Team Control Number

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

**1924628**

Problem Chosen

**C**

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

---

**2019 Mathematical Contest in Modeling (MCM) Summary Sheet**

# Predictions and Countermeasures for the Opioid Crisis with Socio-Economic Factors


## Summary

We developed a set of data analyses to determine the opioids source-locations in Ohio, Pennsylvania, Virginia, West Virginia, and Kentucky, to predict the opioids consumption trends in years up to 2027, to determine the strongest socio-economic factors that influenced the opioids consumptions trends, and to propose a countermeasure that would possibly ease the Opioid Crisis.

First, we processed the data from the drug identification counts by eliminating redundant data and by filling the missing data. We chose data values under "Percent" for our future models, because they are the most representative in analyzing opioids consumption. We filled the missing data by applying the curve fitting and cubic spline methods to data missing between 20%-50% and under 20%, respectively.

Second, with our processed data, we analyzed the locations of the most probable opioid source by utilizing the Bayesian Networks with distance as the determining factor. Graphs representing the locations of several sources are plotted. We then proceed to predict the trends of the top four opioids consumption amount and percentage with Grey Box model. We give a detailed and concise analysis of the opioids types trend.

Third, we took account of the impact of major socio-economic factors in opioid consumption by increasing efficiency with Principal Component Analysis and by upgrading the previous Grey Box prediction with Vector Autoregression (VAR). VAR allowed us to take account of the socio-economic factors along with the data in opioids consumption in 2010-2017.

Last, we proposed that disability is a major factor in determining the opioid consumption trend. We performed the VAR with a decreasing factor of 1% each year and observed positive result confirming our proposal.

In conclusion, we suggest funding for institutions with opioids rehabilitation as a countering measure against the opioid crisis. We also discussed about the significant parameters that bounds the success and failures of our models, as well as the strengths and weaknesses of our models.

# MEMORANDUM

**To:** Chief Administrator of DEA/NFLIS Database

**From:** Team 1924628

**Subject:** Opioids Consumption Status and Countering Measures

**Date:** January 28, 2019

---

Dear Chief Administrator of DEA/NFLIS Database, We are honored to inform you of our achievements in data analyses and modeling.

In processing the given data in several measures, we mapped the most-consumed-opioids locations of counties with the Bayesian Networks, and projected the given data unto as far as 2017 with the Grey Box model. We received the following conclusions:

- The U.S. government should watch carefully for the consumption rates of Heroin and Fentanyl. They exhibit high consumption rates, and, if uncontrolled, can lead to a dramatic increase in abused consumption among the U.S. citizens.
- Within 1 or 2 years, the consumption amount for Heroin and Fentanyl would have reached to a threshold level of 1.5 times than in 2018.
- Ohio and Pennsylvania are states where such consumption increase is most noticeable.

With the provided U.S. Census socio-economic data, we performed a Principal Component Analysis and modified the previous Grey Box model with Vector Autoregression on evaluating the relationship between the socio-economic factors and the opioids consumption. We identified disability status and divorce condition as the major factors that dramatically increase the amount of opioids consumption.

Finally, to determine how much the disability status could affect the potential of opioids abuse, we performed the Vector Autoregression model with a 1% decrease in the disabled opioids consumption every year. The results show a corresponding decrease in opioids consumption. We suggest U.S. government funding for institutions with opioids rehabilitation. Although its benefits are small in the short-run, and thus funding are difficult to win, its long-term benefits will contribute to overall increase in the welfare of the U.S.

We hope your work for the benefit of the U.S. people will have positive impacts,

Sincerely,

Team 1924628

# Contents

# 1 Introduction

## 1.1 Problem Background

Synthetic and non-synthetic opioids abuse has been a national crisis for the U.S. People use these for treatment, management of pain, or recreational purposes. On the one hand, medical departments such as the Centers for Disease Control cannot easily prevent the negative effects of drugs while saving lives. On the other hand, the Federal Bureau of Investigation cannot simply enforce laws across all counties within a state. Opioid abuses make positions requiring precise, sensitive, and trustworthy skills difficult to fill, as well as raising costs for health care and assisted living facility staffing.

Using the Drug ENforcement Administration 2010-2017 drug identification counts for opioids in counties from five states of Ohio, Kentucky, West Virginia, Virginia, and Tennessee, and the U.S. Census Bureau 2010-2016 socio-economic factors from the five states, we will determine the most probable drug source location from the five states.

## 1.2 Our Goals and Thinking

Based on our understanding of Part 1-3, we set the following goals and utilized the following models:

- Process a set of data that would be most efficient for our purposes.
- Use Bayesian Networks to determine a connection between opioid transmission based on distance.
- Use Grey Box model and predict the future trends of the top four opioid consumption up to 10 years.
- Use Principal Component Analysis to reduce socio-economic data dimensions for efficiency in calculations.
- Analyze socio-economic data impact with Vector Autoregression.
- Determine the major socio-economic factor and come up with a countermeasure.

