A New Regional Prediction Method for Opioid Abuse using K-means Method

Xueda Li[1, a], Ying Xiao[2, b]

[1] Dalian University of Technology, Panjin, China;

[2] Dalian University of Technology, Panjin, China.

[a]1298359043@qq.com, [b]3600382231@qq.com

Abstract

The United States is experiencing a national crisis of opioid abuse. Opioid abuse has serious negative effects not only on the health of the American people, but also on important sectors of the U.S. economy. Therefore, in this paper, we aim to identify any possible sites for specific opioid use. We developed a macro model of drug epidemic in states and counties. In the state model, we get the drug spread curve of each state, while in the county model, we cluster the counties of each state through the k-means clustering method. By analyzing the characteristics of the clustering center, we get the regions that may have started to use specific opioid drugs.

Keywords

K-means; curve fitting.

1. Introduction

Opioids have long been used to manage pain (legal, prescription use). In recent years, the number of cases of infectious diseases caused by the use of opioids has continued to rise, and the relevant departments in the United States are gradually aware of the hazards of opioid abuse.

To solve the problem of opiates, we need to establish models to predict the trend of the spread of opiates abuse. At the same time, we need to speculate that the most serious phenomena of opiates abuse occur in the States and counties.

In order to solve the possible predicament of opioids in the States and their subordinate counties, we first established the spreading models of the States and counties, and then analyzed the spreading characteristics of opioids and heroin drugs in their respective states and counties through these models, and predicted the future situation, and found out the possible time and place of the problems.Organization of the Text

2. Model Establishment

2.1 State model

Because the number of States is generally small, we can accurately fit the spread curve, and then analyze its characteristics according to the fitting curve. The main process is as follows.

First, we preprocess data from all states. We add up the different types of DrugReports of Substance Name each year in each state to get the new construction variable Sum of DrugReports. We use sum of DrugReports to represent the sum of opioid and heroin use cases in each state. After eliminating the repetitive data from the annual Total Drug Reports State of each state, we divide sum of DrugReports by TotalDrugReportsState to get another new constructive variable Persent. We use Persent to represent the proportion of opioid and heroin use cases in each state in the total number of drug use cases.

2.2 Spread model of county

Because of the large number of counties, it is more troublesome to deal with, so we first need to separate the data according to the different states. After that, we process the split data. In order to cluster the data better, we need to reconstruct the indicators. First, we select Drug Reports from each

county every year and add them to get a new variable, Sum of Drug Reports (COUNTY), which is used to represent the number of opioid and heroin cases in each county. Total Drug Reports County of each county is selected as the second clustering index, which indicates the number of drug cases in each county every year. At the same time, we choose the trade between Sum of Drug Reports (COUNTY) and Total Drug Reports County as our third variable, Persent (COUNTY), which is used to indicate the proportion of opioid and heroin cases that occur annually in each county in the total drug cases in that county.

Then we deal with the missing values of the three variables and use linear interpolation to fill the missing values. Then, in order to avoid the impact of years on clustering, we use the exponential smoothing method to process the data of these three variables from 2010 to 2017. By trying different smoothing indices, we get the optimal smoothing index of 0.3, thus predicting the values of each county in 2018. These data are used for clustering.

Finally, we clustered the data processed by each county by K-means. We designated m-clusters. According to the results of K-means clustering, we got the clustering indicators (Sum of Drug Reports (COUNTY), Total Drug Reports County, Persent) of the areas where drug abuse might occur. We also distinguished the tasks that may have begun to use specific opium classes according to the results of the final clustering center. Where is the possible location? Moreover, the model can also adjust the threshold values of the three indicators of possible locations according to the actual adjustable size of m, so that the model has stronger practical application value.

3. Simulations and Experiments

We used data from five U.S. States and economic data from the U.S. census to test.

3.1 Verification of State Model

Due to the small number of states, there are five in total, so we can conduct accurate spread curve fitting for them, and then analyze their characteristics according to the fitted curve. Firstly, we preprocessed all the state data. We will in each state every year a different type of Substance Name Drug Reports add and get Sum of DrugReports TotalDrugReports State will each state every year at the end of repeated data.

