Drug abuse models based on large data analysis

Bingze Chen, Yaoran Huang

Xiamen University, Xiamen, Fujian, 361001

Email:649851335@qq.com

Keywords: Drugreports; Curve Fitting; Correlation; PCA

Abstract: In our report, we summarized and analyzed the data on the number of drug cases and related factors given by the NFLS database since 2010. After modeling, we get the models of the number of drugreports changing with time in five states, and modify and optimize them. Based on this model, the number and distribution of drug cases in each state in the future are predicted, and suggestions for controlling drug transmission are put forward.

1. Introduction

Trump made the announcement while on vacation at his golf club:“The situation is critical and the government will escalate the drug abuse problem into an international emergency.” The epidemic of the opioid is sweeping the United States like a wind now. Opioids which can be used for the treatment can also have negative influence on people such as opioid use disorder, hepatitis, and HIV infections. Even the burden on the important economic sector will become heavier If the opioid crisis spreads to all walks of life.

Merely enforcing existing laws is a complex challenge for government agencies. Therefore, we design a model by analyzing the related data about Ohio, Kentucky, West Virginia, Virginia, and Pennsylvania and present a strategy to help the government take appropriate actions.

2. Problem Analysis

We set the following goals in order to meet the requirements of the problem.

1) Use the NFLIS data provided to build a appropriate model to describe the spread and characteristics of opioid incidents in and between the five states and their counties. And use the model to deduce the possible location where specific opioid use started in five states.

2) Use the model to predict what will happen in the future and when and where they will occur. And determine at what drug identification threshold levels do they occur.

3) Use the U.S. Census socio-economic data provided to identify whether the use or trends-inuse somehow associated with the socio-economic factors. We found that it is indeed like this. Then modify the model to include important factors from the data set.

4) Use a combination of our Part 1 and Part 2 results, put forward a possible strategy for countering the opioid crisis. Test the effectiveness of this strategy. Identify any significant parameter bounds that success (or failure) is dependent upon.

2.1 Ideas of Solving the Problem

We first analyzed and processed the data on drug cases and related factors from 2010 to 2017 from the NFLS database. After modeling with equation fitting model, we obtained the model of drug reports quantity changing with time in five states, inferred the possible origin of drugs in each state based on this model, and predicted the number and distribution of drug cases in each state in the future. Then we revise and optimize the model considering social and economic factors, and put forward some suggestions on controlling drug transmission in combination with the two parts. After all the models have been established, we have carried out sensitivity testing and analysis of advantages and disadvantages. Finally, a memorandum was written to the Chief Administrator.

DOI: 10.25236/isrme.2019.119

2.2 Data Preprocessing

The tables in the original data are in chronological order, listing the FIPS numbers and the number of drug reports of each state and its subordinate counties, but it is inconvenient for data visualization. So we compiled a program to get tables of the number of opioids reported in each state and county over time, and to calculate the changes of the total number of states. In each state, we used 5% of the counties with the largest number of drug reports as the criterion, and abandoned counties with the average number of drug reports less than that criterion to get a table of quantities. For example, the following figure shows the number of major counties in Ohio where drugs are reported.