# 2 Assumptions and Notations

## 2.1 Assumptions

Due to the lack of some data, we make the following assumptions in our model calculations. These assumptions form the basis in our subsequent analysis.

- The collected statistics are representative of the general population: they can be treated as parameter values of opioids consumption for a county population.
- Margins of error are negligible, and can be disregarded in our models.
- The data collected in 2010-2016 or 2010-2017 follow a steady pattern. This entails that no events will result in dramatic changes in the upcoming years, and opioid consumption in the upcoming years will behave like the past years.
- Opioids consumption rates are highly dictated by addiction. Because medical uses are mainly dictated by injuries, opioids consumption rates should not change by dramatic increase in the amount of injuries. This assumption allow us to make judgements on where and what type of opioids addiction increased.
- Addiction is competitive. People will choose the most addictive opioid to consume. In other words, addiction is relatively compared. When we say "an opioid is not addictive", we meant that it has likely been outcompeted by a more addictive opioid.
- The number of data is directly related to the contribution to their correlated factor. In other words, we can disregard factors with many missing data, and the accuracy of our prediction would not be affected.
- The data provided are realistic. The results based on these data will allow us to draw a solid conclusion.

## 2.2 Notations

The primary notations used in this paper are listed in **Table 1**.

Table 1: Notations

| Symbol | Definition |
|---|---|
| $R_{Missing}$ | Data Missing Rate |
| $C_{Missing}$ | The number of missing cells in a specific column |
| $C_{Total}$ | the number of cells in that specific column |
| $f_{Intpl}(x)$ | Interpolation Function |
| $X_{max}$ | maximum value of the same factor |
| $X_{min}$ | minimum value of the same factor |
| $X_{new}$ | value after normalization |

# 3 Data Processing

The attached file "MCM_NFLIS_Data.xlsx" provides us the complete counts of drug identification in all the counties from these five states in years 2010-2017. However, there is a large amount of redundant and missing data in the additional seven files with information about family status, education level, and social structure for the counties of these five states in 2010-2016. Since the missing data could seriously affect the accuracy of our model, we need to process the data by cleaning, filling, and normalizing the data based on various algorithms.

## 3.1 The Redundant Data

We are given 611 features of socio-economic factors in the counties of the five states. For each factor, we have four features: estimate, margin of error, percent, percent margin of error. Margin of error and percent margin of error are small enough to be disregarded for most factors, 0%-3%. This elimination process left us with roughly 151 data values of "percent". "Percent" is the most representative and useful data for us. They can be immediately applied to our models.

## 3.2 The Missing Data

To address the missing values in the data files, we use the missing rate $R_{Missing}$ to handle the missing data. The missing rate is defined as:

$$R_{Missing} = \frac{C_{Missing}}{C_{Total}} \tag{1}$$

where $C_{Missing}$ is the number of missing cells in a specific column and $C_{Total}$ is the number of cells in that specific column. We come up with the following strategies to address the problem:

- We delete data in columns whose missing rate is over 50%, because the columns with the big amount of missing values will highly affect the accuracy of our model.

- We apply the curve-fitting to fill the missing data, for columns whose missing rate is between 20% and 50%,
- We fill the blanks for columns whose missing rate is less than 20% with the interpolation method. Specifically, we create an interpolation function $f_{Intpl}(x)$based on our existing data and replace the missing value with the prediction of our interpolation function $f_{Intpl}(x)$.

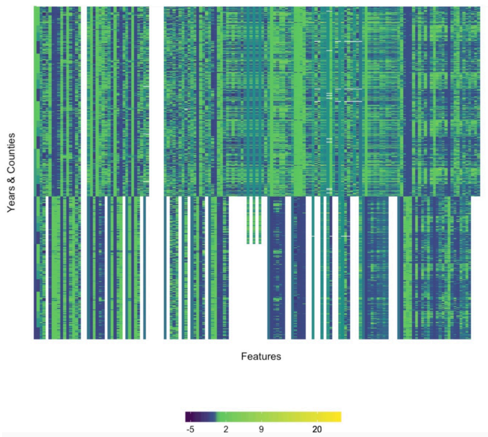



Figure 1: common set of socio-economic factors (with missing data)

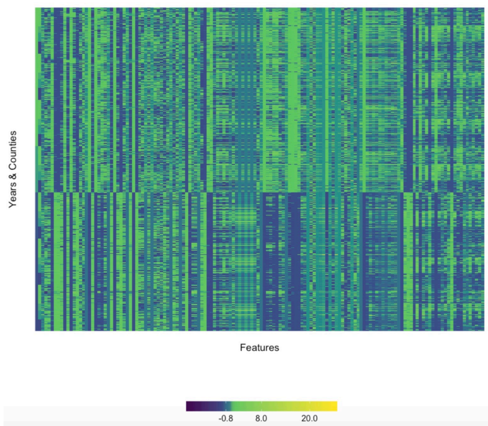



Figure 2: common set of socio-economic factors (without missing data)

