For office use only	Team Control Number	For office use only
T1 _____	**1909434**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**C**	F4 _____

2019
MCM/ICM
Summary Sheet

Summary

In the paper, we aim to explore and interpret drug data from five states, make predictions and recommend possible strategies through mathematical models.

Firstly, according to the characteristics of different opioids, we divide all reported opioids into three categories: heroin, fentanyl and its derivatives and other drugs. We use data visualization to depict the drug use profile, including quantitative and geographical trend of opioid use, through which we make some useful inferences to help modeling and make thorough analysis on internal rules of the provided data.

The next step, we build a modified vector autoregressive(VAR) model to describe the process of opioid drug use variation from a view of time series. We apply our model to backcast possible opioid use conditions in past years and then identify possible origin locations of a specific opioid use, finding that fentanyl might have started in Ohio and Pennsylvania. We then predict the future opioid use. The result shows that the degree of heroin use will decrease at a rapid speed in all 5 states, and the total number of heroin reports will be less than 10000 in 2019, and there will be only one county in Ohio above the identification threshold we set. Fentanyl and its derivatives abuse will be further severe in 2018, especially in Pennsylvania and Ohio. What's more, the fentanyl abuse is getting worse. The total number of fentanyl and its derivatives report will be more than 66000 in 2018 and 110000 in 2019. The fentanyl and its derivatives use has already exceed the reasonable identification threshold in most counties in five states. As for other opioid drugs, the density of use is slowly increasing, but the degree of abuse of these drugs will not rise to a severe level in a short time.

After proper and precise data preprocessing, we apply decision tree and correlation coefficient to explore the correlation between opioid use and socio-economic data. We list top categories of attributes that have linear correlation with opioid use, including educational attainment, marital status, etc. We also list top categories of attributes that have non-linear correlation with opioid use, including fertility, veteran status, etc. We also identify the opioid use's correlation with subdivision or sub-subdivision categories of socio-economic attributes. We modify our model by applying decision tree, which can slightly correct the prediction given by our first model.

Then we propose a possible strategy, which includes strengthening prohibition against illegal drugs, enforcing road transportation management and control by State Highway Patrol, advertising drug use education in campus, being more careful about prescribing addictive drugs and so on. We respectively test our strategy from overall and local perspectives, finding that some actions are effective for reduce opioid use, and it's usually more effective to take multiple actions cooperately. Finally we make detailed analysis on parameters of our model.

Keywords: **Vector Autoregressive Model; Data Visualization; Correlation Coefficient; Decision Tree**

Contents

Memo

To: Chief Administrator, DEA/NFLIS Database
From: Team # 1909434
Date: 28 January 2019
Subject: Opioid Use Profile and Recommended Stategies in 5 States

---

Dear Chief Administrator, we are honored to inform you our summary and recommendation after data analysis and modeling.

Opioid Drug Use Profile

Here is the opioid use profile in Ohio, Pennsylvania, Virginia, West Virginia and Kentucky:

In general, heroin use increased rapidly from 2011 to 2015 and then decreased since 2015. Fentanyl use started to increase since about 2013 and the increasing speed was fairly drastic. Other opioids use didn't show drastic variations during 2010-2017. Thus, we infer that the fentanyl and its derivatives became a substitute for heroin. Quantitatively, without policy interventions, heroin use may continue to decrease and fentanyl use may continue to increase while other opioids use keep stable or decrease slowly in the future. Another notable find is that opioid drug use accounts for an increasing proportion of total opioid use.

Opioid abuse is a more serious problem in Ohio and Pennsylvania than other three states, and the number of drug reports is increasing rapidly in Ohio while decreasing in Pennsylvania. The variation trends of Kentucky, Virginia and West Virginia are relatively stable. The number of opioid reports in Kentucky almost equals that in Virginia although Virginia has nearly twice the population of Kentucky.

As we predict, the degree of heroin use will decrease at a relatively rapid speed in five states, and the total number of heroin report will be less than 10000 in 2019. However, fentanyl abuse is getting worse. The total number of fentanyl and its derivatives report will be more than 66000 in 2018 and 110000 in 2019. As for other opioid drugs, the use density is slowly increasing, but the degree of abuse of these drugs will not rise to a severe level in a short time.

Recommended Stategies

Our analysis also shows that opioid use is related to several socio-economic factors. Here we give our recommended strategies based on important socio-economic attributes identified.

- Strengthen the prohibition against addictive opioids like fentanyl(and its derivatives) in five states.
- Enforce road(to Virginia, West Virginia and Kentucky) transportation management and control by State Highway Patrol to contain the spread of opioid drugs.
- Turn attention to fighting against fentanyl and its derivatives compared with other drugs because fentanyl compounds are fairly potent and fentanyl use are drasticly increasing.
- Further popularize education for all citizens to reduce the number of school dropouts and strengthen drug addiction prevention and control education in campus.

- Regularly hold community anti-drug propaganda and experience exchanging meetings.
- Social welfare agencies should focus on children without parents' caring, and give them appropriate psychological counseling and economic support if necessary.
- Any individual or institution with qualification to prescribe should be careful about prescribing addictive opioid drugs.

1 Introduction

1.1 Background

Since the intoxication of opium was first discovered thousands of years ago, humans have consumed opiates to treat pain and to get high[1].

Drug abuse has been a severe issue in the U.S.. Being that some drugs are illicit and most people are not very open to talk about their drug habits, the degree of drug abuse in America may be more severe than what statistics shows. In addition to traditional analgesics, some new painkillers such as fentanyl are more effective and addictive. A extremely dangerous example is carfentanil, also known as Elephant Tranquilizer. Carfentanil is about 10000 times more potent than morphine [2]. More and more individuals tend to use drugs which are more potent for the purpose of recreation or relieving pain. In a word, drug abuse is an urgent problem to be solved.

1.2 Problems Restatement

To suppress the spread of opioid drug abuse, it's necessary to study the internal rule of this phenomenon. We are going to finish following tasks based on given data.

1. We visualize our data to make a opioid drug use profile, in which we find some internal rules and make some inferences to help modeling.

2. Build a proper model to describe the spread between counties and states, and reflect the characteristics of opioid drug use in five states. Our model has following features:

- Taking the county as the unit, the opioid drug use data of a county and other counties is added into the model to predict the opioid drug use data of the next year of this county.
- The distribution of opioid use at different times can be calculated, including predicting the future and backcasting the past. Our model can show the trend of opioid drug spread over a period of time.
- Reflect the relationship between the changes of the quantity of various opioid drugs.

3. On the basis of the above model, the influence of social and economic factors on the distribution and spread of opioid drugs is further considered. The attributes that have significant impact on opioid drug use are given. Then add socio-economic factors to our model to make the model more accurate.

Then we propose possible strategies for countering the opioid crisis and make thorough analysis on them. We also test the effectiveness for several strategies and make sensitivity analysis towards parameters in our model. Finally we list the strengths and weaknesses of our modeling.

2 Assumptions

- All opium drugs use except heroin and fentanyl and its derivatives follow the same trend, which is proved to be reasonable in drug use profile in Chapter 3.

- Opioid drug use spread is related to the distance between counties and states and the number of drug reports of every county.
- Potential policy intervention and any competitive new drugs that may emerge in the future are not taken into consideration.

3 Opioid Drug Use Profile

In this part, we will visualize the original data, then find the inherent rules of the data. Some inferences or conclusions will be used for modeling, which will make the model more reliable.

3.1 Quantitative Trend of Drug Use

MCM_NFLIS_Data.xlsx contains opioid drug identification counts in years 2010-2017 for narcotic analgesics(synthetic opioids) and illicit drug heroin in each of the counties from five states. In order to identify opioid drugs which varies greatly from year to year, we draw a line chart to show usage changes of each drug(left). Having noticed that there are many fentanyl-related compounds and derivatives such as acetyl fentanyl and carfentanil(an extremely potent drug), we combine all fentanyl-related compounds as one line and get a more intuitive figure(right).

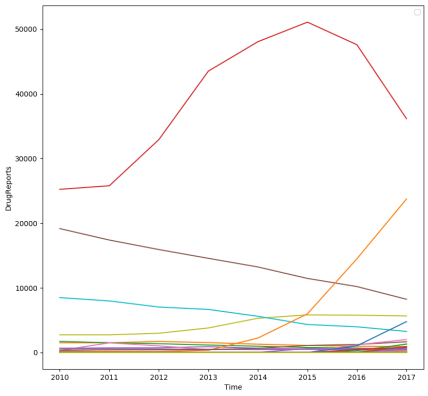

Figure 1: Opioid Drugs Use Variations

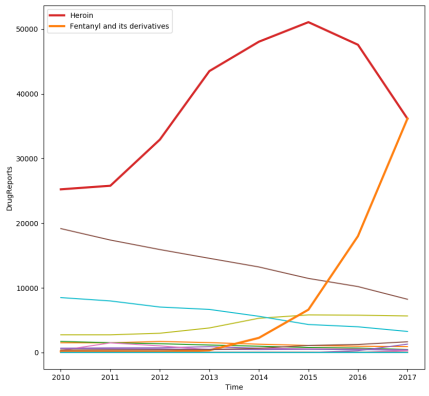

Figure 2: Combine Fentanyl-related Compounds

According to the line chart, we can find:

- Heroin(the top line) use increased rapidly from 2011 to 2015 and then decreased since 2015.
- Fentanyl(the second top line) use started to increase since about 2013 and the increasing speed was getting faster. Its development is likely to be described by $g \circ f$, where g is an exponential function and f is another primary function.
- Other opioid drugs use didn't show obvious(or drastic) variations during 2010-2017.

