Team Control Number
1918240

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

Problem Chosen
C

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

2019
MCM/ICM
Summary Sheet

Opioid Drug Spread Patterns and Its Causes for Government Strategies Making Based on Data Insight

Summary

After performing data analysis and constructing models, we finally identify the spread routes and features of the reported synthetic opioid and heroin incidents, therefore make some reasonable predictions. Further exploration of the triggers for the opioid crisis despite the known danger is also conducted so that we can enact effective (tested by the model) targeted strategies to counter the crisis.

First, we construct the Pathway Model based on Clustering(PMC) and Drug-Invasion & Grey Precaution Model(DIM+GPM) to identify the inter-state and inner-state spread patterns respectively. To discover the path law of inter-state drug spread, we first sort out the year when each drug appeared in each state, based on which states are sorted. Then, six path rules are clustered through hierarchical clustering, verified by K-means clustering and displayed with 2 visualization methods. As to patterns within states we construct the DIM+GPM model in which DIM is more powerful in explanation of drugs with increasing proliferation while GPM model has better prediction capacity on the condition that drug spread, complementary with DIM to some extent. By applying the model, we identify the year when each drug outbroke or will outbreak within each states and predict the spread conditions in next 10 years as reference for governmental decision makers.

Next, to figure out fundamental factors, after data preprocessing, we construct the XGBoost+LR with PCA model(XLP) in which we first accurately and objectively filter variables with XGBoost, based on which linear regression model is constructed to explore their influence. Besides, we also use principal component analysis(PCA) to categorize the variables and finally find out that population size, complexity of household structure and external population are the key contributors.

Then, we give out possible strategies according to the key factors we find above which are residence permit points system and divorce penalty policy. To test the effectiveness of these strategies, we construct the polynomial regression as modification of XLP and visualize it through surface and level plot and identify parameter bounds that success depends on with example.

It is worth mentioning that after the model test and sensitivity analysis, no matter how precise the DIM+GPM model is, its estimation and policy suggestions are still reasonable, justifying the robustness of the model.

KEYWORDS: Clustering, XGBoost, PCA, Regression, Drug Invasion and Grey Precaution Model, Sensitivity Analysis

Contents

MEMO

From: Team 1918240, MCM 2019
To: The group of Governors
Date: January 27,2019
Subject: Current situations and possible strategies of opioid crisis

Dear governors, we are honored to inform you our achievements after performing data analysis and modeling.

Fist, we provide you the inter-state and inner-state spread pathways respectively with visualized maps and line graphs obtained through our PMC model. You can check if with interest.

Then, we introduce the current situations of the opioid crisis. Among the 5 states(Ohio, Pennsylvania, Kentucky, Virginia and West Virginia), Ohio and Pennsylvania suffer from the most severe opioid drug misuse. And Heroin which used to be the dominant drug used according to the reported incidents is gradually substituted by other synthetic opioids in all the 5 states except West Virginia.

By applying GIM+GPM model, we find that:

- Fentanyl should be the key control subject in its possible sources Hamilton, Montgomery and Allegheny respectively in Ohio and Pennsylvania.
- In Kentucky, heroin should be controlled in Jefferson mostly.
- In Virginia and West Virginia, buprenorphine should be focused on in its possible sources Logan, Kanawha and Wise.

Some of our suggestions are offered:

- For those counties where outbreak has already happened:
  - Set up inspection teams to remove drug trade den, supplemented with necessary advocacy on drug boycotts.
- For those counties where outbreak has not yet happened, focus more on precaution measures such as:
  - Strictly monitor the drug trading channels and production departments of those counties that may be the sources of spread
  - Reinforce science education on drug misuse to improve people's knowledge about drug harm, thus making it possible to prevent the tendency of drug transmission before outbreak.

Besides, we figure out that population size, complexity of household structure and external population are the key contributors through our XLP model and accordingly give out possible strategies which you can refer to:

- Design and carry out residence permit points system to control migrant population born abroad to US.
- Conduct divorce penalty policy and strengthen marriage education to improve the complexity of family construction.

Also, you can easily test the effectiveness of the policies you enact with our model since we modify it with a polynomial regression and data visualization.

I. INTRODUCTION

A. Research Background

Opioid misuse has been exacerbated and spread quickly in the United States in recent years. Drug overdose deaths quadrupled from 2000 to 2014, even exceeding automobile crashes as a cause of death since 2014. If that is not troubling enough, the National Survey on Drug Use and Health estimated that more than 10 million people in the United States used prescription opioids for nonmedical use in 2014.

Overreliance on opioid medications is emblematic of the pursuit for quick, simplistic answers to complex physical and mental health needs. In an analogous way, simplistic measures to cut access to opioids negatively offer illusory solutions to this multidimensional societal challenge.

This worsening of drug misuse has drawn increasing attention at the national and state levels. At the federal level, the US Department of Health and Human Services and the Office of National Drug Control Policy have undertaken various initiatives to address the hard nut to crack.

B. Problem Restatement

Given the NFLIS data provided, our team work out 3 mathematical models, which are Pathway Model based on Clustering(PMC), Drug-Invasion Model (DIM) and Grey Precaution Model(GPM).They respectively explore the inter-state and inner-state spread patterns. Additionally, we construct a series of variables to depict the characteristics of the reported synthetic opioid and heroin incidents over the time period from 2010 to 2017.

With regard to the possible locations where specific opioid use might have started, we assume that the earliest state to adopt the specific drugs and the county with most and simultaneously earliest reported incidents are the sources respectively at the state and county level.

To make our models more practical, we construct a ratio of spreading rate to purifying rate as the drug identification threshold level to help us better predict and judge where and when the patterns and characteristics will occur in the future.

Along with drug supply as a direct factor, we posit that the crisis is fundamentally fueled by economic and social upheaval, its etiology closely linked to the role of opioids as a refuge from physical and psychological trauma. So we use the U.S. Census socio-economic data provided to figure out the fundamental factors and add them into the model for modification.

Finally, based on all of our research results, we try to make some suggestions for governments and find out a possible and effective strategy for countering the opioid crisis.

II. NOTATIONS & HYPOTHESES

A. Notations

Since we altogether construct 4 models, the specific notation is listed in each part.

B. Hypotheses

1. Common Hypothesis

- We hold the assumption that the drugs which have existed in all the 5 states only proliferate within the state, which is the object of the intercounty study; while the interstate one only considers the drug use

which underwent a growth out of nothing. That is to say, we divide the whole sample into two parts to study separately.

2. Hypotheses for PMC

- The conditions in which the figure of specific drug in specific county drops from non-zero to zero are considered meaningless and thus neglected.
- The year of occurrence is recorded as 2010 for drugs that has already existed before 2010 and as 2017 for those which has not appeared until 2017 in a similar way. This approximation does not affect the final ranking result.

3. Hypotheses for DIM

- We believe that the use of opioids and thus addiction is like infection with infectious diseases, so we adopt the epidemic model for reference and follow its signs.
- During the circulation of drugs, each county is allowed to use every drug listed in the data set, that is to say, the number(N) of counties with drug reports is unchanged. And we don't consider the conditions that a drug cannot be circulated due to policy or economic reasons. Therefore all the counties are divided into 2 categories: unreported ones (similar to infectious diseases) and already reported ones. The proportions of the 2 types in the total number of counties in the year t is recorded as s(t) and i(t), respectively.
- We record $\hat{\lambda}$ as the average number of counties that each reported counties can effectively contact each year. When the reported county has contact with the unreported ones, the latter will transform into the former one.