### 3.3 Data Normalization

We normalized all the socio-economic factors by scaling all values in the range [0,1]. This is to increase the accuracy of our principal component analysis (PCA) in later calculations. We used the following equation as the normalizing method:

$$X_{new} = \frac{X - X_{Min}}{X_{Max} - X_{Min}} \tag{2}$$

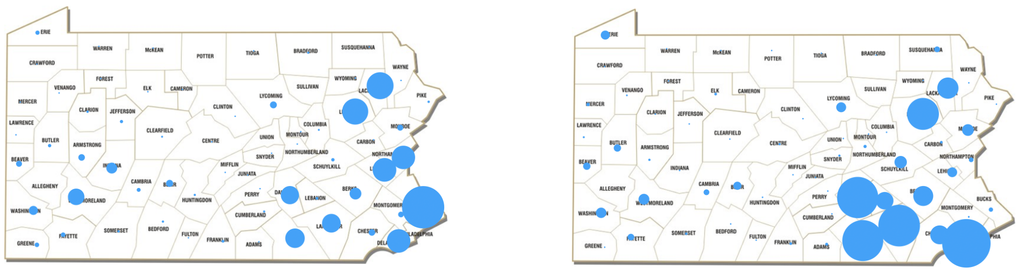
where $X_{Min}$ and $X_{max}$ represent the minimum and maximum value of the same factor, respectively.

## 4 Prediction of Drug Source

We used Bayesian network to help ourselves in determining the drug transmission between different counties. We later used Grey Box model to complete our data set and predict future opioids consumption rate.

Based on our predictions and analyses, we suggest the U.S. government to beware of Heroin and Fentanyl consumptions. Their trends suggest they are substances with high potentials of addictive abuse. Fentanyl especially is a source of danger, whose use, if uncontrolled, can have exponential consumption.

The threshold levels for these drugs occur at 1.5 times from 2010. Based on our predictions, Heroin consumption amount would reach this threshold during 2027, and Fentanyl has reached this threshold. Fentanyl would reach such 1.5 times from 2018 during 2020. Heroin consumption amount increased most visibly within Ohio, and Fentanyl consumption amount increased most visibly within Pennsylvania.



Total Heroin Distribution in PA (2010) Total Heroin Distribution in PA (2017)

Figure 3: Counties in Pennsylvania with circles representing opioids influence. The area of circles is directly related to the opioids influence.

Opioids consumption are depicted Pennsylvania, where the southwest are the most influenced regions (Figure 3).

## 4.1 Bayesian Network

A Bayesian network[1] is an annotated acyclic graph that represents a joint probability distribution (JPD) over a set of random variables. Arrows in the bayesian network indicate whether the two random variables are conditionally or non-conditionally independent. Two nodes connected together by a single arrow indicate that one of the nodes is "parent" and the other is "offspring or child". For this problem, we utilized Bayesian networks (BNs) to create directed acyclic graphs that represent the conditional probability of drug transmission between different counties. For one specific county, all the other counties may affect its drug source. For example, there are n counties in a certain state and each county has n-1 values, T(for drug source) and F(for not drug source). The joint probability function is defined as:

$$P(X_1, X_2, X_3, ..., X_n) = P(X_1|X_2, X_3, ..., X_n)P(X_2|X_3, ..., X_n)...P(X_n) \quad (3)$$

where $X_1, X_2, X_3, ..., X_n$ are the set of counties.

## 4.2 Solution and Results of Bayesian Network

We utilized Bayesian networks (BNs) to create directed acyclic graphs that represent the conditional probability of drug transmission between different counties. To increase the accuracy of BNs calculation, we removed irrelevant counties in the following way. We drew heatmaps for each of the five states. The intensity of the color is directly related to the drug consumption statistics (Figure 4 and 5). We inferred that the light-colored counties are less likely to be the drug-source counties, because they follow a steady pattern and have consumed drugs at a low rate. We also inferred that counties with high variance of drug detection are most likely to be the drug sources, because drug transferring has likely to have occurred in such patterns.

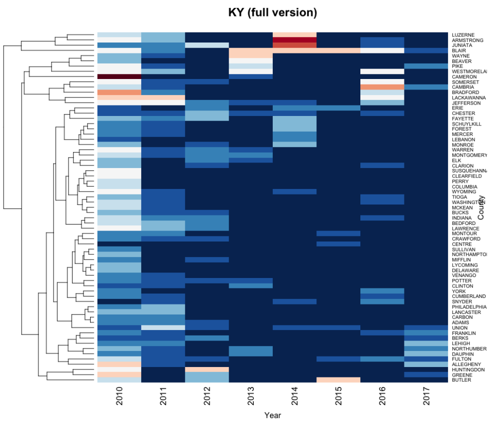



Figure 4: Sample heatmap of KY counties drug consumption from 2010 to 2017.