Thus, we speculate that the decrease of heroin use has correlation with the increase of fentanyl use. Meanwhile, other drugs use variations have few correlation with heroin and fentanyl use. Quantitatively, without policy interventions, heroin use may continue to decrease and fentanyl(and its derivatives) use may continue to increase while other drugs use keep stable or decrease slowly in the future.

The figure below describes variations of total drug reports, opioid reports and ratio. It's noteworthy that the ratio of total drug reports to opium reports was generally increasing from 2010, which indicates that more and more people tend to use opioid drugs.

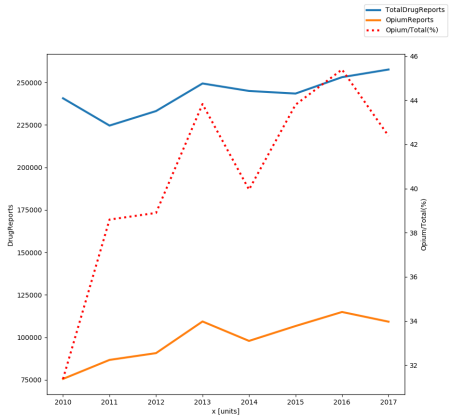

Figure 3: Variations of total drug reports, opioid reports and ratio

3.2 Geographical Trend of Drug Use

Firstly, we draw a line chart to describe variations of total drug reports in each of the five states. As the figure shows, the number of total drug reports in Ohio increases in a linear trend approximately. The number of total drug reports in Pennsylvania fluctuates through years and shows a slightly downward trend, while the number of drug reports in other states are relatively stable.

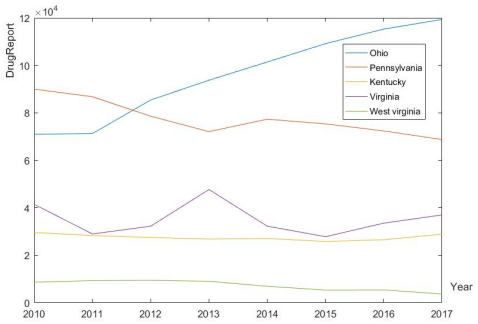

Figure 4: Variations of Total Drug Reports in Each State

Based on the U.S. map and MCM_NFLIS_Data.xlsx, we can draw sketch maps which indicates density variations of opioid drug users during 2010-2017. Because of diversity of variation characteristics, we divided density variations into three parts:

- density variations of heroin users

- density variations of fentanyl(and its derivatives) users

- density variations of other opioid drugs users

Corresponding to these three parts, we draw three images to show the density variations of opioid drug users geographically.

As for heroin users, the user density of two northern states(Ohio and Pennsylvania) is significantly higher than that of three southern states(Kentucky, Virginia and West Virginia). The increasing speed of user density of three southern states is obviously higher than that of two northern states. It can be inferred that both internal factors and influence of two northern states caused the rapid increase of heroin user density in the three southern states.

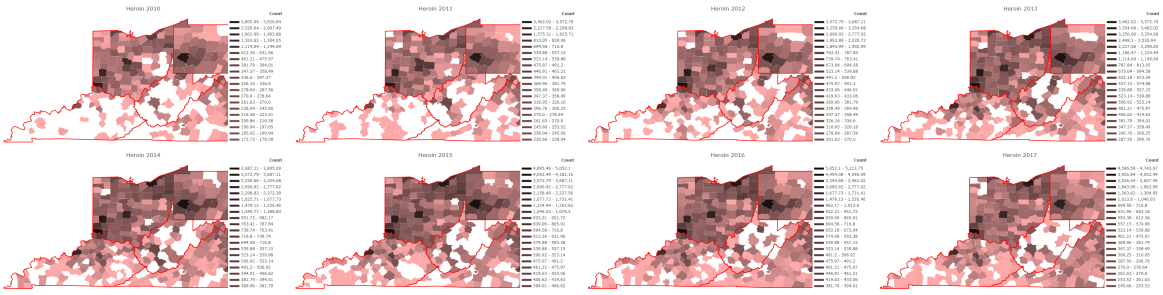

Figure 5: Variations of Heroin User Density

As for fentanyl(and its derivatives) users, neither of these five states had a high user density before 2014. However, the situation sharply turned to be severe since then.

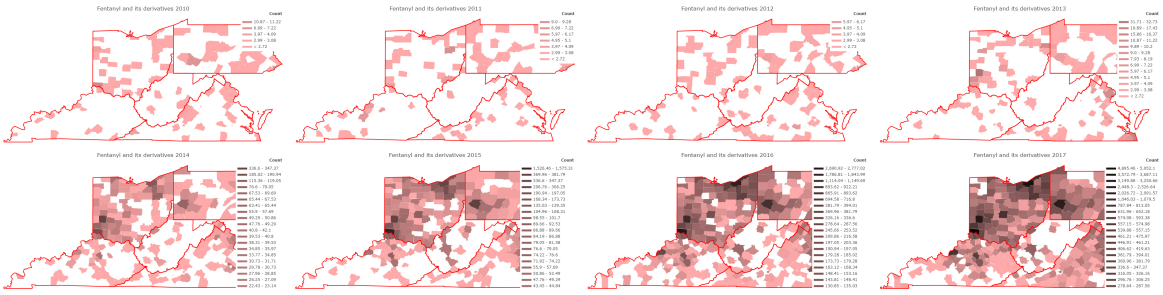

Figure 6: Variations of Fentanyl(and its derivatives) User Density

As for other opioid drugs users, the density variation tendency is relatively stable. More specifically, if you look carefully at the legend of every small picture on the right, you will find the user density was slightly decreasing from year to year. A reasonable guess is that some people abandoned the use of these drugs out of fear of addiction, and some became to use heroin or fentanyl compounds.

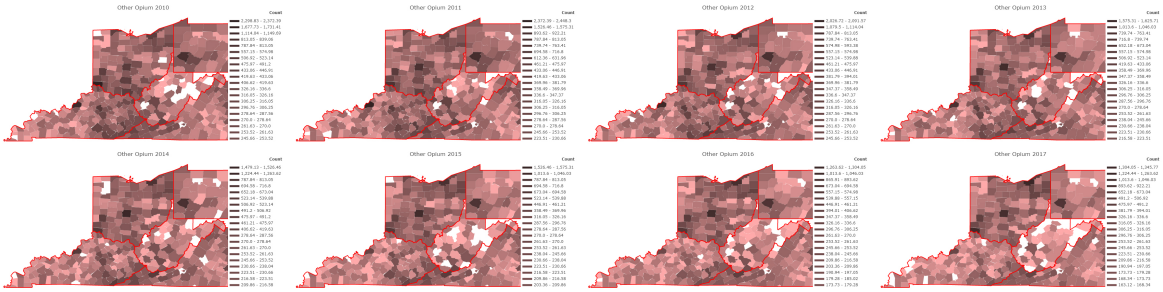

Figure 7: Variations of Other Drugs User Density

4 Part 1: Modeling, Backcasting and Prediction

4.1 Data Preprocessing

In this chapter we will bulid a modified multivariate AR(n) model. Thus, before we process this model, we are supposed to obtain stationary time series by differencing.

In Section 4.2, there will be three time series family denoted as $\{H_t^i\}$, $\{F_t^i\}$ and $\{O_t^i\}$ referring to variations of reports of heroin use, the sum of fentanyl and its derivatives use and other opioid drugs use relatively. From Figure 1 and Figure 2, we analyzed the characteristics of variations(details are in last chapter). Now we use different method to differentiate these three time series in order to obtain stationary time series.

- $\{F_t^i\}$: take logs to get a series with a linear trend, then difference to get a stationary series.

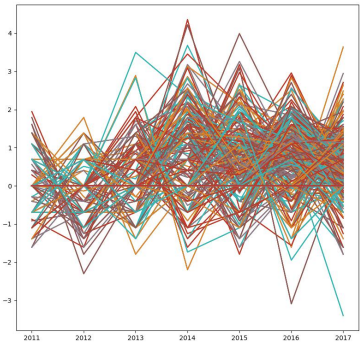

Figure 8: $\{F_t^i\}$ after taking logs and differencing

- $\{H_t^i\}$: difference one time; $\{O_t^i\}$: already stationary, no need to difference.
- Each i corresponds to $county_i$.