Then, we analyzed the data from these tables, and to understand the situation of synthetic opioid and heroin among the states, we first investigated the relationship between Sum of DrugReports and TotalDrugReportsState, and obtained the following figure:

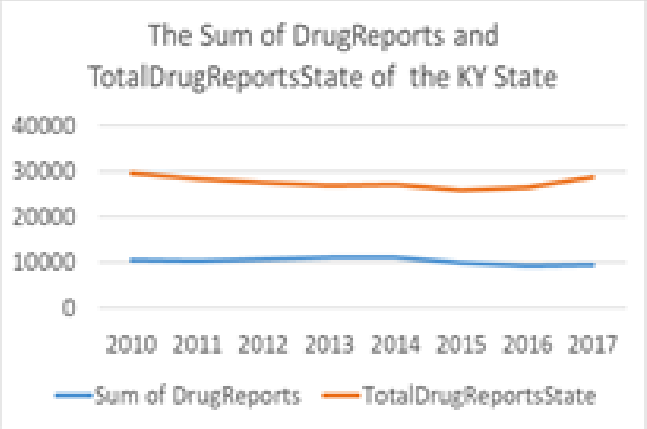

Fig. 1 KYstate.

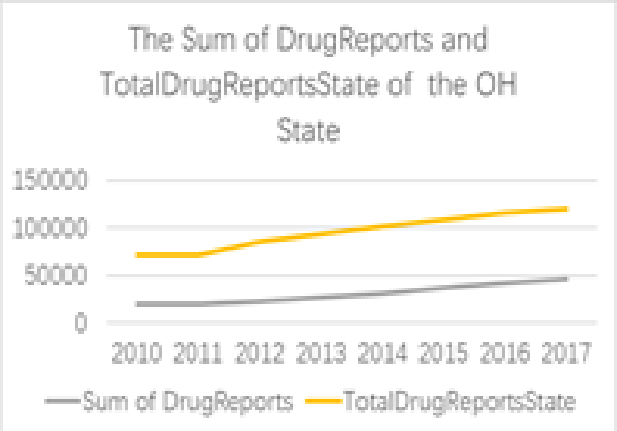

Fig. 2 OH state.

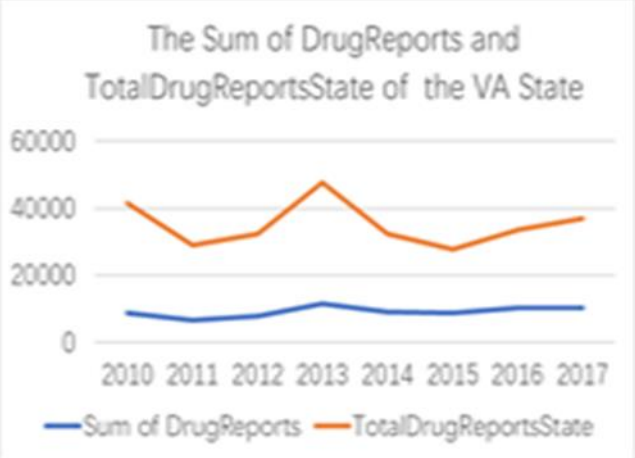

Fig. 3 VA state.

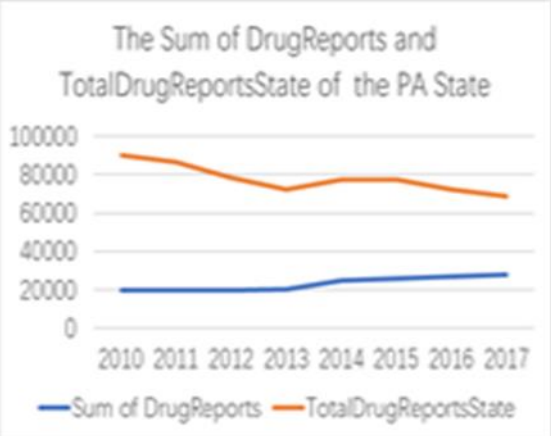

Fig. 4 PA state.

