# Stop the U.S. Opioids Epidemic!

## Summary

The opioids epidemic spreads to all cross-sections of the U.S. population, greatly influences the economy and health-care system in many ways, while being influenced by many social-economic factors at the same time. Here we are studying the intercounty distributions shift model to provide strategies based on the economic-social patterns to stop the opioids epidemic.

Our drug epidemic model contains two sub-models: the epidemic prediction model and the social-economical effection model.

The former one focuses on the synthetic opioids and herion drug reports. We measure the epidemic by the number of the drug reports of a drug in a county, and we apply the **vector autoregression model** and the **multivariate time series prediction model** based on the NFLIS data. We also figure out that Cuyahoga County in Ohio, Philiadelphia County in Pennsylvnia may be the possible starting of the opioids use.

The second one tries to find the correlationship between the social-economic factors and all types of opioid-report distribution. Data dimension is reduced by **model based ranking**. We use the features of 2017 to predict the drug reports that year to test our model. It truns out that those live alone may be more likely to take opioids drugs.

Finally, concluding from the results of the two models, we try to promote the first submodel and form its relationship with the socal-economic factors. We also propose an **advance and retreat strategy** to counter the opioids epidmic crisis.

Key Words: Multivariation time series prediction model; Model based ranking

# Contents

# 1 Introduction

## 1.1 Background

From 1999 to 2017, almost 400,000 people died from an overdose involving any opioid, including prescription and illicit opioids. [1]It is estimated in 2017 that over 130 people died from opioid-related drug overdoses in America a year. [2]

According to the Centers for Disease Control, epidemic is fuled by prescription pain medications, and those addicted to prescription pills may turn to cheaper and more avaliable street drugs like heroin and fentanyl. The doctors are intending to help patients with pain, but they may underestimate how addictive these drugs are and overestimate how helpful they could be when prescribing long-term opioids. Also, the opioid epidemic is the unintended consequence of the increased use and the acceptance of prescription opioids makes it worse.

Since composition of the reasons for the drug addicts is complex, we are looking into the drug reports and social-economic data from the five states in America, hoping to find the trend and provide effective strategies to control the crisis.

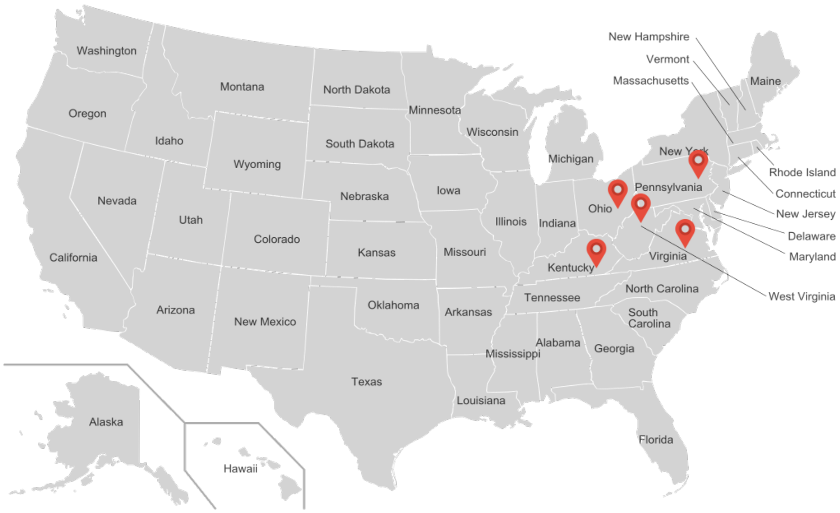


Figure 1: five states we focus on

### 1.2 Restatement of the Problem

In this paper we want to find a model to discribe the spread and characteristics of the opioids epidemic. Our main task is to predict the distribution of the opioids and heroin drug in country, and we start with the counties in the five chosen states. The social-economic factors are considered essential to the growth or shrinkage of the drug distribution, so they become the tool to do further predictions. Finally we want to offer the best strategy for the government to control the opioids epidemic based on what we conclude from the data.

## 2 Assumptions and Symbol Description

### 2.1 General Assumptions

1. The drug report recorded is non-repeated. That means the same addicted person will not be recorded twice in two years.
2. Since we could hardly predict a new substances of the opioids that does not appear in the NFLIS data, we assume that no new substances will appear.
3. The counties in the same state are more likely to spread in the same pattern.
4. The county could be considered the minimal unite of drug epidemic.