4.2 Modified VAR(n) Model

Traditional models to describe spread rule and predict include SI epidemics model based on ordinary differential equation and birth-death process based on stochastic process. However, these models priorily set the law of development, ignoring the information that our source data indicates. We should bulid a model which restore the rules

hidden in the data itself, so we will find another way.

The excel table contains opioid drug use reports for each county and state from 2010 to 2017, so we regard it as time series data. We will model based on elementary theory of time series analysis. The first method we think of is AR(n), which regards the sum of all opioid drugs reports as one time series. However, as mentioned in last chapter, there is diversity of variation characteristics between those three kinds of opioid drugs, so we set three times series family to represent the variation of these three kinds of drug use.

- Time series family 1: variations of reports of heroin use, denoted as $\{H_t^i\}$.
- Time series family 2: variations of reports of the sum of fentanyl and its derivatives use, denoted as $\{F_t^i\}$.
- Time series family 3: variations of reports of other opioid drugs use, denoted as $\{O_t^i\}$.
- Each i corresponds to $county_i$

Therefore, we should consider a multivariate time series, which is a natural extension of the univariate autoregressive model, and often provides superior forecasts to those from univariate time series models[3].

To determine the basic expression of the model, we need to do some analysis:

- Firstly, we will decide the number of previous values that the current value is related to. Because we only have data of 8 years, the order cannot be too high so that overfitting may occur. Let order $n = 1\ or\ 2$. When n=1, we test the reliability by applying time series 1(heroin use) as a univariate time series, then apply AR(1) model, the result shows the fitting of the train set is too poor. Thus, we choose $n = 2$, which is also validated by AIC criteria.
- To describe the spread and characteristics of the reported opioid drug cases in and between the five states and their counties over time, we should take influence of surrounding counties into consideration. We assume that the distance(D_{ij}) is inversely proportional to the degree of influence, and one county's number of drug reports is positively correlated to the number of opioid drug cases of other counties.
  For instance, the number of heroin reports in year t in county i(H_t^i) can be expressed as

  $H_t^i = \{a_{11}^1 H_{t-1}^i + a_{12}^1 F_{t-1}^i + a_{13}^1 O_{t-1}^i + a_{11}^2 H_{t-2}^i + a_{12}^2 F_{t-2}^i + a_{13}^2 O_{t-2}^i\}$ (county i's history)

  $+\{\sum_{j\neq i} \frac{(b_{11}H_{t-1}^j + b_{12}F_{t-1}^j + b_{13}O_{t-1}^j)}{D_{ij}^2}\}$(surrounding effects)$+\epsilon_{it}$(random error)

  where H_{t-1}^i, F_{t-1}^i and O_{t-1}^i are the number of three kinds of drugs respectively in county i in year $t-1$, H_{t-1}^j, F_{t-1}^j and O_{t-1}^j are the number of three kinds of opioid drugs respectively in county j in year $t-1$, and every ϵ_{it} is a random error and independent and identically distributed(i.i.d.) which follows $\mathcal{N}(\mu = 0, \sigma^2)$.
- We use map data(State Files with Federal Codes) from U.S. government[5] with longitude and latitude, then we approximate the distance scale by

$$D_{ij} = \sqrt{(longitude_i - longitude_j)^2 + (latitude_i - latitude_j)^2}$$

To sum up, we can write this model as a vector auto-regressive(VAR) equation as follow:

$$SelfHistory_t^i = \sum_{k=1}^{2} \begin{pmatrix} a_{11}^k & a_{12}^k & a_{13}^k \\ a_{21}^k & a_{22}^k & a_{23}^k \\ a_{31}^k & a_{32}^k & a_{33}^k \end{pmatrix} \begin{pmatrix} H_{t-k}^i \\ F_{t-k}^i \\ O_{t-k}^i \end{pmatrix}$$

$$Spread_t^i = \sum_{j \neq i} \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \begin{pmatrix} \frac{H_{t-1}^j}{D_{ij}{}^2} \\ \frac{F_{t-1}^j}{D_{ij}{}^2} \\ \frac{O_{t-1}^j}{D_{ij}{}^2} \end{pmatrix}$$

$$\begin{pmatrix} H_t^i \\ F_t^i \\ O_t^i \end{pmatrix} = SelfHistory_t^i + Spread_t^i + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \end{pmatrix}$$

The next step, we will apply data fitting to determine the coefficient matrix A_1, A_2 and B, the result is

$$\begin{pmatrix} a_{11}^1 & a_{12}^1 & a_{13}^1 \\ a_{21}^1 & a_{22}^1 & a_{23}^1 \\ a_{31}^1 & a_{32}^1 & a_{33}^1 \end{pmatrix} = \begin{pmatrix} 1.067 & -0.149 & -0.243 \\ 0.002 & 1.795 & 0.005 \\ 0.017 & 0.055 & 0.716 \end{pmatrix} \begin{pmatrix} a_{11}^2 & a_{12}^2 & a_{13}^2 \\ a_{21}^2 & a_{22}^2 & a_{23}^2 \\ a_{31}^2 & a_{32}^2 & a_{33}^2 \end{pmatrix} = \begin{pmatrix} -0.076 & -0.522 & 0.372 \\ 0.134 & -0.143 & -0.092 \\ 0.006 & 0.191 & 0.106 \end{pmatrix}$$

$$\begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} = \begin{pmatrix} 2.48E{-}5 & -3.82E{-}5 & 1.79E{-}5 \\ -9.66E{-}6 & -6.33E{-}6 & 1.00E{-}6 \\ -3.05E{-}5 & -2.06E{-}5 & -2.56E{-}5 \end{pmatrix}.$$

4.3 Identify Origin Locations of Drug Use

1. Estimate Specific Drug Use in Previous Years

Based on the model bulit in last subsection, we have a recursive formula of a vector time series. Therefore, we can use this recursive formula(with simple transformation) to derive any specific opioid drug use in any previous year before 2010 as follow:

$$\begin{pmatrix} H_{t-2}^i \\ F_{t-2}^i \\ O_{t-2}^i \end{pmatrix} = \begin{pmatrix} a_{11}^2 & a_{12}^2 & a_{13}^2 \\ a_{21}^2 & a_{22}^2 & a_{23}^2 \\ a_{31}^2 & a_{32}^2 & a_{33}^2 \end{pmatrix}^{-1} \left(\begin{pmatrix} H_t^i \\ F_t^i \\ O_t^i \end{pmatrix} - Spread_t^i - \begin{pmatrix} a_{11}^1 & a_{12}^1 & a_{13}^1 \\ a_{21}^1 & a_{22}^1 & a_{23}^1 \\ a_{31}^1 & a_{32}^1 & a_{33}^1 \end{pmatrix} \begin{pmatrix} H_{t-1}^i \\ F_{t-1}^i \\ O_{t-1}^i \end{pmatrix} \right)$$

From the recursive formula above, obviously the estimations of the number of heroin(H_{t-p}^i) and fentanyl compounds(F_{t-p}^i) use reports in any previous year can be recursively obtained by calculating the first and second element of the vector on the left side of the equation above.

However, for a specific opioid drug use information except heroin and fentanyl compounds in previous years, we cannot directly obtain the estimation by the recursive formula above. We will discuss this situation below.

For instance, now we don't know the Methadone use data of county i in any year before 2010, say, 2009. Because of our assumption, the separate time series of Methadone use reports(denoted as $\{Me_t^i\}$) follows the same rule as $\{O_t^i\}$, so we can substitute each value in Methadone use reports into each $\{O_t^i\}$ without changing the coefficient matrix. Apply the recursive formula above, we get

$$\begin{pmatrix} H_{2009}^i \\ F_{2009}^i \\ Me_{2009}^i \end{pmatrix} = \begin{pmatrix} a_{11}^2 & a_{12}^2 & a_{13}^2 \\ a_{21}^2 & a_{22}^2 & a_{23}^2 \\ a_{31}^2 & a_{32}^2 & a_{33}^2 \end{pmatrix}^{-1} \left(\begin{pmatrix} H_{2011}^i \\ F_{2011}^i \\ Me_{2011}^i \end{pmatrix} - Spread_{2011}^i - \begin{pmatrix} a_{11}^1 & a_{12}^1 & a_{13}^1 \\ a_{21}^1 & a_{22}^1 & a_{23}^1 \\ a_{31}^1 & a_{32}^1 & a_{33}^1 \end{pmatrix} \begin{pmatrix} H_{2010}^i \\ F_{2010}^i \\ Me_{2010}^i \end{pmatrix} \right)$$

We can estimate the number of Methadone use reports of $county_i$ in 2009(Me^i_{2009}) by the expression on the left side of the equation. It is worth noting that this expression also gives the estimations of the number of heroin and fentanyl compounds use reports, but these two estimations are meaningless because $\{O^i_t\}$ is substituted into $\{Me^i_t\}$!