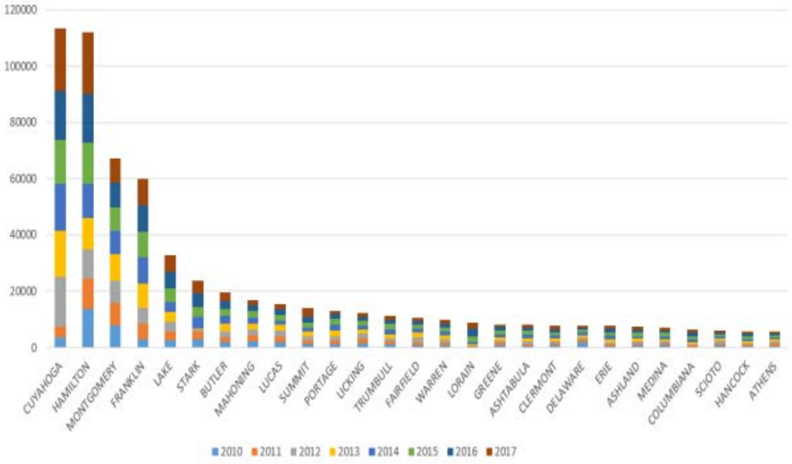


Figure 1. Quantity Map of Major Drug Reporting Counties, Ohio, 2010-2017

3. Symbol Description

Owing to the limitation of our knowledge, we make some assumptions to perform the model that is more reasonable and accurate. And they are the premise for our consecutive analysis. The assumptions and justifications are as follows.

1) Assume that the number of drug reports which occur in an area over a period of time is proportional to and fairly close to the number of local drug users, so in this model the number of drug reports can represents local drug users.

2) Assume that the natural birth rate and natural death rate in an area are roughly equal. And during the application of the model, the five states do not have mass migration of drug users with the external environment. So they have no effect on the local drug population.

3) Assume that we do not take into account the large impact of a pandemic, pestilence, or natural disaster. The local population do not change much over a period of time.

4) Assume that the attractiveness of drugs to people (the possibility of a person being addicted to drugs) does not change as a result of technological development or the varying resistance of different people.And a person who has successfully quit drugs has the same possibility of being spread as a drug addict. It can simply the model and eliminate extraneous factors.

We use some symbol to simplify representation:

TABLE I. Symbol Description

Symbol	Description
N	The number of total population in one state
S	The number of synthetic opioid users
i	The number of Non-synthetic opioid users
Λ	The proportion of people who begin to be addicted to drugs
μ	The proportion of people who quit drug addiction
υ	The number of people who begin drug abuse caused by one other people
t	Time
Δt	Small change in time

4. Preliminary Model Construction

4.1 For One State

Take the simplest model with only one states as an example, we are capable to plot the relation between different groups of people (See The Figure 2).

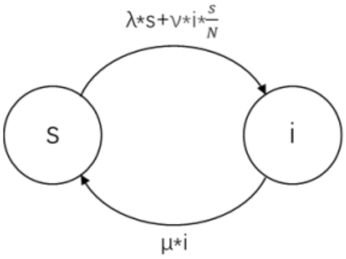


Figure 2: Relation Between Different Groups people

Our model is based on the model of SIS on heredity but they have obvious differences. SIS model infected people only from the contact of infected people, and people who do not use opioids are also likely to become users, so they can not be equated with each other. We change the meaning of some symbols. Λ represents the proportion of people who begin to be addicted to drugs, μ represents the proportion of people who quit drug addiction and υ represents the number of people who begin drug abuse caused by one other people.

Now we consider changes in the number of people using opioids over a short period of time:

$$i(t+\Delta \mathrm{t})-i(t)=\left[\lambda \mathrm{s}(\mathrm{t})+\mathrm{vi}(t) \frac{s}{N}-\mu i(t)\right] \Delta \mathrm{t}$$

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

$$\frac{d i}{d t}=-\frac{v}{N} i^{2}(t)-(\lambda+\mu-\mathrm{v}) i(t)+\lambda \mathrm{N}$$

4.2 Compute Model

Unlike the SIS model, which only produces exposure to infected people, even people who have never used opioids before are likely to become users, so we can't ignore this proportion.

$$\frac{d i}{d t}=-\left[a i^{2}(t)+b i(t)+c\right]$$

After equivalent transformation:

$$\frac{d i}{a i^{2}(t)+b i(t)+c}=-d t$$

By integrating both sides of the equation at the same time, we can get:

$$\int \frac{d i}{a i^{2}(t)+b i(t)+c}=\int-d t=\text { Const }-t$$

The left integral can be obtained by factorization:

$$\ln \frac{i(t)-m}{i(t)-n}=\mathrm{d} *(\text { Const }-t)$$

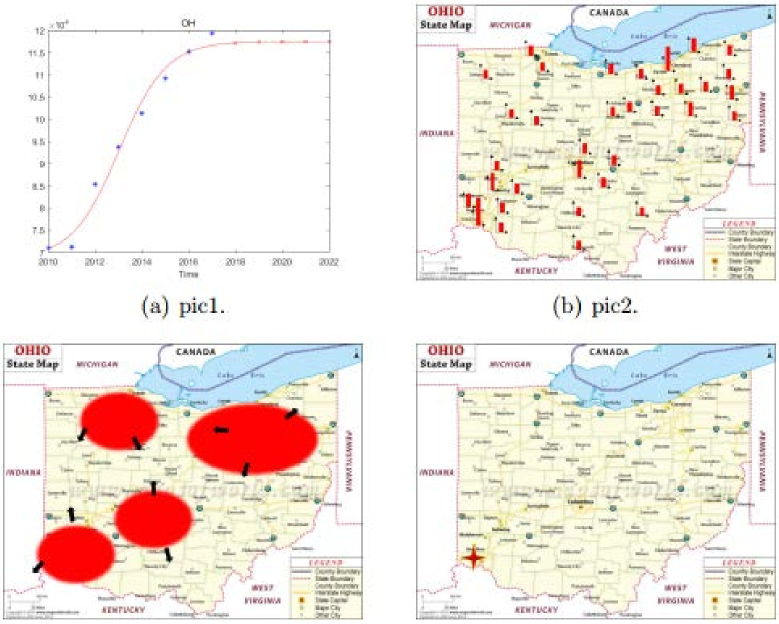
$$i(\mathrm{t})=\frac{-n d e^{C-t}-m}{d e^{C-t}-1}$$

So the model chosen for fitting is:

$$i(\mathrm{t})=-\frac{\mathrm{n}+m}{d e^{c-t}-1}-n$$

4.3 The Equation Fitting Model

In each state, the number of reported opioid abuse incidents varies in different counties, we can get four pictures(see the Figure 3) below from data preprocessing and the above formulas.

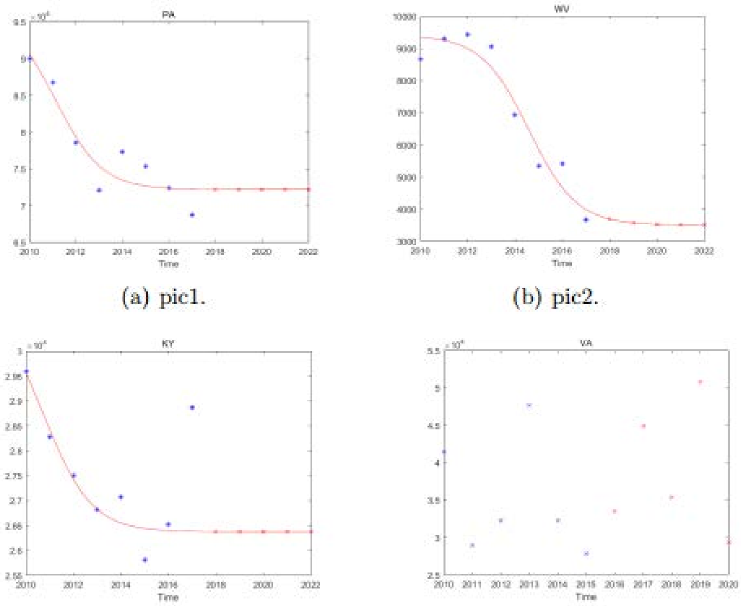


(a) pic1.

(b) pic2.

Figure 3: Four Pictures A

Figure 4 is the fitting model of the other four states. We can see that the fitting of VA is not very good. The reason is that the data changes too dramatically. There are many reasons for this change, but the biggest possible reason is the large-scale population flow.



(a) pic1.

(b) pic2.

Figure 4: Four Pictures B

4.4 Solve the Problem

4.4.1 Locations Where Specific Opioid Might Start

It can be inferred from the distribution of reported number of drugs in each state that drugs in Ohio originated from CUYAHOGA, drugs in Pennsylvania originated from PHILADELPHIA, drugs in Kentucky originated from JEFFERSON, drugs in Virginia originated from FAIRFAX, and drugs in West Virginia originated from KANAWHA. You can find the locations of these locations in the trend map.