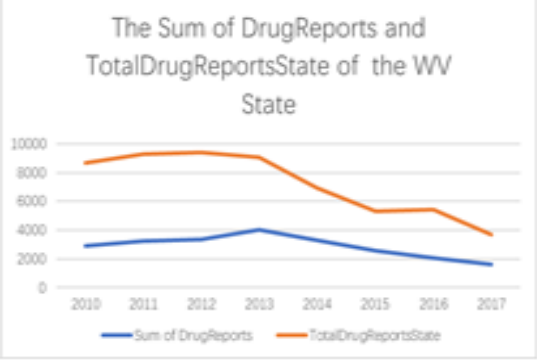

Fig. 5 wv state.

We can see that the change trend of Sum of DrugReports and TotalDrugReportsState is the same. Therefore, we can conclude that the change of TotalDrugReportsState causes the change of Sum of DrugReports. Therefore, we only need to analyze the relationship between Sum of DrugReports and year, and then build the model.

Therefore, we need to conduct curve fitting for the Sum of DrugReports and year. We first preprocessed the data, removed the outliers, and then conducted curve fitting. After with all kinds of curve fitting, we found that the Ozzie and Harriet state and exponential curve fitting, the remaining four states and linear fitting optimal, The results of the fitting curve are as follows:

Fig. 6 OH.

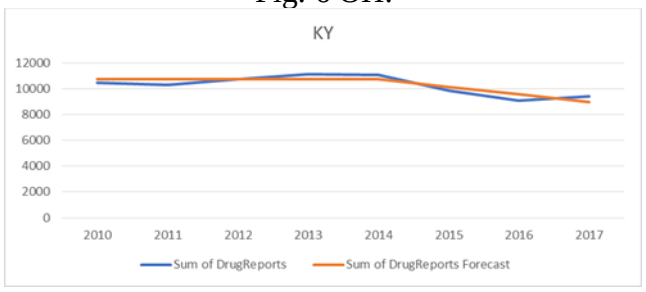

Fig. 7 KY.

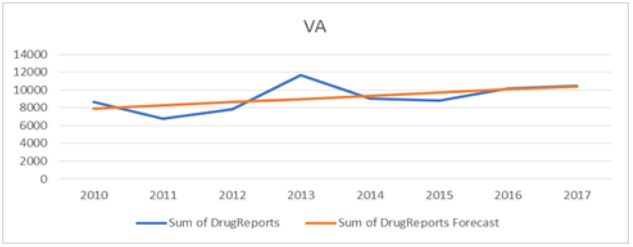

Fig. 8 VA.

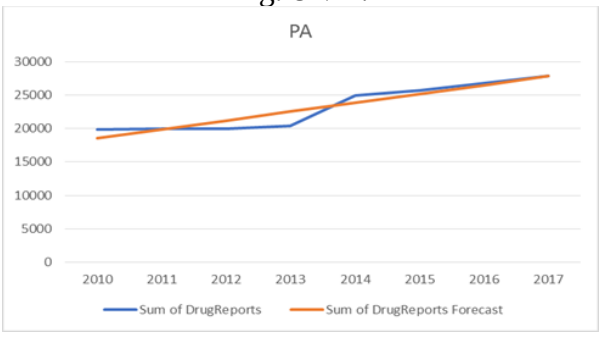

Fig. 9 PA.

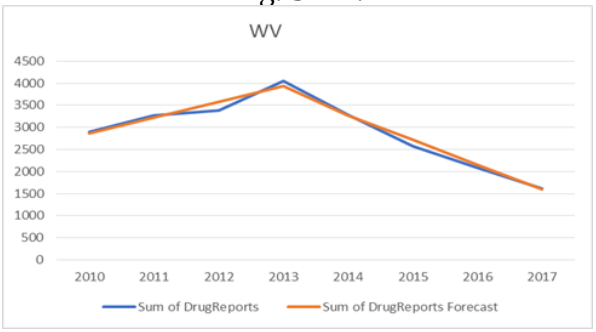

Fig. 10 WV

From the spread curve, it can be seen that in OH state, the usage of synthetic opioid and heroin increases exponentially, and the abuse of synthetic opioid and heroin is very likely. In VA and PA states, the usage of synthetic opioid and heroin increases linearly in recent years, and the abuse of synthetic opioid and heroin is more likely. Some of the heroin consumption, via synthetic opioid and heroin, decreased in KY and WV states.