### 2.2 Notation

**Drug condition in county j** is what kinds of drug reports county $j$ have and the number of the reports.

**Distribution of the drug k** is the distribution of the drug containing substance $k$ in the counties.

### 2.3 Symbol Description

| Symbol | Description |
|---|---|
| $I$ | the set of the known states |
| $J$ | the set of the known counties |
| $K$ | the set of the kinds of drug reports |
| $J_i$ | the set of the known counties in state i |
| $O_{ij}^{Y}(k)$ | the total number of opioid drugs containing substance k reported in county j, state i in year Y, where $i$ is in $I$ and $j$ is in $J_i$ and $k$ is in set $K$ |
| $y_{j,t}$ | the total number of opioid-drug containing substance k reported in county j, state i in year t. Here y means $O_j(k)$ |
| $R_{ij}^{Y}$ | the total number of the drugs reported in county j, state i in year Y, where $i$ is in $I$ and $j$ is in $J_i$ |
| $C_i$ | the total number of the counties in state i, where i is in set $I$ |
| $rat_{ij}^{Y}(k)$ | the percentage of the total number of opioids drugs containing substance k reported by the total of the reported drugs in county j, state i, year Y: $j = 1; 2...C_k$ |

## 3 Model Overview

Our dynamic epdemic prediction model focuses on the sythetic opioids and consists of a self-predicting model and a social-economic influencing model. The first one aims at two goals:

1. Finding the variation of $O_{ij}^{Y}(k)$ with the same $Y$ by $k$, representing the static distribution characteristics of the reported opioids cases in and between five states.

2. Finding the variation of $O_{ij}^{Y}(k)$ with the same $k$ by $Y$, representing the timingly spreading characteristics of the reported opioids cases in five states through years.

Since the same drug addict will not be reported twice, $O_{ij}^{Y}$ could reflect the variation and we introduced the multivariate time series prediction model to do the prediction. While applying the equation to the second goal, the data is adjusted so that it could be trained better.

The second sub-model aims at finding the correlationship between the opioids using level and the social-economic level. We category the social-economic indicators and apply **model based ranking** to select the features. Finally we use several prediction models to test our selections and find the dealing social-economic features.

Then the two subs are merged together into our opioids epdemic and the merged model points out the advance and retreat strategy to face the opioids epidmic.

# 4 Sub-model I: Multivariate Time Series Drug Reports Difference Equation model

When considering the spread and characterstics of the reported synthetic opioids and hreoin incidents, there are many complicated reasons why the doctors would provide the patients with the opioids, and the difference equation model can consider the impact of different factors on the size of the independent variables, only requiring the initial data. For this reason, we woud not use other predicting methods.

## 4.1 Data Pre-processing

We want to find the data that represent the opioids epidemic. Considering drug conditions in counties, we ignore the counties with totally 1 or 2 drug reports appeared in the 8 years to focus on the main counties. Considering the distributions of the drugs, we ignore the drugs that only appeared in 2017.

Then as to the representitive of the drug, we choose the number of the drug reports in a county in a year. The smaller the unite we choose, the bigger the dataset is, the better result we would get. We once want to choose

$$rat_{ij}^{Y}(k) = \frac{O_{ij}^{Y}(K)}{R_{ij}^{Y}},$$

where $R_{ij}^{Y}$ comes from the $TotalDrugReportsCounty$, because it could reflect the intension of the doctor to prescribe better. However, the index may include other opioids substance than the drug reports we get, and the statical loss may be misleading. So in the end we take $O_{ij}^{Y}$ (also equals to $y_{i,t}$). as the index of a certain drug.

## 4.2 Data Predicting

We set the time step for the model is one year. In order to deal with the great number of the counties, we choose to apply the Vector autoregression model(VAR) for the time series model, where every variable is a linear function of all of its former values and all the other variables' former values, written as:

$$\begin{pmatrix} y_{1,t} \\ \vdots \\ y_{n,t} \end{pmatrix} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} + \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ \vdots \\ y_{n,t-1} \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

Here $e_i$ denotes to the errors; $a_{ij}$ is the weight by the former data; $c_{ij}$ is the const value; $y_{i,t-1}$ is the lag value of the opioid drug reports the year before, where $i = 1; 2; ...; |I|$. We use it because it could predict the number of a certain kind of drug report the next year year in a certain county based on all the drug report distribution we have got. The sensitive analysis of the weights matrix will be analyzed later.

## 4.3 Model Implement and Results

We predict the distributions of all the preidictable drugs in 2018. According to the predicted distributional data. We picture the distribution spreading map, and the most representive substances are Oxycodone, Hydrocodone, and Heroin, which are shown as the epidemic spreading maps in figure 2(intercounty distribution) and figure 3(interstate distribution).