4. Hypotheses for DIM

- Unlike the previous Drug Invasion Model(DIM), we believe that there are external factors such as policies or economic elements that prevent drugs from circulation. So the situation is permitted to exist that invaded counties are purified in the next year.

  Since GPM is the modification of DIM, the rest of the assumptions are the same as DIM.

III. DESCRIPTIVE STATISTICS

We first sum up the reported synthetic opioid and heroin incidents respectively of each state in each year, trying to figure out some spread rules and characteristics. The results are shown in the line graphs below:

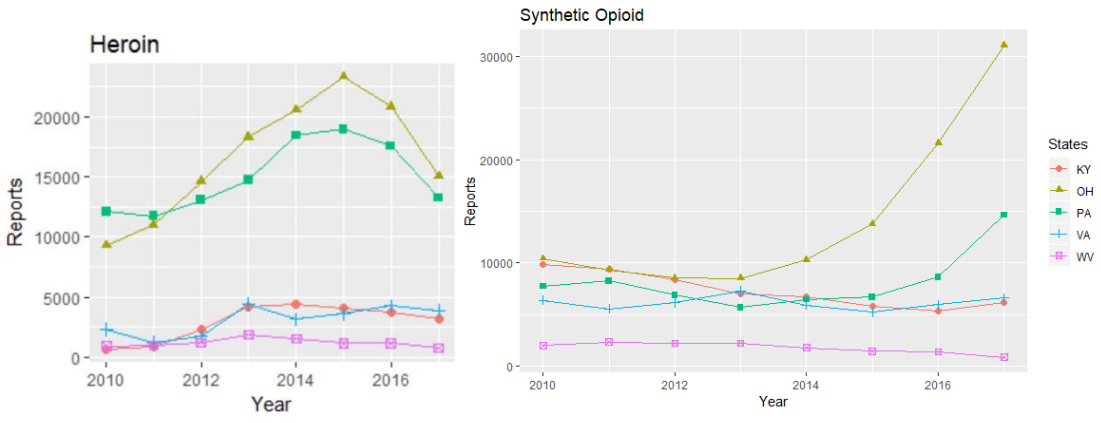

Figure3_1: Descriptive Statistics of Reported Synthetic Opioid and Heroin

From the charts, it is apparent to see that in the first several years of the research period, Heroin was a dominant drug used in Ohio and Pennsylvania in terms of the reported incidents. However, since 2015, Heroin was largely substituted by other synthetic opioids in all the 5 states except West Virginia which can be learn from the downward and upward trend of reported incidents of heroin and synthetic opioids respectively.

IV. SPREAD PATTERNS & ROUTES

To figure out the spread patterns, we construct 3 models to explore the inter-state and inner-state pathways separately which are Pathway Model based on Clustering(PMC), Drug-Invasion Model (DIM) and Grey Precaution Model(GPM).Among them, PMC is applicable to inter-state spread, while PMC and DIM are mainly used to identify the inner-state(inter-county) spread route.

The HYPOTHESIS to which all our models confirm to is that the drugs which have existed in all the 5 states only proliferate within the state, which is the object of the intercounty study; while the interstate one only considers the drug use which underwent a growth out of nothing. That is to say, we divide the whole sample into two parts to study separately.

A. Inter-state Patterns with PMC

SUBJECT: the initial year for each drug to appear in each states.

PURPOSE: Our objective is to find out the possible pathways for specific drugs to spread between the five states and try to conclude some rules applicable to a variety of drugs.

HYPOTHESES:

- The conditions in which the figure of specific drug in specific county drops from non-zero to zero are considered meaningless and thus neglected.
- The year of occurrence is recorded as 2010 for drugs that has already existed before 2010 and as 2017 for those which has not appeared until 2017 in a similar way. This approximation does not affect the final ranking result.

DATA PRE-PROCESSING

Since we only consider the sequence in which each drug appears in each state, so we make some adjustion so that the year of occurrence is recorded as 2010 for drugs that has already existed before 2010 and as 2017 for those which has not appeared until 2017 in a similar way. This approximation doesn't affect the final ranking result. The final result is partly displayed in Table 1.

Table 4_1: The Occurrence Sequence of Drugs

Note: The year of occurrence is recorded as 2010 for drugs that has already existed before 2010 and as 2017 for those which has not appeared until 2017 in a similar way.

	state	Codeine	Dihydrocodeine	Opiates	Opium	Oxycodone	p-Fluorobutyryl fentanyl	Pethidine	Phenyl fentanyl	Thebaine	Tramadol	U-47700	U-48800	U-49900	U-51754
1	VA	2010	2012	2010	2010	2010	2017	2011	2010	2010	2010	2016	2010	2010	2010
2	OH	2010	2014	2010	2013	2010	2015	2011	2017	2015	2010	2016	2017	2017	2017
3	PA	2010	2010	2015	2011	2010	2016	2010	2017	2011	2010	2016	2017	2010	2010
4	WV	2010	2010	2010	2010	2010	2010	2010	2017	2010	2012	2016	2010	2010	2010
5	KY	2010	2011	2010	2017	2010	2016	2010	2010	2010	2010	2016	2010	2010	2010

METHODS

After the data pre-processing, we formally conduct the study by using hierarchical clustering and k-means clustering:

1. Hierarchical Clustering

Firstly, the distance is calculated by using Ward Linkage Method. Therefore, we figure out the classification result which divides the total 70drugs into 6 categories, which is shown in the picture below.

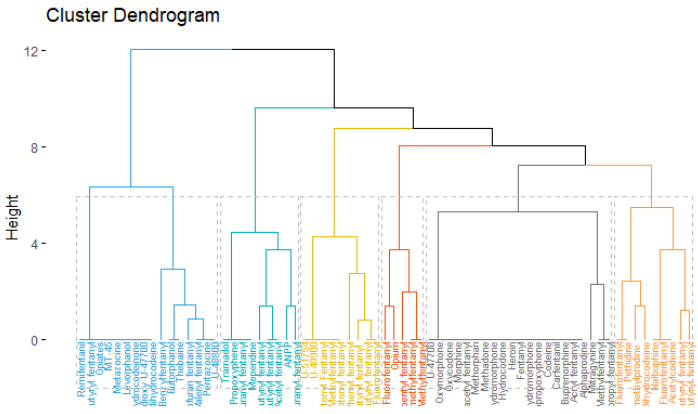

Figure 4_1: Clustering Results of Hierarchical Clustering

2. K-Means Clustering

According to the result in the hierarchal clustering, we then use k_Means (k=6) in advance. And the clustering results are similar to the hierarchical clustering, confirming our classification.(Figure 4_2)

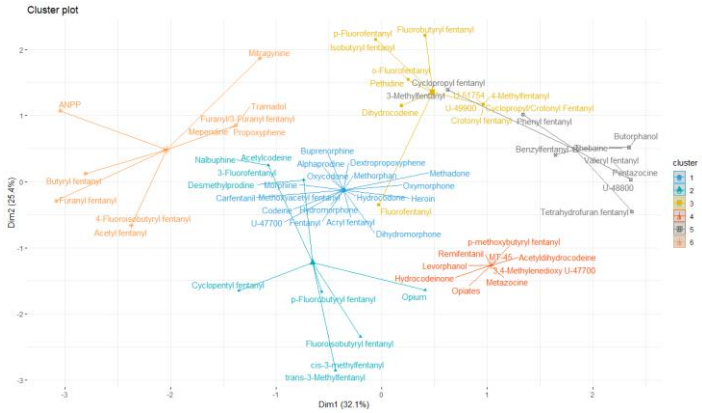

Figure 4_2: Clustering Results of K-Means(k=6)

Then we visualize the path characteristics of each clustering result:

The mean sequences of occurrence of 6 types in each state are firstly calculated as the representative order of this class. The county where each type of drug first appears is the possible source. of the spread. The result is displayed in Figure 4.