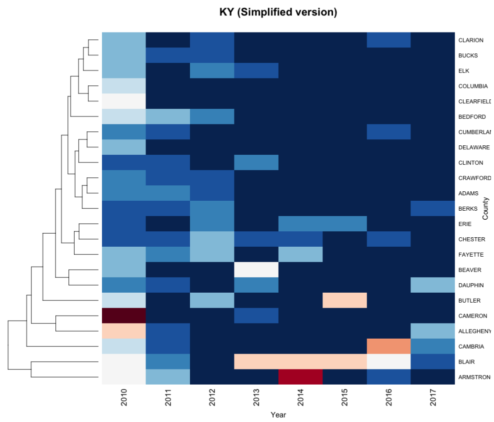



Figure 5: Simplified heatmap of KY counties drug consumption from 2010 to 2017.

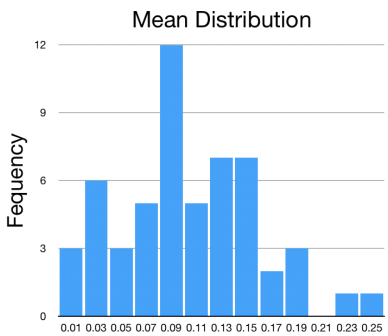



Figure 6: Average indicated substance (2010-2017) for every county

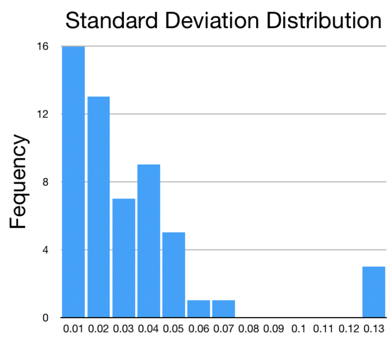



Figure 7: Standard deviation of indicated substance (2010 - 2017) for every county

To be more rigorous, we defined two thresholds to determine whether they are numerically significant. The two thresholds we used are the average and standard deviation. As shown in Figure 6, we calculated the average drug consumption of each county.

We kept the highest drug-consumption counties, because they are most likely to be the drug sources.The standard deviations of counties are determined in the same way (Figure 7).

We removed the counties that do not have statistics higher than both thresholds. For our BNs, we used distance between counties as our initial factor in determining drug transfer relationships. They are calculated with the Distance Formula:

$$D = \sqrt{(x_1 - x_2)^2 - (y_1 - y_2)^2} \tag{4}$$

Because the distances are relatively short compared to arc distances on Earth, we did not use the spherically shaped Earth as our model, and did not use the great-circle distance to calculate the distances. The Distance Formula should be sufficient for our purpose.

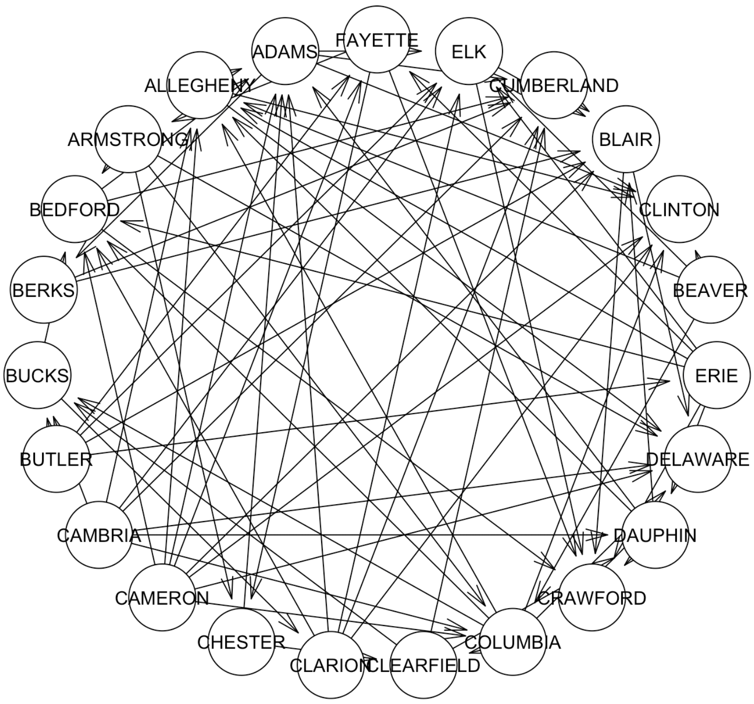



Figure 8: Graph created with BNs. The vertices of the network represent counties, and the edges of the network represent drug transfer.

As shown in Figure 8, each county relates to another. For each county, we calculated the sum of distances for all relationships one county has with another. Counties are listed from the least to the greatest sum. This gives us a rank from the most influenced to the least influenced drug county. We illustrated these values by creating circles on the counties with the greatest values (Figure 9). The area of circles is directly related to the magnitude of values.