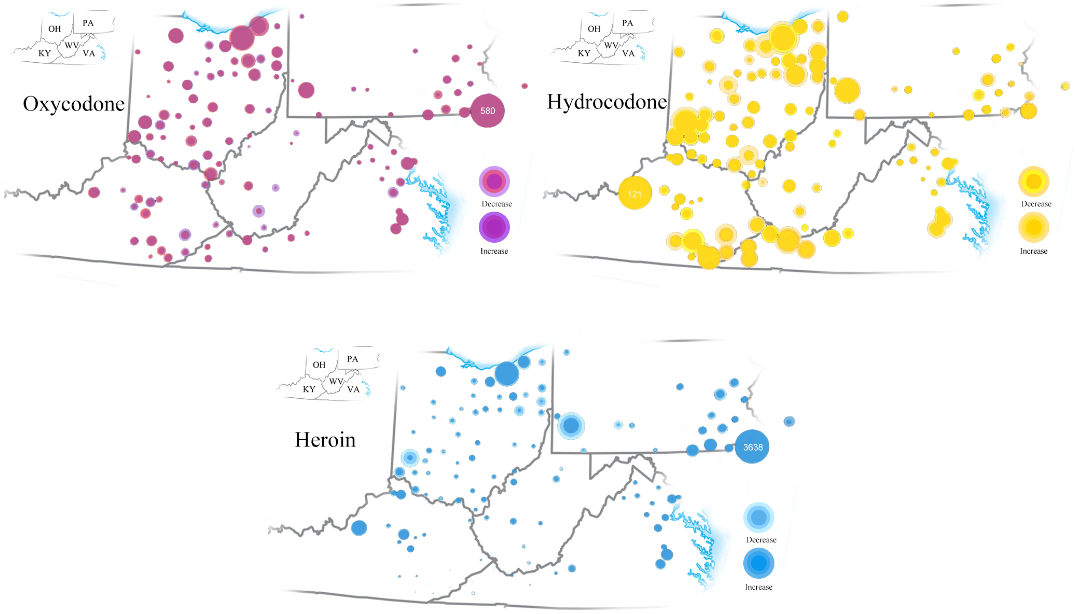


Figure 2: Epidemic spreading map of the top-3 in 2016, 2017, and 2018(predicted)

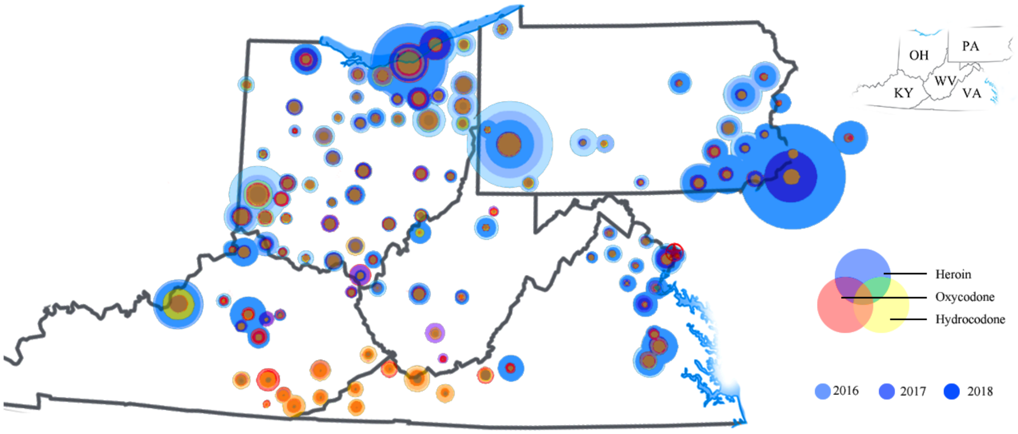


Figure 3: Combined Epidemic spreading map between States in 2016, 2017, and 2018(predicted)

From figure 2 we could see that the distribution remains steady in 2018 generally, with new counties hardly appearing. In Ohio, the county distribution tends to be clustering in the northeast and south west parts. In 2018, some counties whose numbers were small grew bigger:

- Fayette County, Kentucky
- Roanoke city, Virginia
- Richmond city, Virginia
- Scott County, Virginia

However, the trend in big cities is not so obvious.
Oxycodone, Hydrocodone and Heroin are the three most prescribed opioids drug, and the use of them continues to grow in most of the counties. Figure 3 illustrates the features of the interstate evolution of the drugs by taking more kind of drugs' county distribution into account. We could see the drugs reported in Ohio is the most dense one, and West Virgina the most sparse.

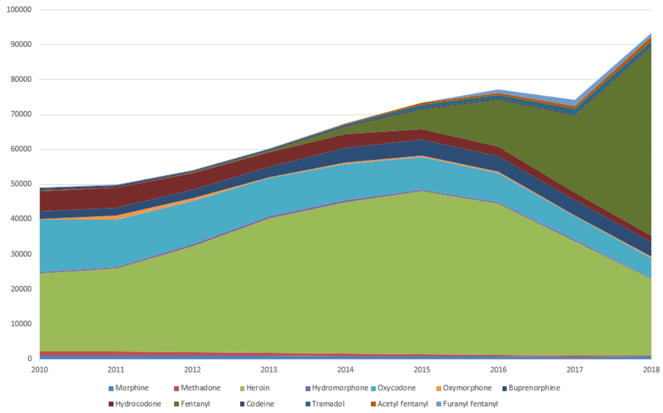


Figure 4: The total top-used opioids numbers variation from 2010 to 2017 and 2018(predicted)

Considering the timing variation in particular, as is illustrated in figure 4, in 2018 there may be a sharp increase in the use of Fentanyl, Hydrocodone, Metheadone, Tramadol, and Morphine, which has long been tops; However, the use of Heroin, which owns a large number, may go down in 2018. There are also "rising stars" like Acetyl fentanyl, which experiences a boom in number in 2016.

Concluding from the drug conditions and distributions above, the counties whose spread patterns of a particular opioid are of high correlationship are on the list of the possible spreading starts of the opioid. Also the counties with a large number of opioids already are suspected, too. We also notice there are counties like Raleigh County, West Virginia, whose number of Oxycodone decreased a lot form 2016 to 2017, but may experience an increase from 2017 to 2018. It may indicate that there a certain suppliers, or it is a producer itself.