2. Identify Origin Locations of Drug Use

There are 462 counties recorded in the excel sheet and each county has a corresponding number($FIPS\_Combined$). Our goal is to identify any possible locations where specific opioid use might have started in each of the five states. It's enough if we can estimate the exact year when a specific opioid have started.

Algorithm

As mentioned before, we can estimate specific opioid drug use in previous years. Let $T_{ij} = \{the\ earliest\ time\ point\ when\ drug\ use\ in\ county_i\ reaches\ 0 | county_i \in state_j\}$

If $county_m$ has the minimum T_{mj} among all counties in $state_j$, then $county_m$ is deemed the origin location of the specific drug use in $state_j$.

If more than one $county_i$ in $state_j$ doesn't exist their T_{ij}s, which means we cannot determine a unique origin location. These counties are all deemed origin locations.

Take Beaver County, Pennsylvania(FIPS:42007) for example, we apply our algorithm to estimate the number of opioid drug use reports from 2001 to 2009.

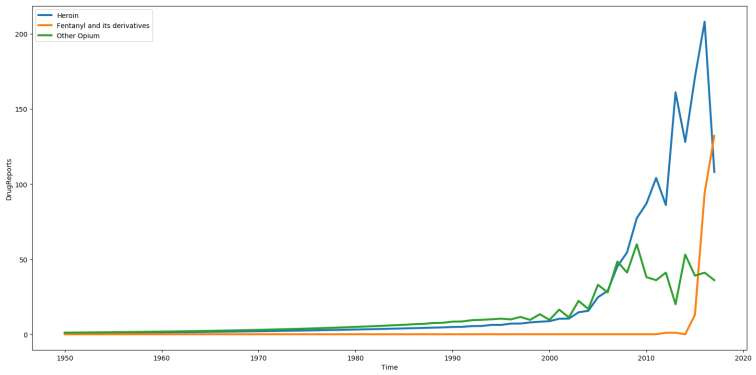

Figure 9: Estimation of Previous Data Before 2010(County FIPS:42007)

This figure shows that the number of fentanyl and its derivatives in Beaver County reaches 0 in 2014, which indicates that fentanyl use in Beaver County started in 2014, so $T_{Beaver,PV} = 2014$ for fentanyl use. Besides, we can see that the estimated number of heroin use reaches 0 in 1950,which indicates that heroin use in Beaver County started in 1950, so $T_{Beaver,PV} = 1950$ for heroin use. After backcasting the fentanyl use in every county, we find that the possible origin locations for fentanyl are Ohio and Pennsylvania.

Similarly, for a specific opioid drug use, we can calculate corresponding T_{ij} of every county then find the county which has the minimum T_{ij} among the state it belongs to. Then we can determine the origin location for a specific drug use in each state.

4.4 Predict Opioid Drug Use

We gave the recursive formula at the bottom of Section 4.2:

$$\begin{pmatrix} H_t^i \\ F_t^i \\ O_t^i \end{pmatrix} = SelfHistory_t^i + Spread_t^i + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \end{pmatrix}$$

where $SelfHistory_t^i$ and $Spread_t^i$ are explained in detail in Section 4.2.

For a random county $county_{51003}$(Albemarle, VA), we can recursively predict future drug use by the formula above. We draw a line chart to depict the predicted future opioid drug use in $county_{51003}$.

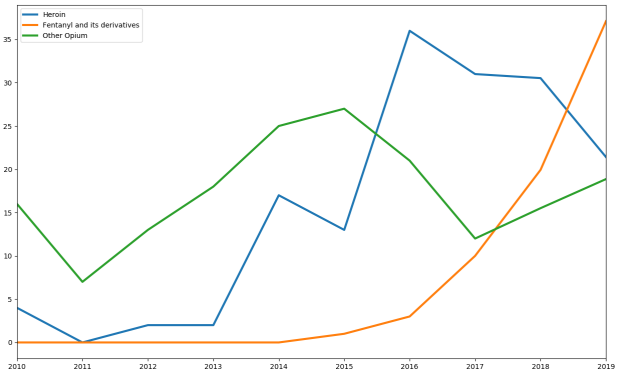

Figure 10: Prediction of Future Drug Use(County FIPS:51003)

Then we circulate similar steps 462(the number of all counties reported) times to predict future opioid drug use of every county in the five states, and visualize them on the map as follow.

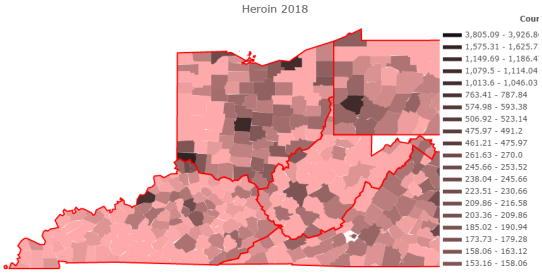

Figure 11: Heroin Use Prediction 2018

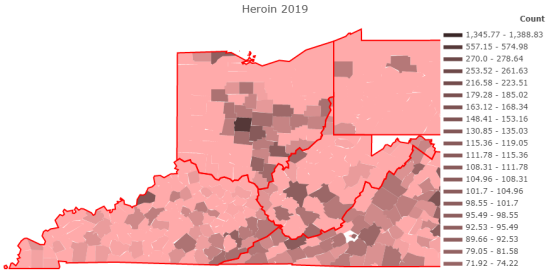

Figure 12: Heroin Use Prediction 2019

According to the figures above which predict heroin use in next two years, the degree of heroin use will decrease at a relatively rapid speed in all 5 states, and the total number of heroin report will be less than 10000 in 2019. If we set the heroin identification threshold as 1000 reports per county per year, there will be only one county in Ohio above this threshold. Thus, U.S. government could pay less attention on restriction against heroin abuse.

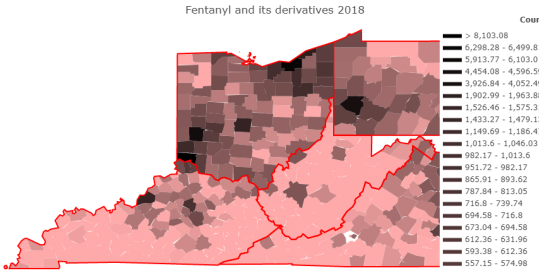

Figure 13: Fentanyl Compounds Use Prediction 2018

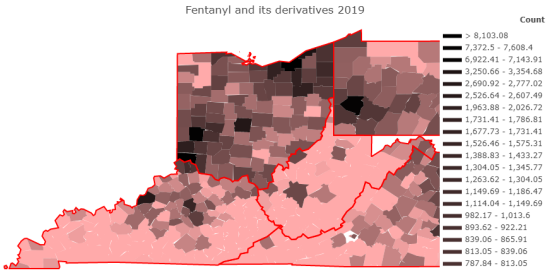

Figure 14: Fentanyl Compounds Use Prediction 2019

As the prediction result shows, fentanyl and its derivatives abuse level is fairly high in 2018, especially in the two northern states Pennsylvania and Ohio. What's more, the fentanyl abuse is getting worse and worse. The total number of fentanyl and its derivatives report will be more than 66000 in 2018 and 110000 in 2019. Whatever the reasonable fentanyl identification threshold is, most counties in Ohio and Pennsylvania and about half southern counties exceed the threshold undoubtedly. Therefore, fentanyl and its derivatives abuse should definitely be the most concern that U.S. government should have. Governments of Ohio and Pennsylvania should put forward strict restrictions on fentanyl compounds use, while the three southern states Kentucky, Virginia and West Virginia governments should be vigilant against the adverse effects of other states, and combat the rise in the number of fentanyl use reports.

Besides, as we assume, we don't take any competitive new drugs that may emerge in the future because it's out of our control. But in actual situation, it's highly possible that there will be a new competitive drug which plays the role that fentanyl is now playing. At that time, the number of fentanyl reports may start decreasing(like heroin) and the number of new drug reports may increase at a drastic speed. It's important to note that this is our speculation out of our model.

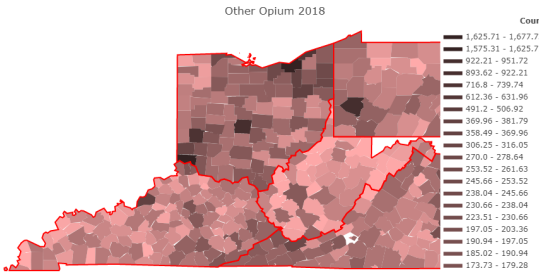

Figure 15: Other Opioid Drugs Use Prediction 2018

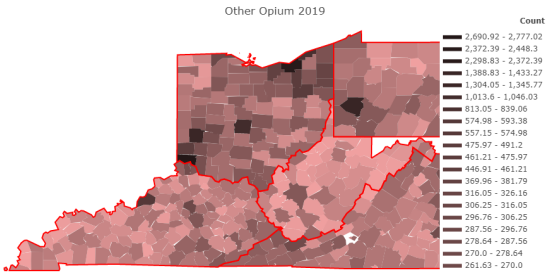

Figure 16: Other Opioid Drugs Use Prediction 2019

As for other opioid drugs, the use density is slowly increasing, so governments should enhance routine measures for the use of these opioid drugs. But according to the model prediction, the degree of abuse of these drugs will not rise to a severe level in a short time. We are not sure about long-term predictions of the use of these opioid drugs because the prediction will be highly unstable.