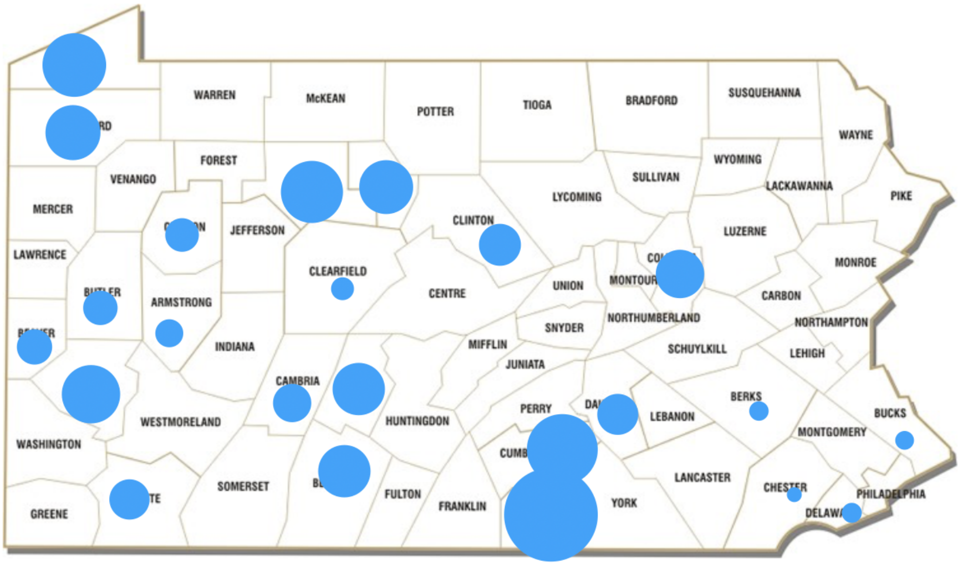



Figure 9: Counties in Pennsylvania with circles representing opioids influence. The area of circles is directly related to the opioids influence.

## 4.3 Analysis and Prediction of Top Four Consumed Drugs

Out of all 69 types of opioids from the U.S. Census Bureau, we find and analyze the top four drugs that have the highest possibility of abuse across counties. The top four consumed drugs are listed with consumption percentage out of other drugs.

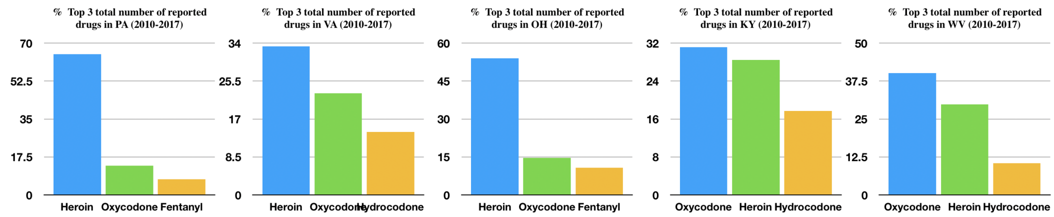



Figure 10: Graphs with the highest three opioid consumption for the five states. Heroin, Oxycodone, Fentayl, and Hydrocodone are the opioids with the highest consumption.

As shown in Figure 10, the high consumption for Heroin, Oxycodone, Fentanyl, and Hydrocodone suggests they are most likely to have been consumed under addiction. We predict the behaviors of these substance consumption in future years with the the Grey Box Model.

## 4.4 Grey Box Model

The Grey Box model[2] completes a model by combining a theoretical structure with data. We chose this model because of its several advantages. First, in dealing with a shortage of feature data, Grey Box model can complete irregular data into patterns with stronger correlation. Second, it suits the lack of enough historical data and its low reliability. It is, however, suitable up to mid-interval of the total interval of data. For the theoretical structure, the Grey Box model assumes the data to consist of sets of feed vectors f, product vectors p, and operating condition vectors c. Vectors c usually extract values from vectors f, with other values. Generally, the Grey Box is modeled with the following relation:

$$m(f, q, p) \tag{5}$$

where function m is a vector and gives errors between data p and model predictions. Vector q fills the variable parameters in the model's unknown parts. The parameter q can be determined with the relation

$$q = Ac \tag{6}$$

where A is a matrix of non-zero unknown coefficients, and c is a linear regression that values of operating conditions to non-linear relations. Coefficients in A are determined by minimizing the error terms over the data in:

$$m(f, p, Ac) \tag{7}$$

The remaining coefficients are determined with non-linear least squares by minimizing equation (7). It is possible to determine the minimizing coefficients with methods such as simulated annealing and evolutionary algorithms.

## 4.5 Solutions and Results of Grey Box Model

Using Grey Box model, we arrived at the following results. For each opioid type, we exhibit trends of Total Consumption in Amount and Percentage vs Time, Consumption by Amount in 5 States vs Time, and Consumption by Percentage in 5 States vs Time. As indicated above, Grey Box model is suitable to relatively short prediction. We therefore forecasted the values up to 2026 or 2027. The following graphs exhibit trends in future years, with some concise analysis and interpretation.