Table 1: The table of the possible starts of the top-3 opioids

| Name of the Opioids | Possible Starts Counties | Possible Starts States | Reason |
|---|---|---|---|
| Oxycodone | Cuyahoga County | Ohio | A large number; clustering. |
| Oxycodone | Philadelphia County | Pennsylvania | A large number; clustering. |
| Oxycodone | Raleigh County | West Virginia | The number decreased from 2016 to 2017, but may increase from 2017 to 2018. |
| Hydrocodone | Laurel County | Kentucky | The number increased from 2016 to 2017, but may decrease from 2017 to 2018. |
| Hydrocodone | Cuyahoga County | Ohio | A large number; clustering. |
| Hydrocodone | Montgomery County | Ohio | A large number; clustering. |
| Hydrocodone | Gefferson County | Kentucky | A large number; seperated. |
| Heroin | Philadelphia County | Pennsylvania | A large number; clustering. |
| Heroin | Cuyahoga County | Ohio | A large number; clustering. |

# 5 Sub-model II: Social-economic Effection Model

In this section, we use the NFLIS data which was provided in the previous section together with the estimated data from the U.S. Census Bureau.

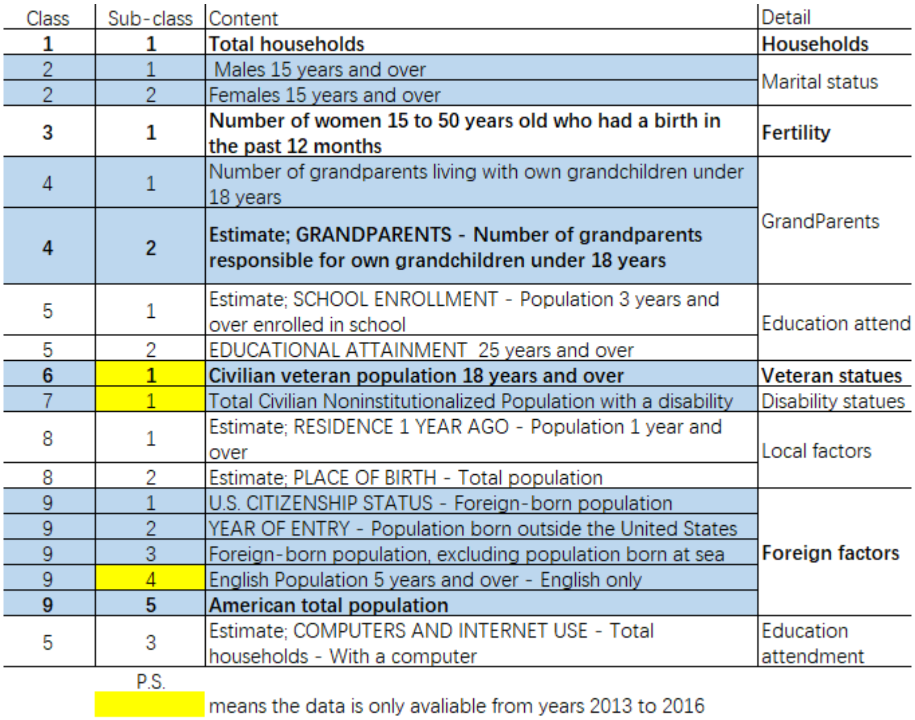
Since the features are all discrete and we want to minimize the loss to the least, we take the training results of randomforest model as the ranking indicator. The result is measured by explained variance score. As to the feature selecting methods in the randomforest model, we test several correlation indicators such as coefficient of determination and mean decrease impurity, and measure by both the maximum values and the average values to avoid the uncertainty of the algorithm. The results are shown in figure 6. The bigger these correlation index is, the better the correlationship the features have with the number of the drug reports. Here we are applying the same method above to have three rounds of selections.

## 5.1 First-round: Rough Selection through Years

According to the data from the World Drug Reports [3] and the data provided, there are 596 possible features but only 460 samples. The sample size is not big enough to find the correlationship among the features. However, we find there are indicators whose concepts could lead to a simpler division. Therefore, we choose them as it is listed in figure 5. We want to find the correlationship between these classes first and then study in more detail. The correlation results are shown in figure 6.

## 5.2 Second-round: Refined Selection through Years

Comparing the charts, we could see sub-classes 1,3,6,9 outstands, representing *Households by types*, *Fertility*, *Veteran statues*, and *Foreign factors*. We then add more features belonging to these classes to do deeper research in order to find more correlationship.