4.4.2 Specific Concerns and Drug Identification Threshold Levels

As can be seen from the trend chart, the number of drugs in Virginia is changing dramatically. In the next two to three years, the government should take measures to control the number of drug crimes in Virginia. The number of drug reports in Ohio will continue to increase in the future. The police can take comprehensive measures to strengthen drug control in Ohio. The strength of crime inspection to control the number of drug-related crimes in the state. The number of drugs in Kentucky has also changed from a previous downward trend to an upward trend, which should be the focus of attention in the next few years. In addition, the number of drugs in West Virginia and Pennsylvania is declining, which is a satisfactory result

In our model, we think that when the number of drug reports in a state enters a relatively stable value, the number of drug reports is its drug recognition threshold level. Among them, Ohio’s drug identification threshold levels appeared in 2019, when the number of drug crimes in Ohio will oscillate near a threshold of 1.2e+5. The Pennsylvania threshold was 7.3e+3 in 2018. Kentucky’s threshold was 2.65e+4 in 2017. The thresholds for Virginia and West Virginia both appeared in 2017, and their thresholds were 4.5e+4 and 4e+3, respectively.

5. Revised Model Construction

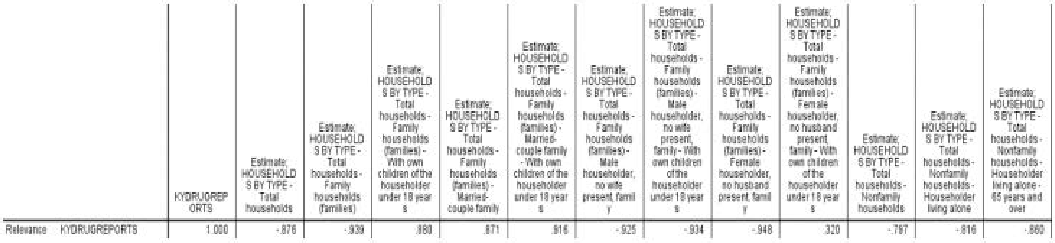
Step 1:We use y to regress x1, x2, x3, ... xk, we can get the equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$$

Step 2:The formula for calculating the complex correlation coefficient is as follows:

$$R = \frac{\sum(y-\overline{y})(\hat{y}-\overline{y})}{\sqrt{\sum(y-\overline{y})^2(\hat{y}-\overline{y})^2}}$$

In this problem, there are many possible related factors, and using SPSS software can greatly simplify our calculation. Here we take the number of drug reports in KY State as an example, assuming that the value is a linear combination of multiple related factors. The statistical moments of social economy and the number of drug reports in the state from 2010 to 2016 are merged into a table and input into spss. The correlation matrix can be obtained by correlation analysis. The first row or list of matrices shows the correlation coefficients between the socioeconomic factors and the number of drug reports in the state. The value of correlation coefficient varies from - 1 to 1. The closer to 1 or - 1, the higher the degree of correlation (positive or negative). Here, we set 0.7 as a threshold, the absolute value is higher than 0.7, we think it is significantly related to the number of drug reports.



		KYDRUGREPORTS	Estimate; HOUSEHOLDS BY TYPE - Total households	Estimate; HOUSEHOLDS BY TYPE - Total households - Family households (families)	Estimate; HOUSEHOLDS BY TYPE - Total households - Family households (families) - With own children of the householder under 18 years	Estimate; HOUSEHOLDS BY TYPE - Total households - Family households (families) - Married-couple family	Estimate; HOUSEHOLDS BY TYPE - Total households - Family households (families) - Married-couple family - With own children of the householder under 18 years	Estimate; HOUSEHOLDS BY TYPE - Total households - Family households (families) - Male householder, no wife present, family	Estimate; HOUSEHOLDS BY TYPE - Total households - Family households (families) - Male householder, no wife present, family - With own children of the householder under 18 years	Estimate; HOUSEHOLDS BY TYPE - Total households - Family households (families) - Female householder, no husband present, family	Estimate; HOUSEHOLDS BY TYPE - Total households - Family households (families) - Female householder, no husband present, family - With own children of the householder under 18 years	Estimate; HOUSEHOLDS BY TYPE - Total households - Nonfamily households	Estimate; HOUSEHOLDS BY TYPE - Total households - Nonfamily households - Householder living alone	Estimate; HOUSEHOLDS BY TYPE - Total households - Nonfamily households - Householder living alone - 65 years and over
Relevance	KYDRUGREPORTS	1.000	-.876	-.939	.880	.871	.916	-.925	-.934	-.948	.320	-.797	-.816	-.860