### 4.5.1 Analysis of Heroin Consumption

As shown in Figure 11, both the amount and percentage of Heroin consumption steadily increase. However, the amount increases at a faster rate than percentage. This indicates that Heroin would have become more addictive, but another opioid, or other opioids would have become more widespread. We therefore can expect another opioid to emerge as a more popular addictive.

As shown in Figure 12, Heroin consumption increase most dramatically in OH. Other states show increase in consumption as well, but WV shows decrease with its small consumption to begin with.

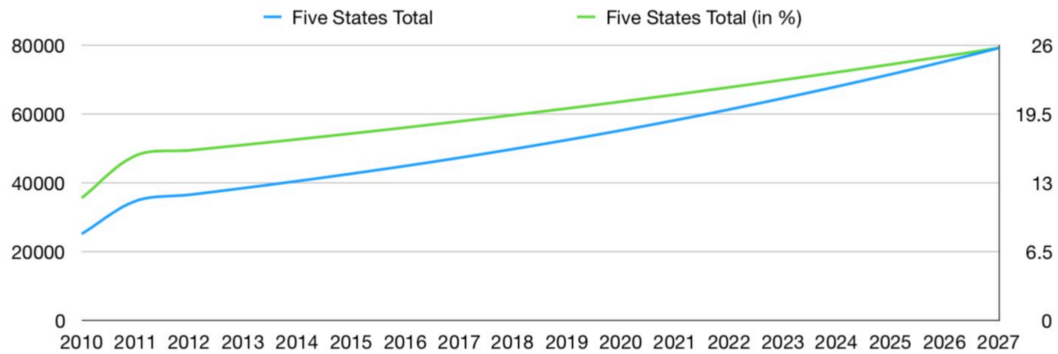



Figure 11: Graph prediction of Oxycodone consumption up to 2027. Blue line is the predicted trend of consumption amount in all states.

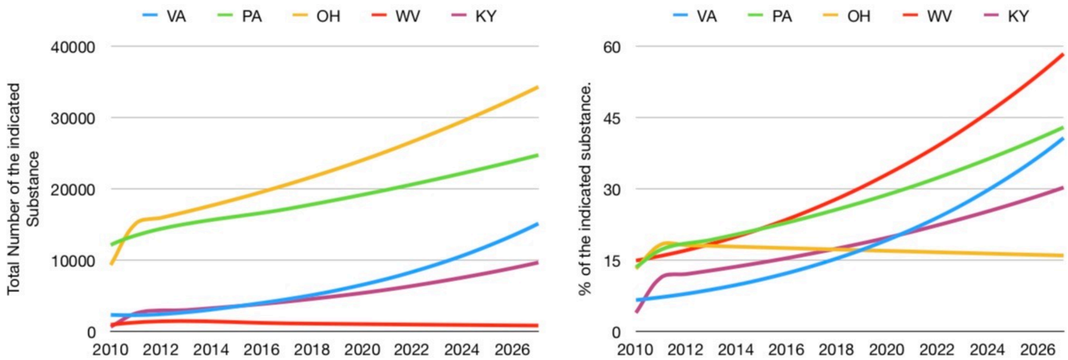



Figure 12: Left is graph prediction of Heroin consumption amount up to 2026. Right is graph prediction of Heroin consumption percentage up to to 2026.

The rank in rates of increase for Heroin consumption percentage shows an interesting contrast. Whereas OH Heroin consumption amount increase was the highest, OH Heroin percentage decrease (Figure 12). This indicate that, while Heroin consumption amount increases dramatically, another opioid, or other opioids consumption outnumbered that of Heroin. This would hint at a higher consumption of opioids in OH than other states. The increase in Heroin consumption amount and the decrease in Heroin consumption percentage suggests that OH has the most opioids consumption compared to other states. While WY Heroin consumption amount decreases, its consumption percentage increases. In other words, when Heroin is the most consumed opioid, the consumed amount decreases. This combination suggests the overall decrease in WY opioid consumption.

#### 4.5.2 Analysis of Oxycodone Consumption

As shown in Figure 13, both Oxycodone consumption amount and percentage decreases steadily. This can have two implications. Oxycodone consumption amount would decrease with an overall decrease in opioid consumption, or it would become a less popular opioid compared to other opioids. However, in either case, it is likely that Oxycodone is not an abused and addictive substance.

