1995;52(6):456-463. doi:10.1001/archpsyc.1995.03950180042006

[10] "Absorption, Distribution, Metabolism and Excretion Pharmacogenomics of Drugs of Abuse." Breast Cancer Management www.futuremedicine.com/doi/abs/10.2217/pgs.10.171