The abscissa is the 5 states and the ordinate is the mean of rank. The smaller the rank is, the earlier the emergence of such drugs. The grouping variable is a category result of 6 k-means.

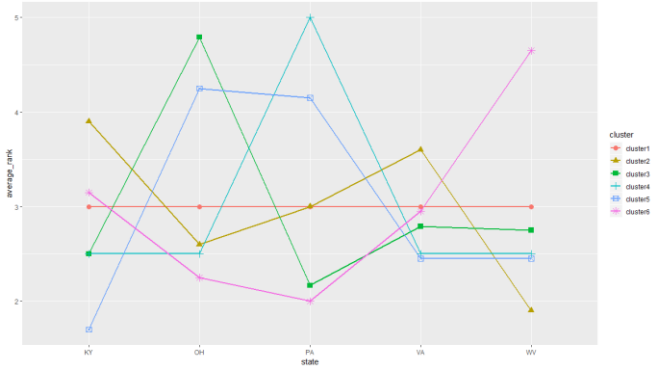

Figure 4_3:Visualized Pathways for 6 Types of Drugs

The results are finally converted to a map for more intuitive display.

Figure 4_4 Spread Routes marked on Map

B. Inner-state Patterns with DIM

The circulation of drugs, especially addictive drugs like opioids, can be interpreted as the infection of addiction symptoms among counties. Each county is like a patient who can be infected. Therefore, we take the infectious disease model for reference and improve it into our drug invasion model(DIM).

SUBJECTIVE

We hold the assumption that the drug-reported counties are those invaded by the specific drug and the drug-unreported ones are uninvaded. Therefore, we focus on the former to set up the DIM model.

PURPOSE

We tried to explore the specific pathway and spread extent of each drug within each state using DIM.

HYPOTHESIS

- We believe that the use of opioids and thus addiction is like infection with infectious diseases, so we adopt the epidemic model for reference and follow its signs.
- During the circulation of drugs, each county is allowed to use every drug listed in the data set, that is to say, the number(N) of counties with drug reports is unchanged. And we don't consider the conditions that a drug cannot be circulated due to policy or economic reasons. Therefore all the counties are divided into 2 categories: uninvaded ones (similar to infectious diseases) and already invaded ones. The proportions of the 2 types in the total number of counties in the year t is recorded as s(t) and i(t), respectively.
- We record $\hat{\lambda}$ as the average number of counties that each reported counties can effectively contact each year. When the reported county has contact with the unreported ones, the latter will transform into the former one.

MODEL

Notation	Explanation
t	We define year 2010 is the year 0,so $t = year - 2010$.
$I(t)$	the class of counties that are invaded by this substance in year t
$U(t)$	the class of counties that are uninvaded by this substance in year t
N	total counties of one state $N = U(t) + I(t)$ for each t
$i(t)$	the percentage of invaded counties in one state in t $i(t) = I(t)/N$
$u(t)$	the percentage of uninvaded counties in one state in t $u(t) = U(t)/N$
λ	the average number of counties that one invaded county can pollute per year
i_0	$i_0 = i(0)$
t_m	the time when the proliferation extent of this substance reaches its zenith in one state

Based on our hypotheses, each invaded county can spread the usage of the specific drug $\lambda s(t)$ counties in year t and the number of invaded counties is $Ni(t)$. So there are $\lambda Ns(t)i(t)$ uninvaded counties are later invaded. In this way, the increase rate of Ni is λsi. Therefore, we have

$$\frac{di}{dt} = \lambda si$$

Because

$$s(t) + i(t) = 1$$

we have

$$\frac{di}{dt} = \lambda i(1 - i), \quad i(0) = i_0 \qquad (1)$$

which is a logistic model, and its solution is

$$i(t) = \frac{1}{1 + (\frac{1}{i_0} - 1)e^{-\lambda t}} \qquad (2)$$

We conduct the regression using (2) and the given data set, and therefore obtain estimation $\hat{\lambda}$of the parameter λ and the sample regression model(SRF). Then we predict the number of counties each drug will invade in each state over the next 10 years on the basis of the SRF.

When $\frac{di}{dt}$ reaches its maximum, the increase of invaded counties will be the fastest, government should take actions in time. From (1), we can easily find out when $i(t) = \frac{1}{2}$, $\frac{di}{dt}$ reaches its maximum.

In this way, we choose $i(t) = \frac{1}{2}$ as the threshold where U.S government should concern. From (5) we can calculate the time when $i(t) = \frac{1}{2}$, which is

$$t_m = \hat{\lambda}^{-1} ln(\frac{1}{i_0} - 1)$$

PARAMETER ESTIMATION & HYPOTHESIS TEST

Although the infectious disease model has been widely used in the fields of infectious diseases and information dissemination, we found in the literature review that most of the existing literatures belong to predictive analysis with given $\hat{\lambda}$. Scarce literature figure out $\hat{\lambda}$ through estimation from the existing data set. Based on the particularity of our problem, there is no existing data to provide us with $\hat{\lambda}$ of each drug in each state. Additionally, considering our research is not a purely theoretical derivation but with a realistic background and problem which remains to be solved, it is inappropriate to set $\hat{\lambda}$ directly without basis. So we estimate $\hat{\lambda}$ through OLS using the given 8-year data from 2010 to 2017.

We can find that in each state, the absolute value of $\hat{\lambda}$ for some drugs is very small. So we go back to re-observe the data and find that the number of drug-reported counties in each state has not fluctuated in 8 years. It can be interpreted as no spread of the drug within the state. Therefore, these drugs don't need too much supervision temporarily. However, some drugs do have really large $\hat{\lambda}$, which indicates that the number of drug-reported counties in each state has increased in the past eight years. These drugs are bound to spread widely within the state and even between states if left uncontrolled, and the consequences could be disastrous.

In order to distinguish essential drugs requiring great attention, we conducted a hypothesis test， in which the null hypothesis is: H0: $\lambda=0$; H1: $\lambda\neq0$. By checking the F-statistic and its p value, we finally pick out the significant $\hat{\lambda}$.(listed in the table below)

OUTCOMES & PREDICTIONS

We divide all the counties into 2 parts according to the plus-minus sign of $\hat{\lambda}$, of which the counties with positive $\hat{\lambda}$ are listed in the first table(Table 4_2(a)) while with negative $\hat{\lambda}$ are listed in the second one (Table 4_2(b)).

Table 4_2(a) depicts the locations where specific opioid use (the 4th column) might have started in each of the five states (the first 3 columns). The larger the $\hat{\lambda}$, the greater the proliferation and spread of the specific drug within the county. Besides, by calculating each drug threshold level of each drug in each county, we figure out the exact year when the specific drug misuse reached or will reach its zenith, which is listed in the "tm" column.

Table 4_2(b) displays the states with negative $\hat{\lambda}$ for specific drug, which indicates that the spread will be suppressed to some extent and the larger absolute value of $\hat{\lambda}$, the greater the extent.

The five pictures(Figure4_5)are the sample graphs for specific drug which implies the trajectory of change and possible future trend in each state.