5 Part 2: Socio-economic Factors

In this part, we will additionally apply U.S. Census Socio-economic Data to analyze the association between this data and opioid drug use, and then modify our model built in Part 1 in order to produce a better effect.

5.1 Data Preprocessing

5.1.1 Data Characteristics

After screening all worksheets, we find that the data has following features:

- The data of each year during 2010-2016 has a large number of attributes.
- There are some differences in attribute name set every year. What's more, sometimes two attribute items with the same meaning have different names. For instance,

$$Attribute\ name\ 1:\ With\ own\ children\ of\ the\ householder\ under\ 18\ years$$

$$Attribute\ name\ 2:\ With\ own\ children\ under\ 18\ years$$

  Roughly speaking, the annual data since 2014 are more subdivided and detailed.
- The data has invalid values and columns.
- The value of 'GEO.display-label' in the metadata worksheet each year is sometimes inconsistent.

5.1.2 Handling Invalid Values

For columns in which almost all(>80%) elements are not valid (e.g. (X), *****, etc.), we regard these columns as invalid columns. Then we screen all worksheets and delete these columns.

For columns in which few elements include '*' (e.g. **, ***, etc.), we set them as '0'.

For columns in which few elements are '(X)', we firstly identify if the attribute name corresponding to the element also exists in the data of other years, if so, apply cubic spline interpolation to substitute the '(X)'; if not, we set the '(X)' as 0.

5.1.3 Handling Attributes with Repetitive Meanings

For any pair of attribute names which has the same meaning, we combine them as one attribute. However, it's beyond our ability to combine all attributes with the same meaning by coding, and it's too much work to combine them manually. To deal with this problem, we have to skip this step in the first place and directly run our algorithm to select important attributes, then manually combine important attributes with repetitive meanings and run our algorithm again.

5.2 Correlation with U.S. Census Socio-economic Data

We will apply decision tree and correlation coefficient in order to explore the correlation.

Correlation Coefficient

In the first place, we calculate the correlation coefficient between 'TotalDrugReportsCounty' and each attribute in each year during 2010-2016 by

$$r_j = \frac{\sum_{i=1}^{n}(x_{ji} - \bar{x_j})(Total_i - \bar{Total})}{\sqrt{\sum_{i=1}^{n}(x_{ji} - \bar{x_j})^2 \sum_{i=1}^{n}(Total_i - \bar{Total})^2}}$$

where r_j is the correlation coefficient between $Attribute_j$ and total opioid drug reports, and x_{ji} is the ith element of $Attribute_j$. The correlation coefficient tells us the degree of linear correlation between the number of drug reports and each attribute.

From our result, almost in every year, there are over 30 attributes(independent variables) that have correlation coefficients larger than 0.8, which means these attributes are strongly linearly related to total opioid drug reports. Among them, some attributes' correlation coefficients are relatively high in every year, so we are convinced that these attributes are indeed highly linearly correlated to opioid drug use.

To intuitively show our results and make it simple to analyze, we calculate the average correlation coefficient of each attribute over 7 years, and rank them in descending order. Then we classify the attributes with higher correlation coefficient as follow:

Category	Subdivision Category	Sub Subdivision Category	Correlation
DISABILITY STATUS OF THE CIVILIAN NONINSTITUTIONALIZED POPULATION	Under 18 years	With a disability	0.857
	18 to 64 years		0.881
	65 years and over		0.853
	Total Civilian Noninstitutionalized Population		0.875
EDUCATIONAL ATTAINMENT	Population 25 years and over	Less than 9th grade	0.757
		9th to 12th grade, no diploma	0.858
		Associate's degree	0.809
		High school graduate (includes equivalency)	0.837
		Some college, no degree	0.829
		Bachelor's degree	0.721
GRANDPARENTS	Number of grandparents responsible for own grandchildren under 18 years	Who are female	0.873
		Who are married	0.836
HOUSEHOLDS BY TYPE	Total households	Family households (families)	0.794
		Nonfamily households	0.875
	Nonfamily households	Householder living alone	0.834
	Family households (families)	Male(Female) householder, no husband(wife) present, family	0.853 (0.827)
LANGUAGE SPOKEN AT HOME	Population 5 years and over	English only	0.832
MARITAL STATUS	Males(Females) 15 years and over	Widowed	0.854 (0.851)
		Never married	0.850 (0.860)
		Separated	0.833 (0.838)
		Divorced	0.836 (0.835)
		Now married, except separated	0.736 (0.735)
PLACE OF BIRTH	Total population	Native	0.832
	Native	Born in United States	0.787
		Born in Puerto Rico, U.S. Island areas, or born abroad to American parent(s)	0.704
RELATIONSHIP	Population in households	Householder	0.834
		Nonrelatives	0.818
		Child	0.804
		Spouse	0.732
		Other relatives	0.801
	Nonrelatives	Unmarried partner	0.796
RESIDENCE 1 YEAR AGO	Population 1 year and over	Different house in the U.S.	0.820
		Same house	0.812
	Different house in the U.S.	Same county	0.787
SCHOOL ENROLLMENT	Population 3 years and over enrolled in school	College or graduate school	0.811
		Kindergarten	0.798
		Nursery school, preschool	0.797

Figure 17: Top Categories of Attributes with Higher Correlation Coefficient

Because of space limitation, we cannot show the complete correlation coefficient rank. By observing the attributes with high correlation coefficient, we find that these

attributes mainly belong to these categories:

1. MARITAL STATUS
2. EDUCATIONAL ATTAINMENT
3. HOUSEHOLDS BY TYPE
4. DISABILITY STATUS OF THE CIVILIAN NONINSTITUTIONALIZED POPULATION
5. SCHOOL ENROLLMENT
6. PLACE OF BIRTH
7. GRANDPARENTS
8. RELATIONSHIP
9. RESIDENCE 1 YEAR AGO

It can be deduced from the result that socio-economic conditions linearly influence the opioid drug use from many aspects. People who have problems in these attributes above are more likely to use drugs.

Decision Tree

In the above analysis, we only take linear correlation into consideration. However, there is nonlinear correlation between variables in actual situation. The accurate way to assess the degree of correlation is to consider both linear and non-linear relationships. As for nonlinear correlation analysis, it's almost impossible for us to quantify as we did in linear correlation analysis. To identify the existence of nonlinear correlation as well as possible, we apply decision tree to process our data.

We use the Classification And Regression Tree(CART), which is a kind of decision tree suited for data regression. In this case, opioid drug report number of each county is set as the dependent variable to be regressed. On each layer of the regression tree, we use mean square error of the opioid drug report number as criteria to decide the optimal partition of the dataset, that is, the mean square error of the sub-tree after partition should be the smallest. Each partition process is as follow:

1. Traverse all feature columns and all sample points of each feature column in turn. According to each sample points, try dividing the data sets into two parts, and calculate the sum of mean square errors of the two subtrees.

2. Find the minimum mean square error, as well as its corresponding feature columns and sample points. Thus, partition attribute and partition value of current layer are decided.

After several rounds of operation, the algorithm stops after an accurate and uniform partition of the data set is formed (that is, the opioid drug report value of the data can be predicted more accurately by using the tree). At this time, all the attributes and partition points which have great influence on the data acquisition can be determined.

We apply this algorithm to the economic and social data of each year, and establish the corresponding regression tree. In order to divide the data evenly, keep the tree simple to find the most important attributes, and prevent the occurrence of over-fitting, we set the maximum depth of the tree to 10, and the minimum sample number of each leaf to 10. Thus, the number of attributes decided by each tree could be less than 20.

We have generated some visualized graphs of trees to make them intuitional. Here's part of the bottom of a regression tree generated from socio-economic data of 2010 (an entire graph is too big for this paper).