| Class | Sub-class | Content | Detail |
|---|---|---|---|
| **1** | **1** | **Total households** | **Households** |
| 2 | 1 | Males 15 years and over | Marital status |
| 2 | 2 | Females 15 years and over | |
| **3** | **1** | **Number of women 15 to 50 years old who had a birth in the past 12 months** | **Fertility** |
| 4 | 1 | Number of grandparents living with own grandchildren under 18 years | GrandParents |
| **4** | **2** | **Estimate; GRANDPARENTS - Number of grandparents responsible for own grandchildren under 18 years** | |
| 5 | 1 | Estimate; SCHOOL ENROLLMENT - Population 3 years and over enrolled in school | Education attend |
| 5 | 2 | EDUCATIONAL ATTAINMENT 25 years and over | |
| **6** | **1** | **Civilian veteran population 18 years and over** | **Veteran statues** |
| 7 | 1 | Total Civilian Noninstitutionalized Population with a disability | Disability statues |
| 8 | 1 | Estimate; RESIDENCE 1 YEAR AGO - Population 1 year and over | Local factors |
| 8 | 2 | Estimate; PLACE OF BIRTH - Total population | |
| 9 | 1 | U.S. CITIZENSHIP STATUS - Foreign-born population | **Foreign factors** |
| 9 | 2 | YEAR OF ENTRY - Population born outside the United States | |
| 9 | 3 | Foreign-born population, excluding population born at sea | |
| 9 | 4 | English Population 5 years and over - English only | |
| **9** | **5** | **American total population** | |
| 5 | 3 | Estimate; COMPUTERS AND INTERNET USE - Total households - With a computer | Education attendment |

P.S.

🟨 means the data is only avaliable from years 2013 to 2016

Figure 5: Social-economic feature-division by 9 classes/ 17 sub-classes

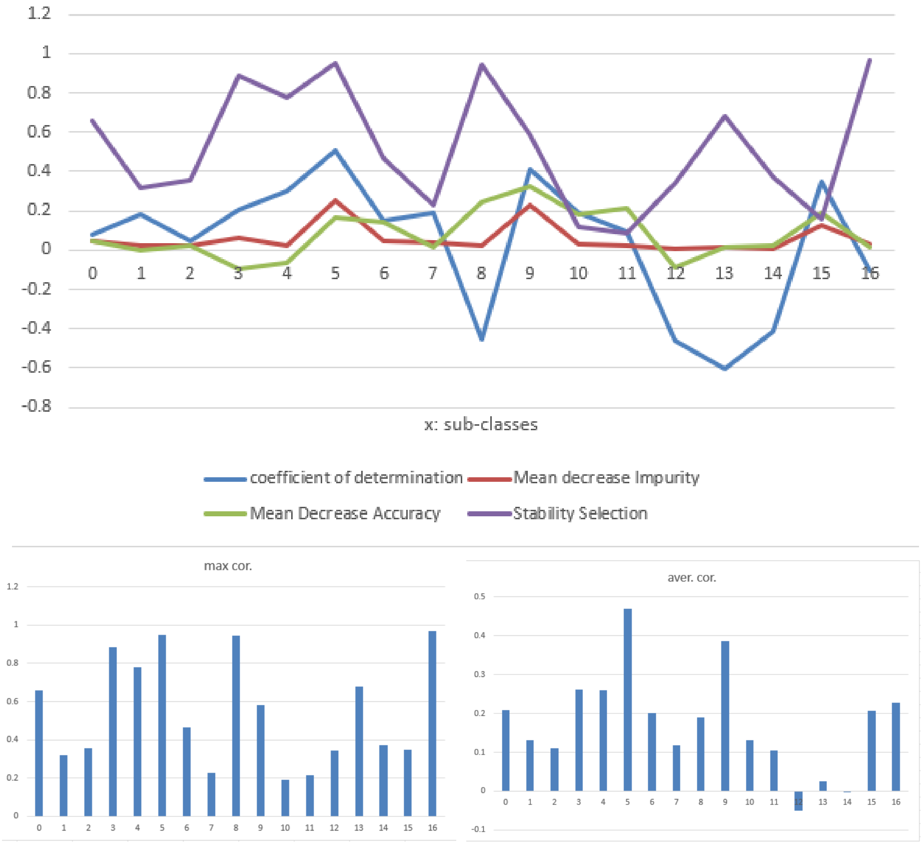


Figure 6: Feature correlations measured by different algorithms for sub-classes

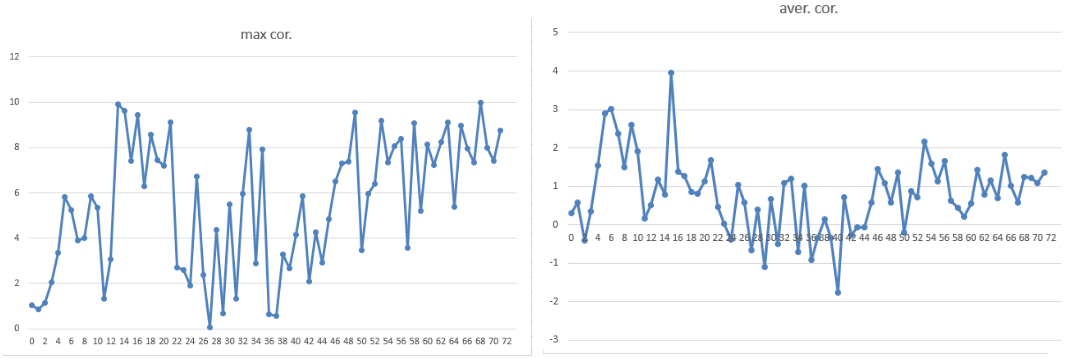


Figure 7: Feature correlations measured by different algorithms for sub-classes' subs

With all the correlation above zero, the categories all belongs to the *Households by types*; with the average correlation above 2 and the maximum correlation above 5, three categories belongs to the *Households by types* and one belongs to *Foreign locations*. Particularly, *male householders with no wife present with own children under 18 years*, *female householders with no husband present* and *Householders living alone* satisfies both of them. So they may be the key features.