State	Source	FIPS_Combined	SubstanceName	λ	tm
KY	HARLAN	21095	Buprenorphine	0.18	2010
	FAYETTE	21067			
	JEFFERSON	21111			
KY	KENTON	21117	Fentanyl	0.43	2016
KY	JEFFERSON	21111	Heroin	0.32	2013
	HAMILTON	39061			
OH	MONTGOMERY	39113	Fentanyl	0.65	2012
	FRANKLIN	39049			
	LAKE	39085			
OH	MAHONING	39099	Tramadol	0.51	2014
	PHILADELPHIA	42101			
PA	ALLEGHENY	42003	Fentanyl	0.47	2011
	TAZEWELL	51185			
VA	WISE	51195	Buprenorphine	0.12	2013
	FAIRFAX	51059			
VA	HENRICO	51087	Fentanyl	0.33	2015
	FAIRFAX	51059			
VA	HENRICO	51087	Heroin	0.17	2009
	LOGAN	54045			
WV	KANAWHA	54039	Buprenorphine	0.07	2014
	WOOD	54107			
WV	CABELL	54011	Fentanyl	0.31	2019

Table 4_2(a) Outcomes with positive $\hat{\lambda}$

State	SubstanceName	λ
KY	Hydrocodone	-0.24
KY	Methadone	-0.21
KY	Morphine	-0.1
KY	Oxycodone	-0.16
KY	Propoxyphene	-0.31
OH	Methadone	-0.12
OH	Oxymorphone	-0.14
OH	Propoxyphene	-0.40
PA	Codeine	-0.06
PA	Hydrocodone	-0.30
PA	Hydromorphone	-0.09
PA	Methadone	-0.22
PA	Morphine	-0.15
PA	Oxycodone	-0.16
PA	Propoxyphene	-0.42
VA	Meperidine	-0.18
VA	Methadone	-0.16
WV	Hydrocodone	-0.26
WV	Methadone	-0.19
WV	Morphine	-0.27
WV	Oxycodone	-0.19

Table 4_2(b) Outcomes with negative $\hat{\lambda}$,

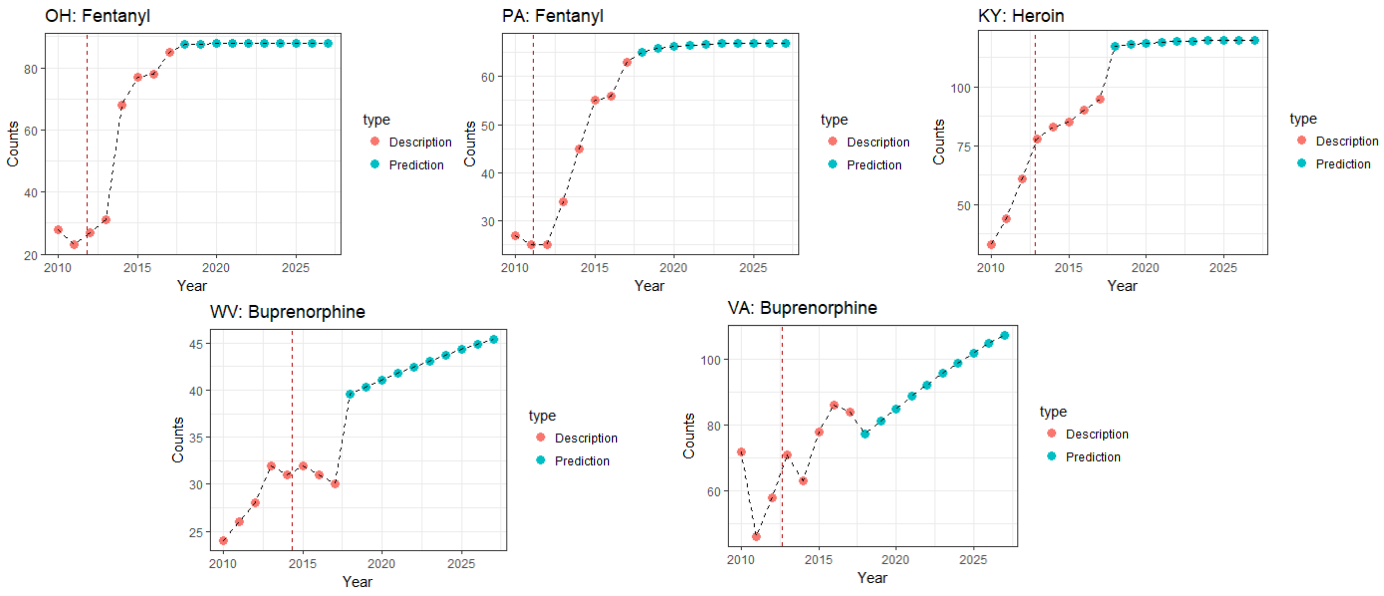

Figure4_5: Sample Graph of Trajectory and Trend with Positive $\hat{\lambda}$

Through the calculation of tm (critical value) when i=1/2, the outbreak of specific drug has occurred in several states, while some have not yet occurred. For those counties where outbreak has already happened, it's necessary to set up inspection teams to remove drug trade den, supplemented with necessary advocacy on drug boycotts. For the latter states, governments should focus more on precaution measures such as strict monitor over the drug trading channels and production departments of those counties that may be the sources of spread and science education on drug misuse to improve people's knowledge about drug harm, thus making it possible to prevent the tendency of drug transmission before outbreak.

MODIFICATION

While applying the DIM model, we find such situation exists that the estimated $\hat{\lambda}$ is negative, which implies that the number of drug-reported counties(invaded countries) in each state is decreasing in 8 years. We can make the explanation that these counties are purified and $\lambda\hat{}$ turns into the purifying rate on this

condition. There are 2 possible reasons for the appearance of purifying rate: the first is that there exists a substitute for this drug and the other may b external factors that inhibit the spread of the drug, which is slightly far-fetched under the DIM assumption that policy and economic factors are not considered. For modification, we use the gray model to predict the diminishing number of invasive counties and we call it Grey Precaution Model(GPM)

C. Inner-state Patterns with GPM

Since GPM is the modification of DIM, the purpose remains the same, while there are slight differences in subjective and premises.

SUBJECTIVE

In GPM, the target we mainly focus on is the drugs that have a downward trend in the number of invasive counties and based on which we establish the new model.

HYPOTHESIS

- Unlike the previous Drug Invasion Model(DIM), we believe that there are external factors such as policies or economic elements that prevent drugs from circulation. So the situation is permitted to exist that invaded counties are purified in the next year.

The rest of the assumptions are the same as DIM.

MODEL

Notation	Explanation
t	we define year 2010 is the year 0,so $t = year - 2010,\ t = 0,1,2 \ldots$
n	The number of years of the original sequence, in our case $n = 8$
$I(t)$	the class of counties that are invaded by this substance in year t
$I^{(0)}$	original invaded county numbers sequence $I^{(0)} = (I(0), I(1), \ldots, I(n))$
$I^{(1)}(t)$	elements of accumulated sequence
$I^{(1)}$	accumulated sequence $I^{(1)} = (I^{(1)}(0), I^{(1)}(1), \ldots, I^{(1)}(n))$
$\hat{I}^{(0)}(t)$	simulated(or predicted) $I^{(0)}(t)$
$\hat{I}^{(0)}$	simulated(or predicted) sequence of original sequence $I^{(0)}$
$\hat{I}^{(1)}(t)$	simulated(or predicted) $I^{(1)}(t)$
$\hat{I}^{(1)}$	simulated(or predicted) sequence of original sequence $I^{(1)}$
a	development coefficient
b	grey input