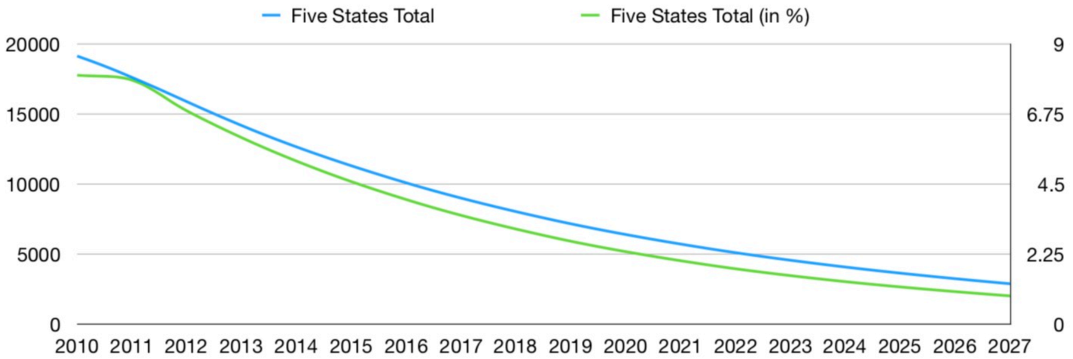



Figure 13: Graph prediction of Oxycodone consumption up to 2027. Blue line is the predicted trend of Oxycodone consumption amount in all states.

As shown in Figure 14, both Oxycodone consumption amount and percentage decreases. This is another evidence for our inference that Oxycodone is not an abused and addictive substance.

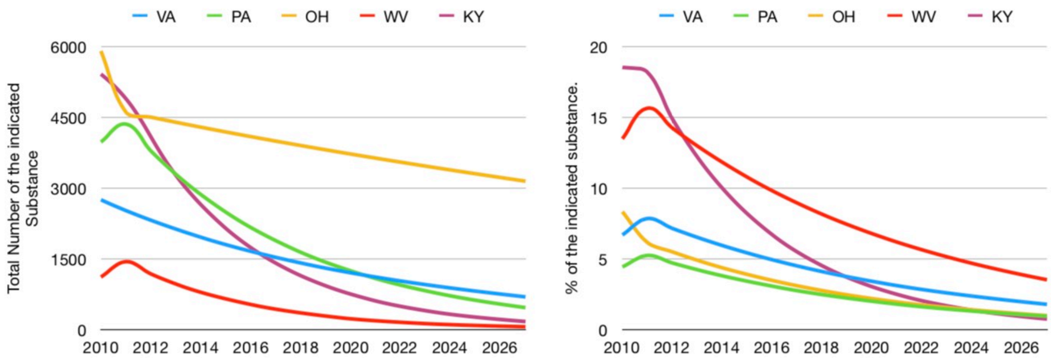



Figure 14: Left is graph prediction of Oxycodone consumption amount up to 2026. Right is graph prediction of Oxycodone consumption percentage up to to 2026.

### 4.5.3 Analysis of Fentanyl Consumption

As shown in Figure 15, both Fentanyl consumption amount and percentage would increase exponentially. Although data collected in years up to date do not show such trend, the predicted trend suggest that Fentanyl is a highly addictive substance.

As shown in Figure 16, both Fentanyl consumption amount and percentage would increase exponentially in all states. Especially in PA, Fentanyl consumption amount increases the most rapidly. PA opioids consumption amount are also ranked top 2 or 3 in Heroin or Oxycodone. These predictions suggest PA consumes a great amount of addictive opioids.

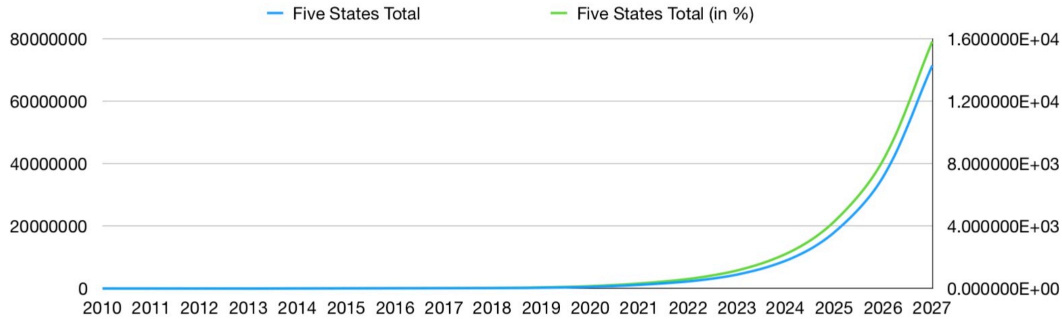



Figure 15: Graph prediction of Fentanyl consumption up to 2027. Blue line is the predicted trend of Fentany1 consumption amount in all states.

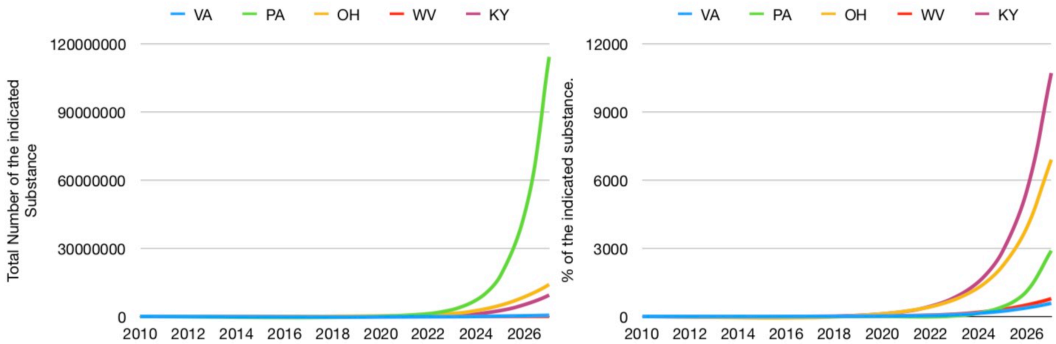



Figure 16: Left is graph prediction of Fentanyl consumption amount up to 2026. Right is graph prediction of Fentanyl consumption percentage up to to 2026.