## 5.3 Third-round: Rough Selection Based on years

Now we hae the possible features which could be the key social-economic factors for the drug use variation from 2010 to 2016. However, we want to konw whether the possible dominant factor would change in a particular year. Therefore, we choose 4 out of result of the first-round selection as the key features every year with the model based ranking method.

The selected features clusterd in class 1(*Households by types*) and class 3(*Fertility*). The unmarried women who have given a birth are full-time present on the list, and those living alone are shifting by the years.

On the other hand, we also choose 12 out of 72 in the second-round selection as the key features every year with the model based ranking method. The most possible features shift from class 1(*Households by types*) to class 9(*Foreign factors*) by 2012. The single households are be the leading feature in class 1, including the male households with no wife present and the female households with no husband present. Those living alone are also on the list. From 2012 to 2016, the *Foreign factors* take the lead and the data clustered in the category *Ancester*. This result is quite surprising and differs from the previous rounds.

## 5.4 Model Testing

To test our chosen features in the previous three rounds, we use the ensemble learning method to predict on the chosen data. We esemble the dicision tree models and applied several forms of it, including gradient boosting, random forest, booststrap aggregating, decision tree learning and K-Nearest-Neighbours.

In order to compare the performance of the prediction models in the three rounds, we import two indicators: the mean-square error and explained variance score.

Mean-square error(MSE) is applied as:

$$MSE_t = \frac{1}{|J|}\sum_{j=1}^{|J|}(y_{j,t} - \bar{y_t})^2, \bar{y_t} = \frac{1}{|J|}\sum_{j=1}^{|J|} y_{j,t}.$$

The smaller the MSE is, the better the model performs. The **explained variance score** explained the variation measuring the proportion to which our models account for the variation of the data based on. The bigger it is, the better the model performs. It is applied as:

$$explained\_variance(y_{j,t}, \hat{y_{j,t}}) = 1 - \frac{Var\{y_{j,t} - \hat{y_{j,t}}\}}{Var\{y_{j,t}\}}$$

The comparison is shown in figure 8 and 9. The solid line(dotted line) in the charts of explained variance score is determined by the results in the second-round(the third round) minus that in the first round. The solid line(dotted line) in the charts of mean-squared error is determined by the results in the first round minus that in the second-round(the third round).

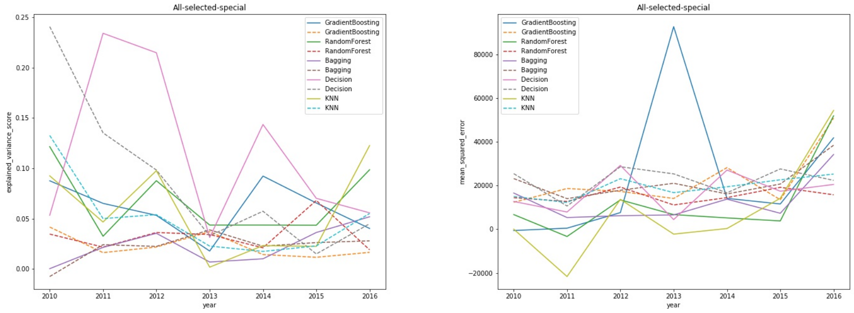


Figure 8: third-round v.s. first-round

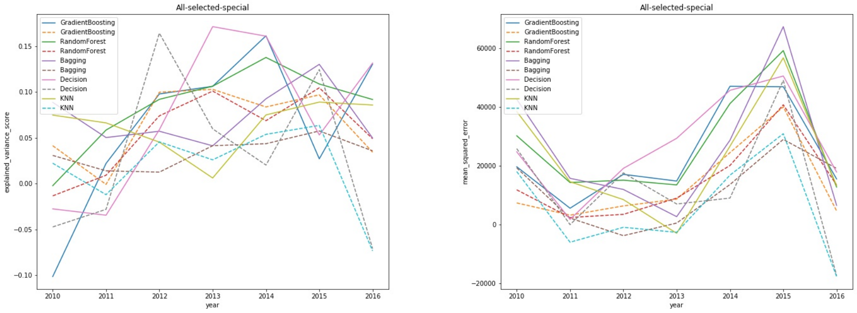


Figure 9: second-round v.s. first-round

From the line chart we could see factors selected in the second round could better at defining the features. It wins completely by the first round, and even if the leading factors may vary from classes in years, the features it chosen remains important.

### 5.5 Conclusion for the Social-economic Effection Model

After the three-rounded feature selecting, we come to the conclusion that those living alone or living with their spouse unpresent are more likely to get prescribed opioids than others. While the total number of men taking opioids drugs is larger than that of women, we find that the fertility of women is also important. This mainly reflects in the young unmarried women who have given a birth in a year.

In order to find the association with the trend across the states, we have a map of the male ratio in the five states in figure 10. Compared with the drug epidemic map in the previous section, we could say there is a greater possiblity that whether one is single or not is an important factor.