1. Data Accumulation

Our original sequence is

$$I^{(0)} = (I(0), I(1), \ldots, I(n-1))$$

Then accumulate the original sequence we have the accumulated sequence $I^{(1)}$:

$$I^{(1)}(t) = \sum_{k=0}^{t} I\,(k),\ t = 0,1,\dots,n-1$$

Define GM(1,1) grey differential equation

$$\frac{dI^{(1)}}{dt} + aI^{(1)} = b$$

2. Model Solution

$$I^{(1)}(t) - \frac{b}{a} = (I^{(0)}(t) - \frac{b}{a})e^{-at},\ t = 0,1,2,3\dots$$

$$\hat{I}^{(1)}(t+1) - \frac{b}{a} = (I^{(0)}(1) - \frac{b}{a})e^{-at} + \frac{b}{a}$$

$$\hat{I}^{(0)}(t+1) = \hat{I}^{(1)}(t+1) - \hat{I}^{(1)}(t),\ t = 0,1,2,\dots,n-1$$

3. Model Test

The original sequence is

$$I^{(0)} = (I(0), I(1), \dots, I(n-1))$$

And using GM(1,1) we hae the simulated sequence which is

$$\hat{I}^{(0)} = (\hat{I}\,(0), \hat{I}\,(1), \dots, \hat{I}\,(n-1))$$

the residual error sequence is

$$\hat{\epsilon}^{(0)} = (\hat{\epsilon}\,(0), \hat{\epsilon}\,(1), \dots, \hat{\epsilon}\,(n-1))$$

$$\bar{I} = \frac{1}{n}\sum_{t=0}^{n-1} I\,(t)$$

$$S_1^2 = \frac{1}{n}\sum_{t=0}^{n-1} (\,I(t) - \bar{I})$$

$$\bar{\epsilon} = \frac{1}{n}\sum_{t=0}^{n-1} \epsilon_t$$

We use small error probability P test method to test the robustness and accuracy of the model.

$$p = P\{|\ \epsilon_t - \bar{\epsilon}\ | < 0.6745S_1\}$$

The larger value p is, the better the model can fit.

4. Prediction

We can use our model to simulate the original sequence and predict the invaded county numbers in the future, and the simulation and prediction sequence is:

$$\hat{I}^{(0)}(t) = (\hat{I}(0), \hat{I}(1), \dots, \hat{I}(n-1), \hat{I}(n), \hat{I}(n+1), \dots)$$

OUTCOMEs & PREDICTIONs

The five pictures(Figure 4_6)are the sample graphs for specific drug which implies the trajectory of change and possible future trend in each state.

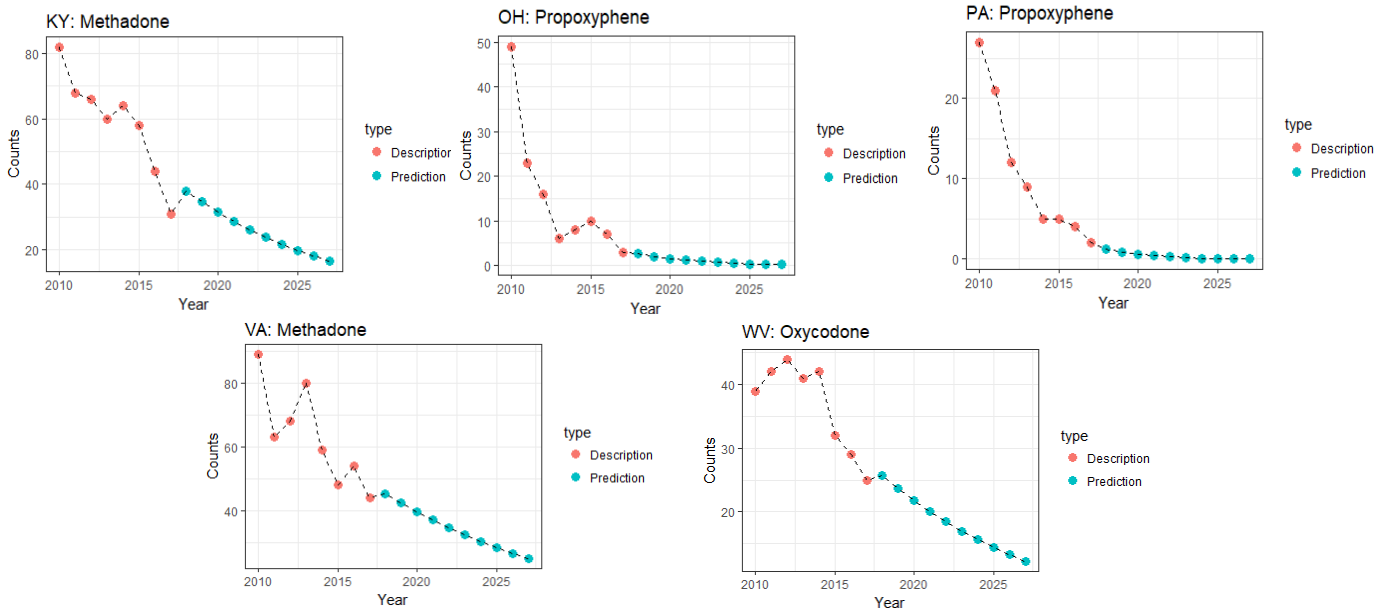

Figure4_6: Sample Graph of Trajectory and Trend with Negative $\hat{\lambda}$

V. EXPLORATION FOR TRIGGERS

There are intuitive causal connections between drug abuse and social and economic factors such as population size, family upbringing etc . So we establish another model called XGBoost+LR with PCA(XLP) using the U.S. Census socio-economic data provided to figure out the fundamental factors.

Considering the importance of each county varies and some essential social-economic factors should be taken into account, we construct a new dependent variable and add key factors into the model along with the trend variable for modification.

PURPOSE:

Our target to construct the model is to find out the fundamental influential factors which make opioid use get to its current level despite its known danger.

SUBJECT:

The dependent variable(Y) is the total amount of reported synthetic opioid and heroin incidents(“drugreports” in the data set) of each county multiplied by the ratio of the amount of reported drugs in this county(“TotalDrugReportsCounty”) to the amount in the entire state.(“TotalDrugReportsState”)

The motivation for constructing the dependent variable is that drugreports is used to only measure the absolute number of drug popularity, which neglects the importance of different counties. So we make adjustments by giving weight. The weight is the ratio of reported drug incident of the county to the incident of the entire state, whose meaning is that the larger the weight, the larger spread threats its county has to other counties in the state.

METHOD:

There are 3 components of the model where different methods are adopted respectively which are XGBoost, Principal Component Analysis(PCA) and Linear Regression(LR) which will be discussed in detail later.

GENERAL IDEA:

We first select essential variables from the data set through XGBoost, and then implement PCA to further explore the classification of variables and finally construct a linear regression model with selected variables. By analyzing the plus-minus sign and absolute value of the parameters, we hope to get inspired and work out some policy suggestions and possible strategy to countering the crisis. The process is described through flow chart shown below.

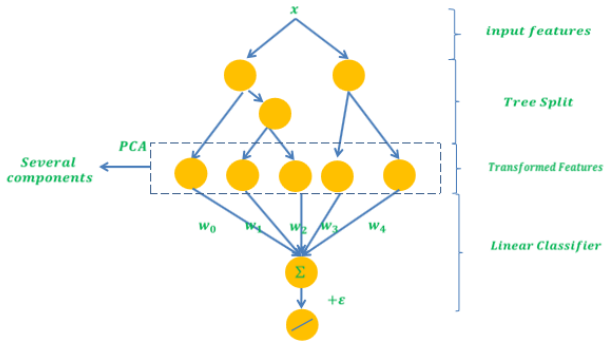

Figure 5_1 The Process of XLP

DATA PREPROCESSING

Since missing and abnormal data can seriously affect the results of the analysis, we pre-process the original data before analysis.

a. Missing Value Processing

We use different methods to process variables with various degrees of data loss:

1. For variables with large amount of data missing, we just delete it, since small data cannot provide enough and valuable information for our modeling.
2. For variables with a small amount of data missing, we use interpolation method to compensate the data.

b. Abnormal Value Processing

We call data which deviates from the mean value by more than two standard deviations an abnormal one. They are identified through descriptive statistics and thus replaced with the mean value.