3.2 Verification of County Model

Started specific opioids Because of the large number of counties, it is difficult to deal with, so we first have to split the data according to the state.

After that, we carried out k-means clustering for counties in each state. We designated a total of m classes to be clustered. According to the k-means clustering results, we obtained the clustering indexes (DrugReports, TotalDrugReportsCounty and TotalDrugReportsState) of the regions likely to have drug abuse, and used them as the threshold to screen the regional indexes likely to have drug abuse. The threshold value can be adjusted according to the actual size of m, so that the model has stronger practical application value. Here, we assume that m is 3, and take KY state as an example, then the clustering results of KY state are as follows:is HENRICO county.

Table 1 Final Cluster Centers of KY State

Final Cluster Centers

	Cluster		
	1	2	3
Sum of DrugReports(County)	57.4974795	1789.2464151	693.1663681
TotalDrugReportsCounty	1319.10654553	74359.46516040	23169.713268675
Persent(COUNTY)	.057456677316 1312	.024062120555 0690	.030643513223 4585

Table 2 Clustering in KY State

Number of Cases in each Cluster

Cluster	1	117.000
	2	1.000
	3	2.000
Valid		120.000
Missing		.000

From the results, we can clearly see that the spread of counties in KY state is divided into three categories, of which the first category and the second category are similar and can be combined into one category, the two categories are about 98% in total, so most counties in KY state are relatively good, which is also consistent with our analysis in KY state. The third category is the area most likely to suffer from drug abuse, accounting for about 2%. We took the third category of indicators as the threshold for screening, and the result showed that JEFFERSON county was the only county that met the criteria. Therefore, we can conclude that in the case of the threshold value of TotalDrugReportsCounty being 4833 and the threshold value of TotalDrugReportsState being 27647, the specific opioid region in KY state may have started to be JEFFERSON county.

And the same thing is true for the other four states when m is 3.

All the counties in OH state have a high possibility of drug abuse. Among them, ADAMS, ATHENS, BROWN and other 89 counties may have started specific opioid use.

Only 1 percent of the counties in PA state are at risk of drug abuse, while other areas are doing well, including PHILADELPHIA county, where specific opioid use may have begun.

The likelihood of drug abuse in counties in WV state is very low, and it is possible that specific opioid-like areas have begun to emerge for HARRISON county.

The situation in the VA state is similar to that in the WV state, and all the counties in the VA state are doing well, and one of the areas that may have started specific opioids is HENRICO county.

4. Conclusion

In this paper, we analyzed all opioid use in five states and their counties. First of all, we established the opioid and heroin class spread model of drugs at the state and its subordinate counties, and through the analysis OH states exponential curve, and PA in VA states into a linear growth, the rest of the two states are declining trend, also came to the states may have to start using specific opioid county, a total of 25 May. We also predict that the number of DrugReports in OH state will increase 5-10 times in 2026. This occurs when the OH state TotalDrugReportsState threshold is approximately 240000-250000.

References

[1] Penm J, Mackinnon N J, Connelly C, et al. Emergency Physicians' Perception of Barriers and Facilitators for Adopting an Opioid Prescribing Guideline in Ohio: A Qualitative Interview Study[J]. The Journal of Emergency Medicine.

[2] Icholas A. Trasolini, Braden M. McKnight, Lawrence D. Dorr. The Opioid Crisis and the Orthopedic Surgeon[J]. The Journal of Arthroplasty.

[3] Yang Yuhui, Xu Xiuli, Zhu Zhu. An overview of opioid abuse and its control measures in the United States [J]. China Drug Warning, 2017, 14 (12): 746-751.