Figure 5: Relevance matrix

The Figure 5 is the first line of the correlation matrix (only give a small part here because the matrix is too long).

5.1 Modify Model

Because there are too many parameters, it is not suitable to write all parameters directly into the model. So we use the principal component analysis method to reduce the dimension of the parameters, and get the three main parameters, which are the linear combination of all the above variables that have an impact on the model. Principal Component Analysis (PCA) is a method of transforming multiple parameters into several unrelated comprehensive parameters by using the idea of linear transformation and on the premise of losing little information. If there are many original variables: x1, x2, x3, ... xk, One or two principal components should be extracted to retain at least 90% of the information of the original variables. The extracted principal components can be written as follows:

$$F_1 = b_{11} \times X_1 + b_{21} \times X_2 + b_{31} \times X_3 + \cdots$$

$$F_2 = b_{12} \times X_1 + b_{22} \times X_2 + b_{32} \times X_3 + \cdots$$

In our model, we use SPSS to make principal component analysis of the obtained parameters. For example, for KY state, the 10 parameters with the greatest correlation with KY state are taken as the original parameters, and the principal component analysis is carried out. We get the Scree plot(Figure 6:The abscissa is the number of components and the ordinate is the eigenvalue).

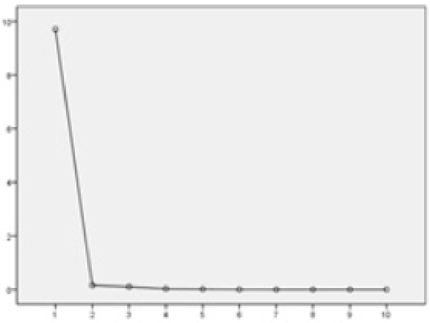
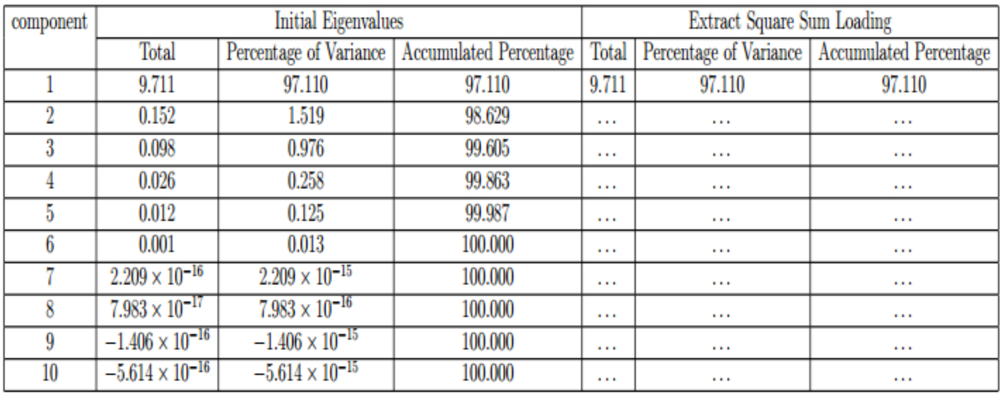