PROCEDURES

1. Variable Selection-XGBoost

Since the number of variables in the original data set is too large to be screened by intuition, a suitable method of variable selection is needed. The existing methods of feature selection can be divided into three categories: Filter, Wrapper and Embedded. Embedded is a more commonly used method at this stage, including two types: one is to add a penalty item to the original model, such as LASSO and the other is based on the Decision Tree Model, such as the Random Forest, GBDT and XGBoost. Since XGBoost has such advantages as fast calculation speed, we finally adopt it.

The principle of XGBoost to sort the importance of variables is that the increment of the division index (ie Gini) measures the importance of variables each time a tree is divided. The variable importance is the accumulation of index increments in each division. The last 10 most important variables are displayed in the bar chart below.

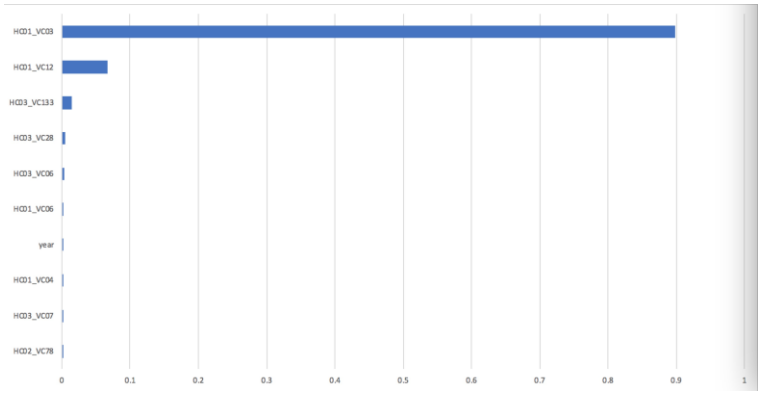

Figure5_2 Importance of Variables Selected

Variable	Code
Estimate/HOUSEHOLDS.BY.TYPE/Total.households	HC01_VC03
Estimate/HOUSEHOLDS.BY.TYPE/Family.households..families/Female.householder..no.husband.present..family/With.own.children.under.18.years	HC01_VC12
Percent/PLACE.OF.BIRTH/Native/Born in Puerto.Rico U.S.Island.areas/or born abroad to American.parents	HC03_VC133
Percent/RELATIONSHIP/Child	HC03_VC28
Percent/HOUSEHOLDS.BY.TYPE/Family.households..families/Married.couple.family/With.own.children.under.18.years	HC03_VC06
Estimate/HOUSEHOLDS.BY.TYPE/Family.households..families/With.own.children.under.18.years	HC01_VC06
year	
Estimate/HOUSEHOLDS.BY.TYPE/Family.households/families.	HC01_VC04
Percent/HOUSEHOLDS.BY.TYPE/Family.households..families/Married.couple.family	HC03_VC07
Estimate/SCHOOL.ENROLLMENT/Elementary.school/grades.1-8	HC02_VC78

Figure5_3 The Variable and its Corresponding Code

The categorization of the 10 variables are shown in the figure below:

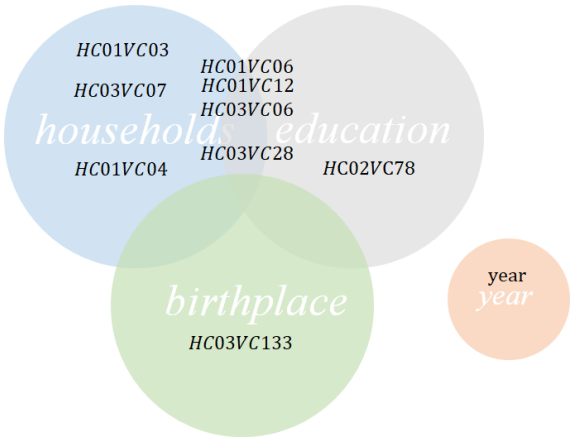

Figure 5_4 Results of Variable Categorization

2. Principal Component Analysis(PCA)

SUBJECT: the 10 fundamental factors selected through XGboost.

PURPOSE: On the basis of variable selection through XGboost, we adopt PCA, trying to figure out the intrinsic structure between variables and further classify variables.

PROCEDURES:

Step1: We first calculate the correlation coefficient matrix of the data set.

Step 2: Based on the matrix, we further perform principal component analysis and thus draw the scree plot (shown in Figure 5_5) and identify the quality of representation for each principle component(shown in Figure 5_6)

The first three principal components explain 83.6% of the total variance, and the elbow point of the scree plot is the third point, indicating that the first three principal components are essential enough to explain the main information contained in the 10 variables.

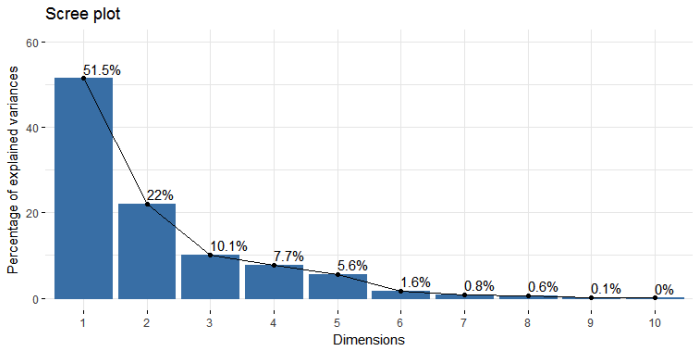

Figure 5_5 Scree Plot of PCA

The figure shows the representativeness of each variable in each principal component. The darker and larger the circle, the higher the representativeness of the variable in this principal component. It can be seen from the figure below that the first principal component is an absolute value, the second one is a proportional value and the third one represent time trend.

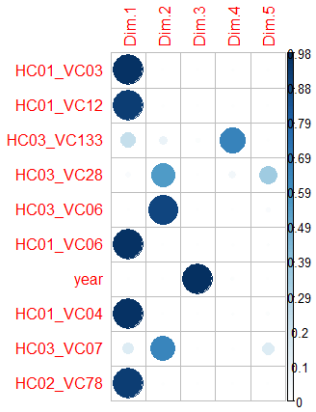

Figure 5_6: Quality of Representation

REFLECTION

The classification results of principal component analysis are different from our initial intuition.(10 variables can be classified according to family, education, birthplace, etc.) Instead, the results more depict different data types. (For example, the first principal component is "estimate..." which is the count data, while the second one is the proportional data).

We think about the possible reasons for this result. Initially, we wonder whether it is because of the difference between data scale (The range of values of each variable is large). However, considering PCA is conducted based on the correlation matrix rather than covariance matrix, which implies standardization, this speculation was ruled out. Therefore, we hold the assumption that perhaps the count data indirectly reflects the size of the county, and the proportional one more focus on the difference in distribution. Therefore, the

first principal component indicates the portrayal of the population size of the county, while the second one ignores the scale and focuses more on the proportion (structure) of the constituent components.