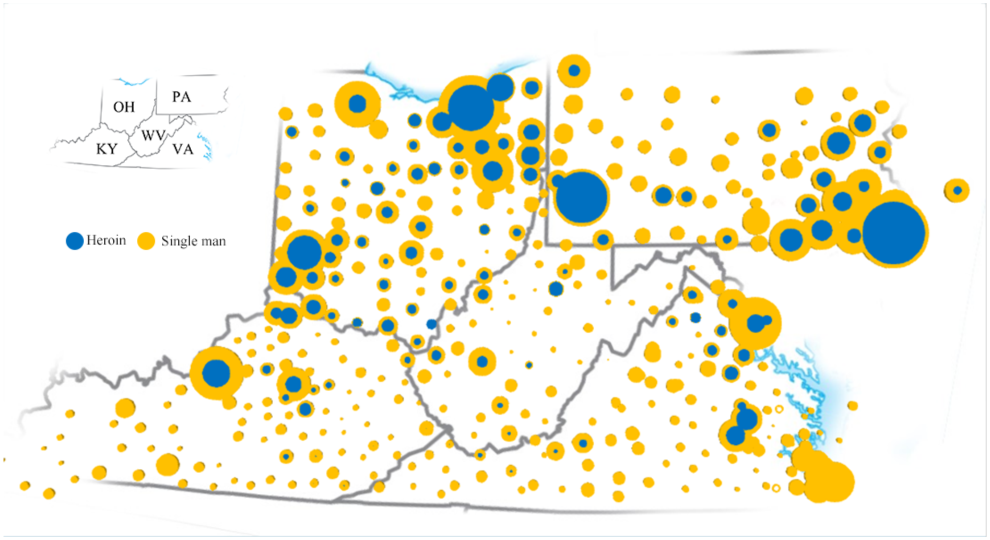


Figure 10: Distribution of the male householders with no spouse presenting in the U.S. constrasting with the distribution of the Herion in 2016

## 6 Model Merging

By implementing the model, we hope to get a value of a particular feature of a county by training with all the features of the given counties other than that in the same year. Take the number of the female households in Cuyahoga, Ohio, 2016 as an example. First we train the predicting model with randomforest, which has been proved to have the best performence in the previous section, with the data from the other counties in the state. (As we assume, the counties in the same state would have a more obvious influence). Then we set a series of paraments as the possible values of the total femal households, and we use the prediction function to predict the value of the total number of the female households in Cuyahoga, Ohio, 2016. We measure the predicted value with the true value to adjust the paraments. When the paraments falls in to a particular range, the true value would be always below the predicted value. Then the parament bound is found: When the true value of the female households in Cuyahoga, Ohio, 2016 fall in the bounds, the number of the Heroin in Cuyahoga, Ohio, 2016 must decrease.

However, while predicting, we do not have the social-economic data of the other counties in the future. So when we use it to predict the conditions in 2018 , we still choose the data in 2017 to train the prediction functions.

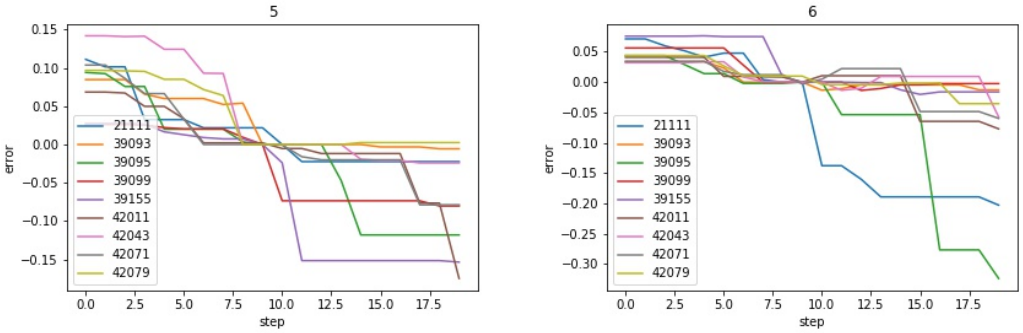


Figure 11: 5:Male householder with no wife; 6:Female householder with no husband present

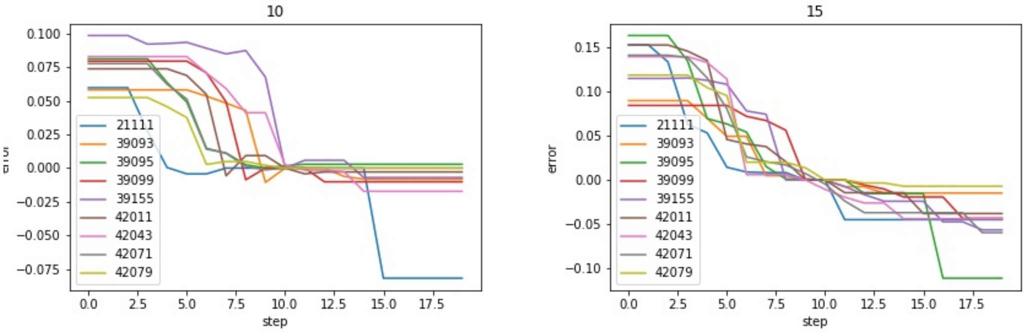


Figure 12: 10: Householder living alone,65 years and over; 15: Number of women 15 to 50 years old who had a birth in the past 12 months