Figure 6: Gravel Maps From Principal Component Analysis of KY State

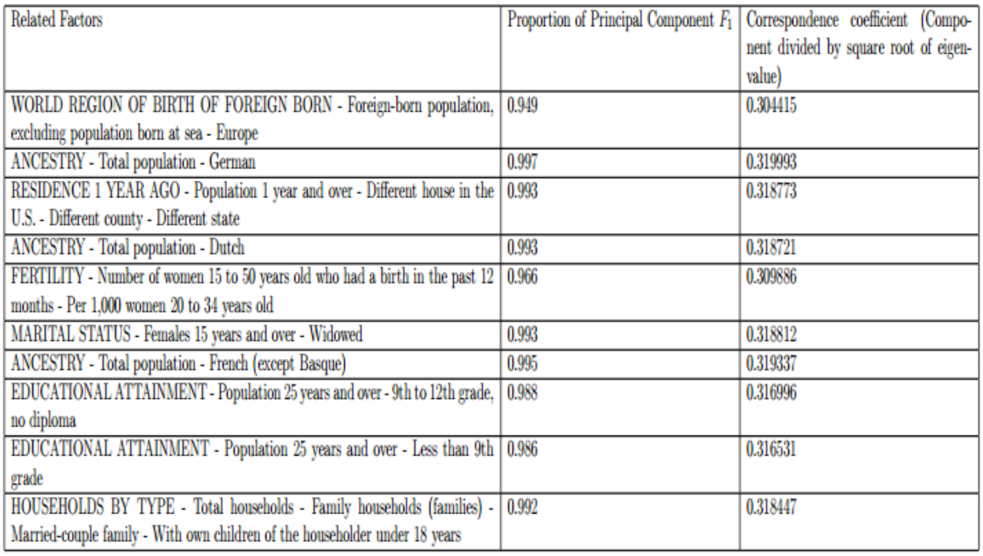
TABLE II. Explained Variance



component	Initial Eigenvalues			Extract Square Sum Loading		
	Total	Percentage of Variance	Accumulated Percentage	Total	Percentage of Variance	Accumulated Percentage
1	9.711	97.110	97.110	9.711	97.110	97.110
2	0.152	1.519	98.629	…	…	…
3	0.098	0.976	99.605	…	…	…
4	0.026	0.258	99.863	…	…	…
5	0.012	0.125	99.987	…	…	…
6	0.001	0.013	100.000	…	…	…
7	2.209×10^{-16}	2.209×10^{-15}	100.000	…	…	…
8	7.983×10^{-17}	7.983×10^{-16}	100.000	…	…	…
9	-1.406×10^{-16}	-1.406×10^{-15}	100.000	…	…	…
10	-5.614×10^{-16}	-5.614×10^{-15}	100.000	…	…	…

As shown in the table II above, a principal component F1 was extracted, and its characteristic value was 9.711, which means that it contributed 97.11% of the total variance, more than 90%, so it can be used as the only principal component after analysis. The matrix contains the corresponding coefficients (the square root of the component/eigenvalue) needed for the next calculation of F1.The component matrix obtained is as Table III.

TABLE III. Explained Variance



Related Factors	Proportion of Principal Component F_1	Correspondence coefficient (Component divided by square root of eigenvalue)
WORLD REGION OF BIRTH OF FOREIGN BORN - Foreign-born population, excluding population born at sea - Europe	0.949	0.304415
ANCESTRY - Total population - German	0.997	0.319993
RESIDENCE 1 YEAR AGO - Population 1 year and over - Different house in the U.S. - Different county - Different state	0.993	0.318773
ANCESTRY - Total population - Dutch	0.993	0.318721
FERTILITY - Number of women 15 to 50 years old who had a birth in the past 12 months - Per 1,000 women 20 to 34 years old	0.966	0.309886
MARITAL STATUS - Females 15 years and over - Widowed	0.993	0.318812
ANCESTRY - Total population - French (except Basque)	0.995	0.319337
EDUCATIONAL ATTAINMENT - Population 25 years and over - 9th to 12th grade, no diploma	0.988	0.316996
EDUCATIONAL ATTAINMENT - Population 25 years and over - Less than 9th grade	0.986	0.316531
HOUSEHOLDS BY TYPE - Total households - Family households (families) - Married-couple family - With own children of the householder under 18 years	0.992	0.318447

Calculate mean and standard deviation by column:

$$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$S_j = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)}{n-1}}$$

Then calculate the standardized data:

$$X_{ij} = \frac{x_{ij} - \overline{x_j}}{S_j}$$

Here, the normalized matrix can be easily obtained by SPSS.