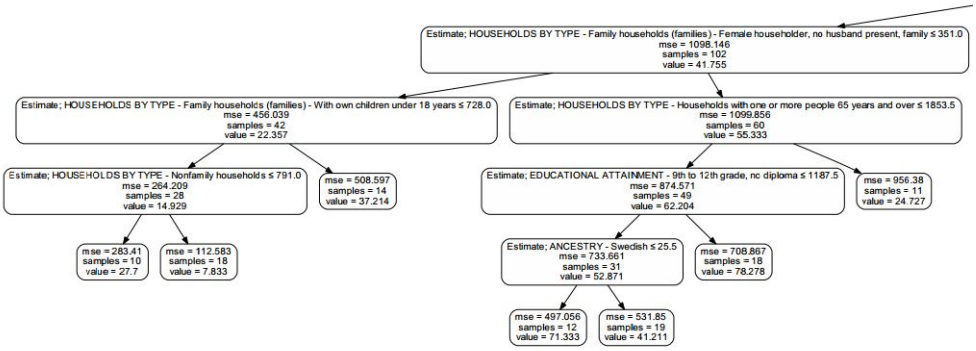

Figure 18: Part of the Bottom of the Regression Tree

In order to check the effectiveness of the regression tree, we could measure the difference between the regression result and the real value. Here's the perfomance of the tree shown above when predicting opioid drug use in Kentucky. The orange points indicate the real value, and blue points indicates the prediction values.

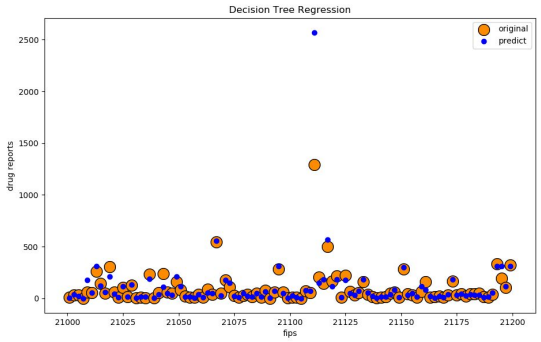

Figure 19: Performance of the Regression Tree

We could see from the picture that the tree performs well with less than one-tenth of total attributes, despite some errors with extreme values. With trees derived from data of different years, a lots of those decisive attributes are the same(or with similar meanings). Retaining recuring attributes, we can confidently say they matter in some ways.

For attributes that are deemed important by our decision tree, we understand them from two perspectives:

- If the attribute has sufficiently large correlation coefficient(>0.6) with total opioid drug reports, then we identify its correlation as linear.
- If the attribute doesn't have sufficiently large correlation coefficient with total opioid drug reports, then we identify its correlation as nonlinear.

For attributes that are deemed not important by our decision tree, we identify them as not related to opioid drug use.

The result shows that the decision tree nearly makes the same judgement we made with correlation coefficient, which means most attributes deemed important by decision tree are the same as Figure 17. But there are still slight differences. We only show the

new attributes that are deemed important by decision tree as follow.

Category	Subdivision Category	Sub Subdivision Category
FERTILITY	Number of women 15 to 50 years old who had a birth in the past 12 months	Per 1,000 women 15 to 50 years old
		Per 1,000 women 20 to 34 years old
VETERAN STATUS	Civilian population 18 years and over	Civilian veterans
GRANDPARENTS	Number of grandparents responsible for own grandchildren under 18 years	Years responsible for grandchildren- 1 or 2 years
		Years responsible for grandchildren- 5 or more years
HOUSEHOLDS BY TYPE	Average family size	
	Households with one or more people 65 years and over	
YEAR OF ENTRY	native	
	Entered 2000 or later	

Figure 20: New Attributes Deemed Important by Decision Tree

There are 3 new attribute categories which are deemed important by decision tree. For example, 'VETERAN STATUS' is deemed important by decision tree while its correlation coefficient with total opioid drug use reports is relatively low. Thus we identify the correlation between veteran status and opioid drug use as nonlinear. Similarly, 'FERTILITY' and 'YEAR OF ENTRY' are also identified as nonlinear correlations with opioid drug use.

Other than these 3 new categories, there are also several new subdivided and categories. For instance, 'average family size' is a new subdivided attribute category under 'HOUSEHOLD BY TYPE', which indicates that, some attributes under category 'HOUSEHOLD BY TYPE' have linear correlation with opioid drug use, some attributes under category 'HOUSEHOLD BY TYPE' have nonlinear correlation with opioid drug use and other attributes under category 'HOUSEHOLD BY TYPE' don't have obvious correlation with opioid drug use. Similar analysis method can be applied to any attribute or attribute category, and we can obtain a detailed and rigorous report. But we don't repeat similar analysis out of space limitation.

(Note: In fact, our model shows that 'ANCESTRY' and 'WORLD REGION OF BIRTH OF FOREIGN BORN' are also important attributes related to opioid drug use. We choose not to discuss issues involved with ethnic groups.)

5.3 Improved Model

To include important socio-economic factors, we need to modify our model built in Chapter 4 to make it more accurate. In Part 1, we modified the vector autoregressive model by adding the opioid drug use reports number and distance factor. Now that we need to take socio-economic factors into consideration, we should make use of socio-economic data to somehow correct the forecast or backcast value of the model built in Part 1. Being that there are too many various socio-economic attributes related to opioid drug use as we analyzed, it's not appropriate to simply add independent variables to the vector autoregressive model. Given that the decision tree can be also applied to predict, we decide to synthesize the prediction given by the decision tree and the prediction given by our autoregressive model. We generate a decision tree only based on important attributes mentioned above. Specifically, we denote the prediction of drug use reports given by the model in Part 1 as y', and denote the prediction of drug use reports given by the decision tree as y''. Then our final prediction y is in this form:

$$y = \alpha y' + \beta y'' + \epsilon$$

where $\epsilon = \mathcal{N}(\mu = 0, \sigma^2)$ is the random error.

In this way we transform the problem of balancing two predictions into a linear model without intercept, where y' and y'' are regarded as independent variables and y is regarded as the dependent variable. The criterion of least squares is to minimize

$$RSS = \sum_{t,j} \epsilon_{tj}^2 = \sum_{t,j} (y_{tj} - \alpha y'_{tj} + \beta y''_{tj})^2$$

where y_{tj} is the true number of drug reports of $county_j$ in $year_t$, y'_{tj} and y''_{tj} are our predictions of $county_j$ in $year_t$ given by two models respectively. Then, the least square estimation of α and β are $\hat{\alpha}$ and $\hat{\beta}$ which make

$$Q_e = \sum_{t,j} (y_{tj} - \hat{\alpha} y'_{tj} + \hat{\beta} y''_{tj})^2 = min_{\alpha,\beta}\{RSS\}$$

The fitting result of this linear model as $\hat{\alpha} = 0.837$ and $\hat{\beta} = 0.241$. So our final prediction is $\hat{y} = 0.837y' + 0.241y''$, which indicates that the modified vector autoregressive model built in Part 1 performs better than the decision tree which only takes socio-economic factors into consideration. Thus, the first model still has a decisive impact on the final prediction and the decision tree can slightly modify the prediction given by our first model. The coefficient of determination R^2 increases 0.03 compared with the first model.

Additional Considerations: In our previous analysis, there is linear correlations between socioeconomic factors and opioid drug use. In dealing with parameter estimation of linear models, because of multicollinearity, the ordinary least squares method is likely to have large mean square error(MSE) in prediction, which make the model highly unstable. To deal with multicollinearity, we could apply Ridge Regression or Lasso. Due to space limitation, we are not going to make further discussion.

6 Part 3: Strategy for Countering the Opioid Crisis

6.1 A Possible Strategy

The possible strategy we propose is a **set** of sub-strategies, for the sake of simplicity of presentation, we refer to 'sub-strategies' as 'strategy' in the following paper. We will explain our strategy as follow.

From a geographical point of view

In Part 1 we identified origin locations and made predictions of opioid drug use in Ohio, Kentucky, West Virginia, Virginia, and Pennsylvania. Any possible strategies proposed should be in view of characteristics of opioid drug use in these five states.

For Ohio and Pennsylvania, opioid drug abuse has already been a severe issue. Thus governments in Ohio and Pennsylvania should definitely strengthen the prohibition against illegal drugs like heroin(although heroin use is decreasing). For illegal drugs, prevention is key. This may sound like a cliché but it's nevertheless true. Prevention is the best way to keep people from exposing to illegal drug sources.

As for Virginia, West Virginia and Kentucky, governments of these states should be vigilant about the spread of opioid drugs from the two northern states. Measures such as enforcing road transportation management and control by State Highway Patrol are possibly useful for containing the spread of opioid drugs. What's more, the U.S. Drug Enforcement Administration(DEA) in these states may turn attention to fighting against

fentanyl and its derivatives compared with other drugs because fentanyl compounds are fairly potent and fentanyl use are drasticly increasing.

From a socio-economic point of view

Based on those important socio-economic attributes given in Part 2(Figure 17 and 20), we are going to propose possible strategies from a socio-economic point of view.

Educational enrollment is one of important factors related to opioid drug use. The U.S. government should find ways to reduce the number of school drop-outs and strengthen drug addiction prevention and control education in campus.

Household type is also highly correlated to opioid drug use, regular community anti-drug propaganda and experience exchanging meetings should be useful for nonfamily households and householders living alone who don't have families around to supervise.