## 6.1 Model Testing and Explanations

We propose to limit the value of the *Female householder, no husband present*, *Male householder with no wife present*, *Householder living alone* and *Number of women 15 to 50 years old who had a birth in the past 12 months* for our merged model. Then we used the data in 2016 to test the results.
As is shown in figure 12, we could conclude the future opioids consition of drug by adjusting parameters concerning category 5,6,10 and 15. The error is measured by

$$\frac{predict(y_{i,t}) - y_{i,t}}{predict(y_{i,t})}$$

So all the four categories shows a negative correlation. Every colored line represents a city. Step of paraments are measured by ratio, according to the sizeof the particular feature chosen(here varying from 5,6,10,15). And the way we calculate the step leads to the lines would cluster near the error zero, witch means "no changes happen".
Surprisingly, in category 10 we could find a sharp discrease along with fluctuation when getting close to the mid-point. That indicates when the parameters are near to the interval [6.0, 7.5], there will be a together decrease in use, which is a best time to control the use of the drugs; when the parameters are near to the interval [7.5, 10], the number of the drug reports would increase with the parameter increase.

## 6.2 Advance and Retreat Strategy

According to the research in the previous sections, we could find the spread patterns of the opioids epidemic in time and space with the first sub-model, and with the help of the second sub-model, we could see the change of the opioid in numbers by county. In order to control the epidemic, we could control the paramenters in the second model to reduce the total number in a county, which is called Advance Strategy; at the same time, according to the first model, we could predict the places where could be the starts of the opioids and we should let them retreat. That is the so-called Retreat Strategy.

# 7 Conclusion

We build a model that could predict the spread of the opioids epidemic with social-economic factors. Figure 11 is our flow chart.

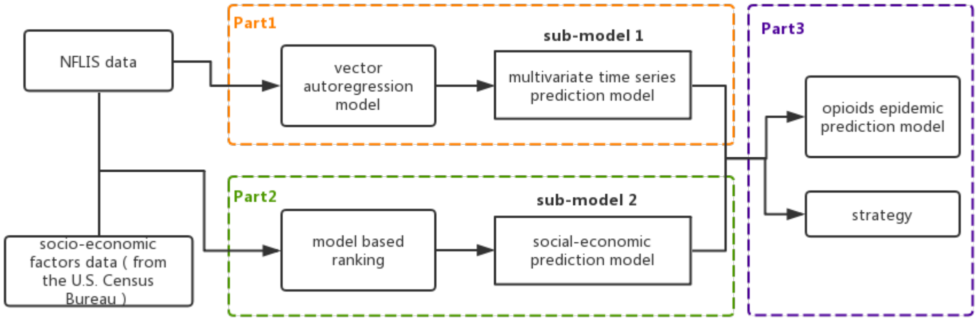

NFLIS data
Part1
vector autoregression model
sub-model 1
multivariate time series prediction model
Part2
model based ranking
sub-model 2
social-economic prediction model
socio-economic factors data ( from the U.S. Census Bureau )
Part3
opioids epidemic prediction model
strategy

Figure 13: Flow chart of the opioids epidemic model

After adjusting our model and determining the influence of possible influencing factors, we finally propose our strategy for the government to control the use and sell of opioids in the cities where there is already a large amount of prescribed opioids drugs reported, such as Cuyahoga County, Ohio, Philadelpha County, Pennsylvania and Montgometry County, Ohio. Also the government should pay attentions to cities like Raleigh County, West Virginia and Laurel County, Kentucky because of their contrasting growing partterns with the counties around.

As to the social-eonomic factors, since the households with no spouse present would under a larger possibility of taking the prescribed opioids drugs, we propose there could be more services to improve the singles' sense of security and happiness index.

# 8 Strengths and Weaknesses

## 8.1 Strengths

In our model, the input data source is trusted and adequate for predicting the five states' spreading prediction. We make full use of tha social-economic data, divide them and merge again to find the patterns begind. We consider the time difference while building the social-economic effecting model. That is, at different period of time, the dominant factors may be different. Besides, we do full error analysis by seperating the training set and dataset by 0.9,0.8,0.7,and 0.6, and take the average result in the end.

## 8.2 Weaknesses

The social-economy data may not be enough to do the further predictiing.The way we measures the geographic diatributions is rough and the method to find the possible starting could be more programmatic. And our second sub-model did not consider households' relationship data. We have no sensitive analyze because we could not get the social-economic data to predict the number of the county drug reports in 2017.

# References

[1] Scholl L, Seth P, Kariisa M, Wilson N, Baldwin G. Drug and Opioid-Involved Overdose Deaths – United States, 2013-2017. WR Morb Mortal Wkly Rep. ePub: 21 December 2018.

[2] https://www.hhs.gov/opioids/about-the-epidemic/index.html

[3] https://www.unodc.org/wdr2018/

[4] ERDOS,P.,R.ENYI,A.On the evolution of random graphs.Publ.Math.Inst.Hung.Acad.Sci,1960,5:17-61.

[5] Opportunities and challenges of complex networks [EB/OL].https://blog.sciencenet.cn/blog-3075-719543.html,2013.

# 9 Appendix