TABLE IV. Table of Major Related Factors of Normalization

WORLD REGION OF BIRTH OF FOREIGN BORN-Foreign-born population, excluding population born at sea-Europe	ANCESTRY-Total population-German	RESIDENCE 1 YEAR AGO-Population 1 year and over-Different house in the U.S.-Different county-Different state	ANCESTRY-Total population-Dutch	FERTILITY-Number of women 15 to 50 years old who had a birth in the past 12 months-Per 1,000 women 20 to 34 years old	MARITAL STATUS-Females 15 years and over-Widowed	ANCESTRY-Total population-French (except Basque)	EDUCATIONAL ATTAINMENT-Population 25 years and over-9th to 12th grade, no diploma	EDUCATIONAL ATTAINMENT-Population 25 years and over-Less than 9th grade	HOUSEHOLDS BY TYPE-Total households-Family households (families)-Married-couple family-With own children of the householder under 18 years
1.89984	1.67101	1.51818	1.62387	1.47392	1.43587	1.45196	1.45421	1.41470	1.43218
0.77568	0.8865	0.99148	0.94101	1.02792	1.06639	1.01160	0.93286	0.97197	1.02727
-0.10410	0.21170	0.19805	0.10766	0.55218	0.36602	0.43714	0.35576	0.38588	0.4552
-0.54399	-0.21073	-0.07277	-0.0035	-0.54011	-0.07601	-0.10804	-0.04744	-0.06436	-0.17849
-0.42698	-0.5406	-0.39430	-0.64402	-0.81710	-0.73855	-0.72114	-0.37498	-0.40270	-0.65474
-0.89797	-0.96276	-1.15051	-0.99088	-1.03149	-1.08591	-1.04146	-0.99679	-0.95150	-0.86804
-0.70247	-1.05516	-1.09013	-1.03405	-0.66531	-0.96781	-1.03004	-1.3236	-1.3539	-1.21338

Let the values of 10 variables in the k-th row of the above matrix be Xkj (normalized values), and the corresponding coefficient is the component value divided by square root of eigenvalue. The expression for calculating F1 can be obtained from the table above:

$$F_1 = 0.304415X_{k1} + 0.319993X_{k2} + 0.318773X_{k3} + 0.318721X_{k4} + 0.309886X_{k5}$$

$$+ 0.318812X_{k6} + 0.319337Xk7 + 0.316996Xk8 + 0.316531Xk9 + 0.318447X_{k10}$$

The revised equation is as follows:

$$i(\mathrm{t}) = -\frac{\mathrm{n} + m}{de^{C-t} - 1} - n + k \times F$$

Among them,F is the parameter obtained by correlation calculation, and the size of F is determined by the parameter with the greatest correlation. Therefore, on the basis of the original equation, k=times/F term is added to the right side of the equation as a correction. K can be calculated by the variation of the difference between the predicted value and the actual value, and the specific value can be given by the following formula:

$$\mathrm{y} - \mathrm{y}' = \mathrm{k} \times \mathrm{F}$$

6. The Strategy for the Opioid Crisis

Through the analysis of the data of five states, we can get more significant correlations.

- Educational attainment: The proportion of older but less educated people in the state is significantly positively correlated with the number of drug reports in the state.
- Marital status: The higher the proportion of married women in widows'families, the higher the number of drug reports. The higher the proportion of divorced families, the higher the number of drug reports. And the more separated families, the higher the number of drug reports.
- Family status: There is a significant positive correlation between the number of solitary

population and the number of drug reports.