We find that the attribute named 'Grandparents living with/responsible for children under 18 years' impacts opioid drug use a lot. We believe that it may be due to the lack of adequate control and care for some children, who are living alone with their grandparents. Therefore, the government social welfare agencies should focus on those children, and give them appropriate psychological counseling and economic support if necessary. At the same time, they should attach importance to their school education and convey positive information.

We also find that veterans and women who had a birth in the past year are more likely addicted to opioid drugs. Old wounds in long-term military training may cause a large number of veterans to use painkillers. Pregnant women also usually suffer a lot of pain before and after childbirth. People who have legitimately needed pain relief often end up addicted to opioid drugs. Thus, any individual or institution with qualification to prescribe should be careful about prescribing addictive drugs.

Similarly, for other factors identified important in Part 2, we can propose possible strategies as well based on its socio-economic characteristics. Specially, as for factors like marital status and year of entry, it is almost impossible for us to control these factors at the social and policy levels. Thus we don't propose strategies based on these factors.

6.2 Strategy Effectiveness Test

To evaluate the effectiveness of a specific strategy, we make some proper changes to corresponding attributes assuming that the strategy has already implemented, then observe the predictions produced by the model and see if the opioid drug use situation gets better.

Overall Effectiveness for 5 States

Firstly, we analyze the effectiveness of strategies from a overall perspective. As we mentioned above, one possible strategy is to strengthen the control of addictive drugs. Assume that this strategy has come into force, so the number of drug reports should decrease, and we want to see how it will influence future. Without loss of generality, we reduced the number of people by 30% in 2016 and 2017, and predict the drug use in 2018 and 2019, then compare the new predictions with our original predictions. The result is as follow:

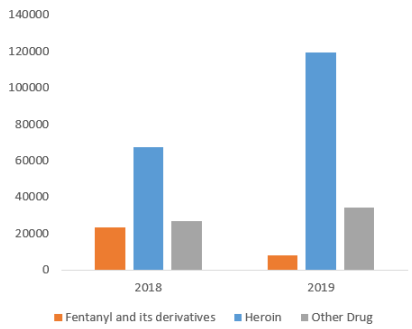

Figure 21: Original Predictions

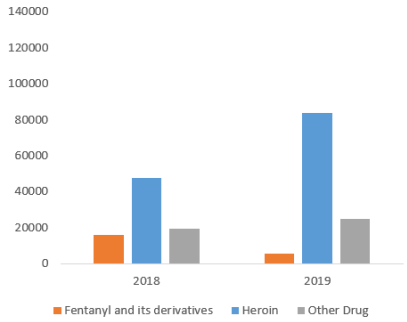

Figure 22: New Predictions after the Strategy Come into Force

As the result showed above, after strengthen the control of addictive drugs(only in 2016 and 2017), the predicted opioid drug use will be much lower than original predictions. Nevertheless, we find that the increase rate won't vary a lot compared with original predictions. In other words, strengthen the control of addictive drugs can only influence the quantity of opioid drug use but not the trend unless keep strengthening the control forever. Also, we similarly test the effectiveness of containing spread of opioid drugs, the result is 'not much effectiveness'.

As for socio-economic factors, we take following 3 strategies as example. The first column in the following picture indicates the original number of opioid drug reports in five states. The second, third and fourth column indicates the prediction value of opioid drug use reports by decreasing the number of adults without high school education(by 30%), decreasing the number of people divorced(by 30%) and decreasing the number of grandparents living with own grandchildren under 18 years(by 30%). The fifth column indicates the prediction value of opioid drug use reports with these 3 strategies coming into force cooperately.

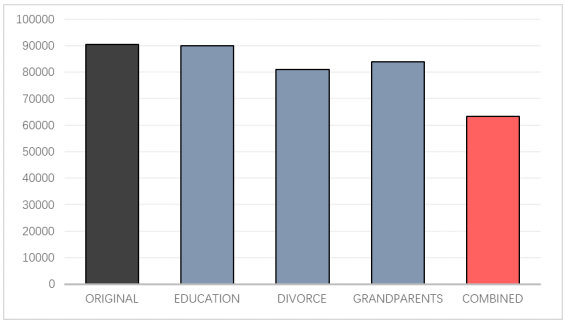

Figure 23: New Attributes Deemed Important by Decision Tree

The result shows that the strategy related to divorce effects the most among the three strategies while the strategy related to education effects the least. Besides, the combined effectiveness exceeds the sum of effectivenesses of the three strategies coming into force separately, which means the government should use multiple strategies cooperately.

Local Effectiveness for a County

Next, being that a strategy may be effective for five states but not for a certain county, we need to analyze of the strategy's effectiveness on each county from a more micro perspective.

For example, assume our strategy to reduce the number of school drop-outs comes

into force, which means the number of people with higher educational attainment will increase. Then, we raise several positive attribute(like 'Bachelor's degree') under 'EDUCATIONAL ATTAINMENT' category by 25% and see what's the prediction of our model. To avoid loss of generality, then we choose other two attributes and repeat similar operations separately. We put three results in the left picture where every dot represents a county. The result shows opioid drug use in most counties is nearly not affected by only changing 'EDUCATIONAL ATTAINMENT' attribute, and only several counties showed positive variations in opioid drug use. Analysis of effectiveness of other two attributes is similar.

The next question is what will happen if multiple strategies come into force at the same time. We raised six positive attributes at the same time(including increasing ratio of high school and above certificate, decrease elders living alone or with children, decrease divorce, etc.), the result is the right picture. The result shows opioid drug use in majority of counties decrease, which means it's effective to implement multiple strategies at the same time. However, the strange thing is the opioid drug use in some counties rises on the contrary. For these strange counterexamples, maybe they reflect the truth or maybe they reflect errors of our model.

(Note: To better observe the result, we only show part of the coordinate system.)

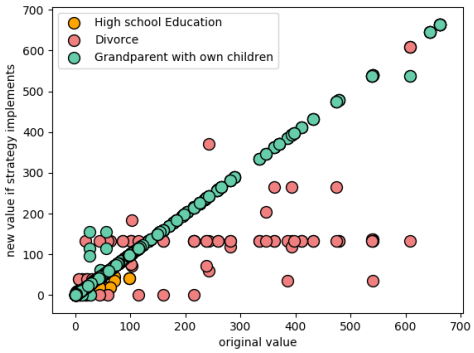

Figure 24: Implement Strategies Separately

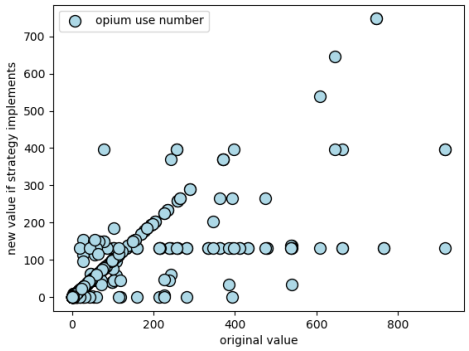

Figure 25: Implement Strategies Cooperately

6.3 Significant Parameter Bounds

In this part, we consider our strategy to be success when it reaches the result that the opioid drug use in all 5 states(including heroin, fentanyl and others) has decreased by 30%. For drug use control, we must decrease the opioid drug use number of the past year by at least 45% to reach the goal. If we modify thedrug use number of past two years, we only have to decrease it by about 30%. For socio-economic attributes, due to their complex correlation, it's hard for us to generate a well-described how-to-modifylist for the strategy to work well. Yet, we still find a general description to evaluate the variation bounds for attributes when they work together to reach success, which is shown below.

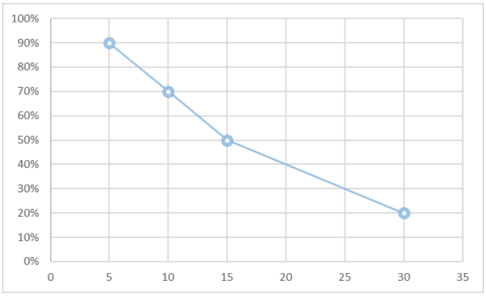

Figure 26: Average Variation Range of Attributes

The x-coordinate indicates the number of attributes working together, and the y-coordinate indicates the average variation range of each attribute measured by percentage.Variation by 20% of 30 socio-economic attributes may still sounds hard to realize. However, each county has different characteristics, which certain few attributes could impact a lot, whereas large variation of other attributes could be useless. We believe that if different sub-strategies have been applied differently in each county, the feasibility and effectiveness of our model would become much better.

7 Strengths and Weaknesses

7.1 Strengths

- Apply data visualization to explain the raw data as well as present our results intuitively and succinctly.
- Our modified VAR(n) model has advantages of vector autoregressive model(VAR), and also includes the interaction of changes in the number of drugs and the geographical distance factor, making it more credible.
- By calculating correlation coefficients and applying decision tree, we make a thorough and quantatitive analysis on correlation with socio-economic factors.