Later, considering that the classification constructed by the principal component is still relatively vague and not so conducive to interpretation if regarded as new variables for regression, so the latter regression model is still carried out with 10 variables instead of three principal components.

3. Linear Regression

Population regression model(PRF) is first constructed with the selected 10 independent variables:

$$\begin{aligned} Y = {} & \beta_0 + \beta_1(HC01VC03) + \beta_2(HC01VC12) + \beta_3(HC03VC133) + \beta_4(HC03VC28) \\ & + \beta_5(HC03VC06) + \beta_6(HC01VC06) + \beta_7 year + \beta_8(HC01VC04) \\ & + \beta_9(HC03VC07) + \beta_{10}(Hc02Vc78) + \varepsilon \quad , \quad \varepsilon \sim N(0, \sigma^2 I), \\ & I \text{ is the identity matrix}, \quad \sigma^2 \text{is the variance.} \end{aligned}$$

Using the U.S. Census socio-economic data provided over time period from 2010 to 2016, we set up the corresponding SRF and estimate parameters through OLS, and the regression outcome is shown below:

$$\begin{aligned} Y = {} & -244.961 + 0.006(HC01VC03) + 0.021(HC01VC12) + 21.280(HC03VC133) \\ & + 2.159(HC03VC28) - 4.343(HC03VC06) - 0.002(HC01VC06) \\ & - 0.062 year - 0.012(HC01VC04) + 5.941(HC03VC07) \\ & + 0.367(Hc02Vc78) + \varepsilon \quad , \quad \varepsilon \sim N(0, 92.353^2 I), \\ & I \text{ is the identity matrix}, \quad 92.353^2 \text{is the variance.} \end{aligned}$$

Results

	Dependent variable:
	Y
HC01_VC03	0.006***
	(0.0004)
HC01_VC12	0.021***
	(0.001)
HC03_VC133	21.280***
	(2.581)
HC03_VC28	2.159***
	(0.724)
HC03_VC06	−4.343***
	(0.800)
HC01_VC06	−0.002**
	(0.001)
year	−0.062
	(3.074)
HC01_VC04	−0.012***
	(0.001)
HC03_VC07	5.941***
	(0.461)
HC02_VC78	0.367***
	(0.025)
Constant	−244.961
	(6,179.397)
Observations	3,642
R^2	0.700
Adjusted R^2	0.699
Residual Std. Error	92.353 (df = 3631)
F Statistic	845.480*** (df = 10; 3631)
Note:	*p<0.1; **p<0.05; ***p<0.01

It is clear to see that except the variable ”year”, all the other nine variables are statistically significant at the significance of 5% and except “HC01_VC06”, the others are even significant at 1% level.

Among the statistically significant variables, those with a positive coefficient are promotion factors that boost the spread of drugs, including HC01_VC03, HC01_VC12, HC03_VC133, HC03_VC28, HC03_VC07 and HC02_VC78. They can be interpreted respectively as the fact that ceteris paribus, the more the households, the more the number of single-parent families with only mothers and children under 18 years old, the greater proportion of citizens born in Puerto Rico US, island, the great popularity drugs have.

Variables with negative coefficients including HC03_VC06, HC01_VC06, which are the household ratio of married couple family, with own children under 18 years etc. They have the effect to suppress the proliferation of drugs.

All the above factors can be categorized into 3 types: family structure, the proportion of minors and the place of birth. The social-economic meaning behind it lies in that:

place of birth. The social-economic meaning behind it lies in that:

1. Children with unhealthy family structure tend to lack parents' care and suffer from psychological trauma. Without the ability to distinguish right from wrong, they thus more likely to take the wrong path and become addicted in drugs.
2. The number of households represents the population size of the county. Therefore, the larger the population is, the more complex the social network is, and therefore the higher proportion of citizens with drug overdose.
3. External population born abroad to US tend to suffer from discrimination and saddled with more pressure. Therefore, we have the speculation that they are more likely to turn to addictive drugs to seek illusionary pleasure.

Since the opioid crisis is not purely triggered by physiology but more caused by social-economic factors, the fact is reasonable that opioid use persists and proliferates despite its known dangers.

VI. STRATEGY & EFFECTIVENESS TESTING

STRATEGIES:

Considering enforceability, we try to work out 2 possible strategies targeted at place of birth and family structure respectively according to the key factors we find above:

1. Design and carry out residence permit points system to control migrant population born abroad to US.
2. Conduct divorce penalty policy and strengthen marriage education to improve the complexity of family construction.

PURPOSE

To make our model more practical, we have a try to give out possible strategies according to the fundamental factors gained from the XLP model and at the same time develop the policy effect testing function so that we can test the effectiveness of strategies and identify significant parameter bounds that success is dependent upon.

SUBJECT

We continue to use the dependent variable in XLP but further select independent variables from the 10 in XLP in terms of the policy enforceability. Finally we mainly focus on household structure and place of birth as the entry point for policy implementation.

METHODS

Polynomial regression and data visualization.

PROCEDURES

Considering the possible interaction and other nonlinear relations between the two independent variables, we modify the linear regression model of XLP by adding the interaction term and quadratic term. Therefore, we construct a polynomial regression model:

$$y = \beta_0 + \beta_1(HC01VC12) + \beta_2(HC03VC133) + \beta_3(HC01VC12)^2 + \beta_4(HC03VC133)^2 + \beta_5(HC01VC12) \times (HC03VC133) + \varepsilon, \quad \varepsilon \sim N(0, \delta^2 I),$$
$$I \text{ is the identity matrix}, \delta^2 \text{ is the variance.}$$

Through OLS, we get the parameter estimation outcomes:

$$y = 18.15 - 5.593 \times 10^{-3} \times (HC01VC12) - 22.33 \times (HC03VC133) + 1.447 \times 10^{-7} \times (HC01VC12)^2 - 7.2 \times (HC03VC133)^2 + 8.277 \times 10^{-3} \times (HC01VC12) \times (HC03VC133) + \varepsilon, \quad \varepsilon \sim N(0, 68.64^2 I),$$
$$I \text{ is the identity matrix}, 68.64^2 \text{ is the variance. } (Multiple\ R^2 = 0.8338, p - value < 2.2 \times 10^{-16})$$

The coefficients of all the terms used are significant(not displayed in the paper), which proves that a nonlinear relationship exists between the two independent variables. Therefore, the strategy must consider these 2 factors simultaneously. To better show their influence on the extent of drug spread, we draw a three-dimensional surface plot and level plot as shown below. The dark purple dot in the surface plot marks the current state calculated from the mean value of given data.

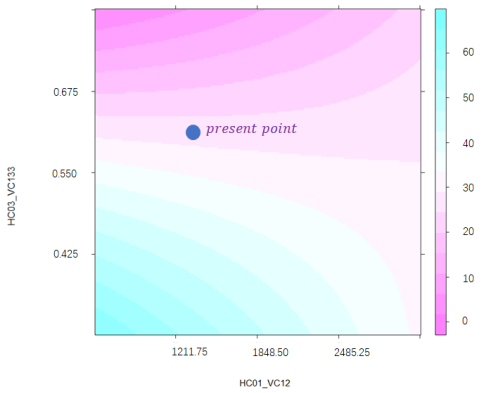

Figure6_1(a) Three-dimensional Surface Plot

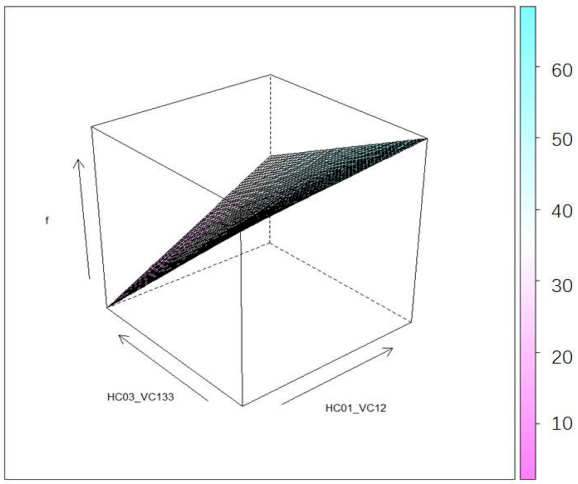

Figure6_2(b) Three-dimensional Level Plot

To prevent the spread, the strategy should be adjusted towards the pink area. The degree of specific adjustment depends on the actual situation. For instance, if the target is to lower the dependent variable Y(which represents the spread extent to 20, the shortest disturbance path should be vertical to the contour. To achieve the goal, through calculation, the mean of HC03VC133 should be improved to 0.633 and HC01VC12 reduced to 1212.5.