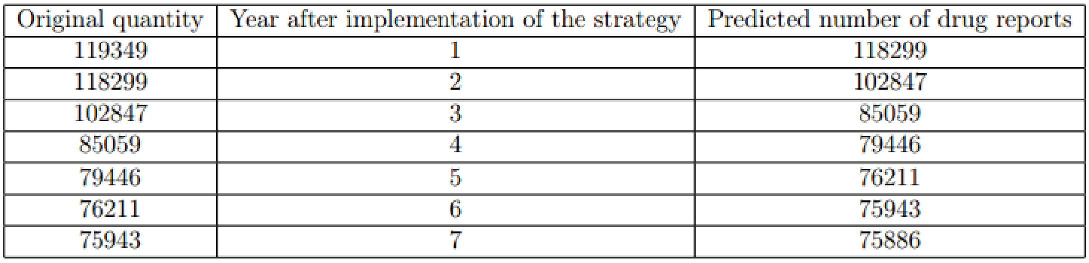
- Military Service Status: There is a significant positive correlation between the number of veterans and the number of drug reports.
- Female fertility: The number of fertile women aged 15-34 is positively correlated with the number of drug reports.

Therefore, we can make the following suggestions to the local government to strengthen the control of opioid use in the state:

- Improve the level of education management, especially the education of young people (18-25 years old). Because in the socioeconomic data of the five states analyzed, we can find that the number of people who meet the age of 25 and above and who do not have a high school diploma is positively correlated with the number of local drug cases. Therefore, we suggest that the government strengthen the control in this respect.
- Conciliation of family contradictions to avoid large numbers of separated or solitary families.
- Strengthen sexual safety education, especially for Female Adolescents.

Based on the implementation of the above strategies, we can estimate the changes of some socio-economic data in this year. Assuming that our strategy is successfully implemented, the number of undereducated adolescents and divorced families in a state is expected to decrease in the coming year, we calculate a new F value based on the above theory and use our model to predict it. The following table is obtained: (Take OH State as an example).

TABLE V. Prediction Table



Original quantity	Year after implementation of the strategy	Predicted number of drug reports
119349	1	118299
118299	2	102847
102847	3	85059
85059	4	79446
79446	5	76211
76211	6	75943
75943	7	75886

It can be seen that in addition to the slow decline in the first year, the number of drug reports has been significantly controlled in the next few years, reaching a new threshold in about seven years, which represents the success of our strategy in controlling the spread of drugs.

7. Conclusions

In our modeling process, we use the analysis method of combining theory with actual data. Firstly, based on the principle of drug transmission, we get an equation of the number of drug cases changing with time, and then we get five different models by fitting the actual data of different states. We successfully used the model to predict the data for the next few years, and gave the possible thresholds for the number of drug cases in each state in the future.

When we analyze the influence of social and economic factors on our model, we use PCA analysis method to reduce the ten most relevant factors to a principal component F, and use F as the only parameter of social and economic factors to write into our model, and complete the optimization of the model. In this process, we find several parameters that have the greatest impact on the model, such as the education level of the population, family environment and so on. We give our strategy on how to optimize these indicators for a state. Tests in our model show that our strategy is effective.

References

[1] Lindsay I Smith A Tutorial on Principal Components Analysis, February 26, 2002

[2] Haiming Lin, Zifang Du Problems Needing Attention in Comprehensive Evaluation of Principal Component Analysis, 2013.Aug.

[3] Jiye Liang, Chenjiao Fen, Peng Song Summary of Big Data Relevance Analysis, 2016, (1):1-18.DOI:10.11897/SP.J.1016.2016.00001

[4] Hangxing Zhou, Songcai Chen Ordered Discriminant Canonical Correlation Analysis, 2014,9.doi:10.3724/SP.J.1016.2011.01820

[5] Shan Wang, Huiju Wang, Xiongpai Tan Architecture Big Data: Challenges, Status Quo and Future, 2011, (10):1741-1752.doi:10.3724/SP.J.1016.2011.01741

[6] Enhong Chen, Jian Yu Frontier of Big Data Analysis Special Journal 2014, (9):1887-1888.doi:10.13328/j.cnki.jos.004652.

[7] LU, X., BENGTSSON, L., HOLME, P.. Predictability of population displacement after the 2010 Haiti earthquake. Proceedings of the National Academy of Sciences of the United States of America, 2012,29(29):11576-11581.