7.2 Weaknesses

- Generally speaking, the decision tree is not precise for prediction, but we directly use the prediction given by decision tree to correct the prediction given in Part 1.
- Our model does not perform well in long-term prediction.

References

[1] https://deserthopetreatment.com/drug-abuse/addictive-opiates/

[2] https://www.washingtonpost.com/graphics/2017/health/opioids-scale/

[3] Zivot, Eric, and Jiahui Wang. "Vector autoregressive models for multivariate time series." Modeling Financial Time Series with S-PLUS® (2006): 385-429.

[4] https://www.cdc.gov/nchs/fastats/drug-use-illegal.htm

[5] https://geonames.usgs.gov/domestic/download_data.htm

Appendices

Because there's too much code, we'll just pick a few to present as follow.

Appendix A Tran.py

```
      import xlrd
import numpy as np
import json

workbook = xlrd.open_workbook("E:/Users/Ryan/Desktop/MCM/MCM_NFLIS_Data.xlsx")
sh = workbook.sheets()[1]

ANS={}
f=open("E:/Users/Ryan/Desktop/MCM/Drug.json","w")

for ind in range(1, sh.nrows):
    Name = sh.cell_value(ind, 6)
    Year = int(sh.cell_value(ind, 0))
    ID = int(sh.cell_value(ind, 5))
    DrugReports = int(sh.cell_value(ind, 7))
    if(ID not in ANS):
        ANS[ID]={}
    if Year not in ANS[ID]:
        ANS[ID][Year]={"F":0,"H":0,"O":0,"ALL":int(sh.cell_value(ind, 8)),"List":[]}
    ANS[ID][Year]["List"].append([Name,DrugReports])
    if(Name =="Heroin"):
        ANS[ID][Year]["H"]+=DrugReports
    elif(Name.strip()[-7:]=="entanyl" or Name=="Carfentanil" or Name=="ANPP"):
        ANS[ID][Year]["F"]+=DrugReports
    else:
        ANS[ID][Year]["O"]+=DrugReports

json.dump(ANS,f)
```

Appendix B forecast.py

```
      import numpy as np
import json
import os
from sklearn.linear_model import LinearRegression
from random import randint
from sklearn import preprocessing
import matplotlib.pyplot as plt


f=open("./Json.json","r")
JSON=json.load(f)

workbook = open("./Drug.json","r")
ALLDrugs = json.load(workbook)

IDs={}
\thispagestyle{empty}
TrainX=[]
TrainY=[]


for X,Y in ALLDrugs.items():
    IDs[X]=[np.zeros(3) for i in range(8)]
    for Year in range(2010,2018):
        if(str(Year) in Y):
            IDs[X][int(Year)-2010] = np.array([(Y[str(Year)]['H']),(Y[str(Year)]["F"]),(Y[str(Y

DIS={}

for dx in IDs:
    for dy in IDs:
        if(dx!=dy):
            DIS[str(dx)+str(dy)]=np.power(np.array(JSON[str(dx)])-np.array(JSON[str(dy)]),2).su

for X,Y in IDs.items():
    for years in range(2,8):
        SUM=0.0
        for dx in IDs:
            if(dx!=X):
                SUM+=(IDs[dx][years-1])/DIS[str(dx)+str(X)]
        TrainX.append(np.array([Y[years-1],Y[years-2],SUM]))
        TrainY.append(Y[years])

TrainX=np.array(TrainX).reshape(-1,9)
TrainY=np.array(TrainY).reshape(-1,3)

model = LinearRegression()
model.fit(TrainX, TrainY)

AANS={}

# PC=0.7

# for X in IDs:
#     IDs[X][6]=IDs[X][6]*PC
#     IDs[X][7]=IDs[X][7]*PC

for tX in IDs:
    TrainX=[]
    TrainY=[]

    tY=IDs[tX]
    STX=tY
```

```
    TrainY=np.array(tY).reshape(-1,3)
    for years in range(8,10):
        SUM=0.0
        for dx in IDs:
            if(dx!=tX):
                SUM+=(STX[years-1])/DIS[str(dx)+str(tX)]
        TrainX=np.array([STX[years-1],STX[years-2],SUM]).reshape(-1,9)
        TX=model.predict(TrainX).flatten()
        STX.append(TX)

    STX=np.array(STX)
    tmp=STX[-2:]
    AANS[tX]=np.where(tmp<0,0,tmp).tolist()
```

Appendix C Tree_Generation.py

```
from sklearn import tree
import numpy as np;
import matplotlib.pyplot as plt
from sklearn.externals.six import StringIO
import pydot;
import json
from sklearn.externals import joblib
import pydotplus

def readData(fileName):
    dataSet = []
    fr = open(fileName)
    head=next(fr);
    head=head.strip("\n").split('\t');
    featurename=head
    line=fr.readline()
    length=len(head)+1;
    while(line):
        if line!="\n":
            curLine = line.strip().split('\t')
            theLine = list(map(float, curLine))   # map all elements to float()
            dataSet.append(theLine)
            if(len(theLine)!=length):
                print("!!!wrong");
                print("oldlength:"+str(length));
                print("newLength:" + str(len(theLine)))
        line = fr.readline()
    return featurename,dataSet

def generateTree():
    featurename,train=readData("data/ACS_16.txt")
    featurename=featurename[1:]
    train=np.array(train);
    trainx=train[:,1:-1]

    ids=train[:,0]
    #total=trainx[:,0]
    # trainx=trainx[:,1:]
    # cols = trainx.shape[1]
    # for i in range(0,cols):
    #     trainx[:, i]=(trainx[:, i])
    trainy = (train[:, -1])
    print(trainy)
    theTree = tree.DecisionTreeRegressor(min_samples_leaf=2, max_depth=20)
```

```
    theTree.fit(trainx,trainy);
    joblib.dump(theTree,"data/tree_model.m")
    idt,y,y_=test(theTree);

    plt.figure()
    plt.scatter(idt, y, s=200, edgecolor="black",
                c="darkorange", label="original")
    plt.scatter(idt, y_,c='b',label="predict");
    plt.xlabel("fips")
    plt.ylabel("drug reports")
    plt.title("Decision Tree Regression")
    plt.legend()
    plt.show()
    graphtree(featurename,theTree,trainy)

def test(theTree):
    featurename, test = readData("data/ACS_16.txt")
    test = np.array(test);
    ids=test[:,0]
    testx=test[:,1:-1]
    testx=testx[:,:]
    # total=testx[:,0]
    # testx=testx[:,1:]
    # cols = testx.shape[1]
    # for i in range(0, cols):
    #     testx[:,i] = (testx[:,i])
    testy=(test[:,-1])
    exporty(testy,ids)
    #testy = testy[0:300];
    y_=theTree.predict(testx);
    exporty(y_, ids)
    return ids,testy,y_

def graphtree(featurename,theTree,trainy):
    data_target_name = np.unique(trainy)
    dot_data = StringIO()
    tree.export;
    tree.export_graphviz(theTree, out_file=dot_data,feature_names=featurename,filled=True, rour
    #graph = pydot.graph_from_dot_data(dot_data.getvalue())
    graph = pydot.graph_from_dot_data(dot_data.getvalue())
    graph[0].write_pdf("iris_16.pdf")
    print('Visible tree plot saved as pdf.')

def exporty(y,ids):
    jsonDict={};
    for index,item in enumerate(y):
        jsonDict[ids[index]]=item
    with open("data/predict_16.json","w") as f:
        json.dump(jsonDict,f)

generateTree()
```

Appendix D load_attributes.py

```
import xlrd
import csv
import json;
import numpy as np
import pandas as pd
```

```
#import xlwt
#readbook= xlrd.open_workbook(r'data/ACS_10_5YR_DP02_with_ann.xlsx');
#sheet=readbook.sheet_by_index(1)


countyDict={}
with open("data/drug.json") as jf:
    numbers=json.load(jf);
    for key,values in numbers.items():
        if "2016" in values.keys():
            countyDict[key]=values["2016"]['F']+values["2016"]['H']+values["2016"]["O"];
edata=[]
print(countyDict)

with open("data/ACS_16_5YR_DP02_with_ann.csv")as csvf:
    csv_reader=csv.reader(csvf);
    header1 = next(csv_reader)
    header2=next(csv_reader)
    for row in csv_reader:
        #print(row)
        nrow=row[3:-1:4]
        for i,item in enumerate(nrow):
            if item=='(X)':
                nrow[i]='0'
        nrow.insert(0,row[1])
        if row[1] in countyDict.keys():
            nrow.append(str(countyDict[row[1]]))
        else:
            nrow.append(str(0))
        edata.append(nrow)
header1=header1[3::4]
header1.insert(0,"id")
header2=header2[3::4]
header2.insert(0,"id")
print(header1)

with open("data/ACS_16.txt","w") as f:
    f.write("\t".join(header2));
    f.write("\n")
    for row in edata:
        row="\t".join(row)
        f.write(row)
        f.write("\n")
```