Therefore, our model fully considers the mutual influence of the two policy adjustment processes, and enables the simultaneous implementation of the two policies to achieve the best possible results with the least possible policy changes.

VII. MODEL EVALUATION

A. Advantages & Innovation

1. Reasonably select the variables and subjects to be modeled.

The large data size and missing information increase the difficulty of modeling. Despite this, we still select objects worthy of analysis with refinement, therefore work out a concise model with information remained to the greatest extent.

For example, there are numerous variables offered to select while constructing XLP, we select 10 variables at last based on the importance rank by XGBoost, which guarantees the accuracy and objectivity to avoid the omissions and speculations of manual screening

2. Choose the appropriate model with modification and synthesis.

We choose a novel and reasonable method to calculate the order of occurrence of drugs between states, and use the clustering method to classify the pathways, hierarchical and k-means clustering are combined to ensure the stability and accuracy of the algorithm.

When studying the spread routes and characteristics, we tailored different models (DIM and GPM) to fit and predict. Both DIM and GPM complement with each other, thus the prediction effect greatly improved.

Besides, after the model test and sensitivity analysis, we find that no matter how precise the DIM+GPM model is, its estimation and policy suggestions are still reasonable, justifying the robustness of the model

3. Visualize to make the results more intuitive.

We use the method of data visualization both when exploring features of original data and handling results. For instance, the spread routes are displayed both in the line graph and through marks on the map, which are more rigorous and more intuitive respectively. Both methods can serve as the reference in policy-making.

4. Focus on the interpretability and practical value of the models.

Machine learning, as a substitution for linear regression adopted in XLP has better prediction effect. However, taken the explanatory ability into account, we finally choose the method of linear regression.

5. Continuous model improvement and adjustment.

The model design in strategy and effectiveness test is not limited to the linear regression used in XLP and the polynomial regression model is constructed considering the interaction between variables.

B. Disadvantages & Possible Modifications

1. The DIM model is not so efficient when λ is negative, so GPM is used for modification. Despite the good effect, we must acknowledge that its reality fit degree is not as high as DIM. Therefore, we have considered the introduction of the purifying rate and the number of purified counties, which can reflect the influence of external factors such as policy intensity, economic reasons, drug efficacy and side effects to some extent. With limit time, it is too late to estimate parameters and fit predictions. If given opportunity to continue to explore later, it is a good entry point.
2. In XLP, we only implement one method to filter variables. If possible, it would be better to combine a variety of selection methods to enhance the robustness of the model.
3. Besides, the analytical expression of parameter bounds is not given in strategy and effectiveness test. Instead, we figure out the bounds with grid search and visualization which lacks generalization capacity.

VIII. SENSITIVITY ANALYSIS

Considering the possibility of error in parameter estimation, we analyzed the sensitivity of the DIM model to investigate the sensitivity of the drug outbreak time to the parameter λ.

Our approach to sensitivity analysis is regard λ as an unknown parameter for the solution for tm:

$$t_m = \lambda^{-1} ln(\frac{1}{i_0} - 1)$$

If there is a relative change in volume, what is the corresponding relative change of tm. We construct the measure $S(t_m, \lambda)$for sensitivity according to the definition of the derivative.

$$S(t_m, \lambda) = \frac{dt_m}{d\lambda} \cdot \frac{\lambda}{t_m} = \frac{ln(\frac{1}{i_0} - 1)}{\lambda \cdot t_m}$$

We then calculated the sensitivity of the drug model within each state, and the results are shown as the table below:

State	SubstanceName	λ	tm	S(tm,λ)
KY	Buprenorphine	0.18	2010	-0.0033
KY	Fentanyl	0.43	2016	-0.0006
KY	Heroin	0.32	2013	-0.0021
OH	Fentanyl	0.65	2012	-0.0026
OH	Tramadol	0.51	2014	-0.0013
PA	Fentanyl	0.47	2011	-0.0029
VA	Buprenorphine	0.12	2013	-0.0022
VA	Fentanyl	0.33	2015	-0.0012
VA	Heroin	0.17	2009	-0.0041
WV	Buprenorphine	0.07	2014	-0.0013
WV	Fentanyl	0.31	2019	0.0009

Figure 8_1 Sensitivity of the drug model within each state

It can be found that the sensitivity is very low, which indicates strong robustness. So even if estimation of λ is not very precise, the tm is still accurate with guidance value for the United States governments' policy-making.

IX. CONCLUSION

Having learned the background and problem, we construct several models based on the purposes which are listed as follows:

- identify the spread routes and characteristics of the reported synthetic opioid and heroin incidents
- make reasonable predictions and suggestions as reference for US governments.
- explore the fundamental factors for the opioid crisis
- enact possible strategies to counter the crisis and test their effectiveness

Accordingly, we construct 3 models: Pathway Model based on Clustering(PMC) and Drug-Invasion & Grey Precaution Model (DIM+GPM) to identify the inter-state and inner-state spread patterns and features respectively, XGBoost+LR with PCA model(XLP) to figure out fundamental factors so that we can give out possible strategies accordingly. The model test and sensitivity analysis justify the robustness of our models to be used for estimation and prediction.

Some primary conclusions and suggestions are listed as follows:

- Conclusion:
  - Ohio and Pennsylvania suffer from the most severe opioid drug misuse which require great attention
  - Heroin which used to be the dominant drug used according to the reported incidents is gradually substituted by other synthetic opioids.

- Suggestions:

- Set up inspection teams to remove drug trade den, supplemented with advocacy on drug boycotts.
- Reinforce science education on drug misuse to improve people's knowledge about drug harm
- Design and carry out residence permit points system to control migrant population born abroad to US.
- Conduct divorce penalty policy and strengthen marriage education to improve the complexity of family construction.

REFERENCE

- [1] M. M. Meerschaert. Mathematical Modeling, Fourth Edition. Academic Press, 2014.
- [2] M. Martcheva. An Introduction to Mathematical Epidemiology. Springer, 2015.
- [3] N. Dasgupta , L. Beletsky , D. Ciccarone . Opioid Crisis: No Easy Fix to Its Social and Economic Determinants. American Journal of Public Health, 2018, 108(2): 182 - 186.
- [4] Q. Jiang, J. Xie, J. Ye. Mathematical Modeling, Fourth Edition. Higher Education Press, 2011.
- [5] T. Chen, C. Guestrin. XGBoost: A Scalable Tree Boosting System. ACM, 2016, pp. 785-794.
- [6] Z. Wu. Application of MATLAB in Mathematical Modeling. Beijing University of Aeronautics and Astronautics, 2014.