#### 4.5.4 Analysis of Hydrocodone Consumption

As shown in Figure 17, both Hydrocodone consumption amount and percentage would decrease. Similar to the analysis to Oxycodone, these trends suggest Hydrocodone is not an addictive substance.

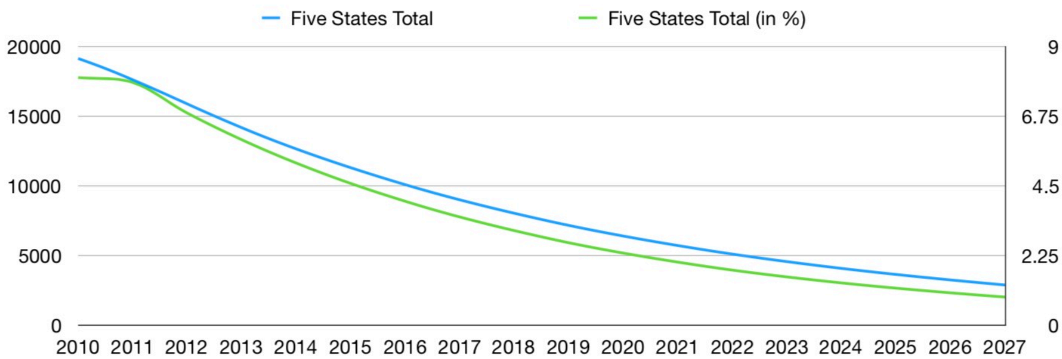



Figure 17: Graph prediction of Hydrocodone consumption up to 2027. Blue line is the predicted trend of Hydrocodone consumption amount in all states.

As shown in Figure 18, Hydrocodone consumption amount and percentage steadily decreases in all states. This is another evidence that Hydrocodone is not an addictive substance.

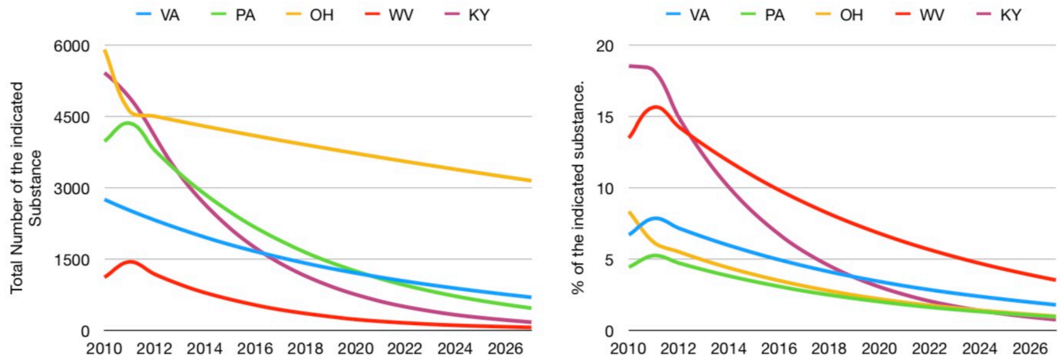



Figure 18: Left is graph prediction of Hydrocodone consumption amount up to 2026. Right is graph prediction of Hydrocodone consumption percentage up to to 2026.

### 4.5.5 Conclusion of Opioids Consumption

The overall opioids consumption trends follow ordinary patterns, either all increasing or decreasing, in all fields (e.g., Hydrocodone consumption amount, etc.), with exception to OH Heroin Consumption Percentage (See section 4.5.1 Analysis of Heroin Consumption). Our prediction models suggest that Oxycodone and Hydrocodone are not addictive, while Heroin and Fentanyl are addictive. Our analysis suggests that the highest opioids consumption states are OH and PA, while WV is one of the least opioids consumption state, if not the least.

# 5 Analysis of Socio-Economic Factors

With Principal Component Analysis (PCA)[3], we scaled the influence of socio-economic factors in [0,1]. We modified the Grey Box model and applied its results in the Vector Autoregression. Our predictions suggest that the socio-economic factors give the prediction a more wave-like behavior, but the overall trend is increasing.

## 5.1 Principal Component Analysis (PCA)

The 151 features are divided into 16 major factors: Marital Status, Disability Status, Educational Attainment, Residence, Year of Entry, Grandparents, Language Spoken at Home, Ancestry, Veteran Status, School Enrollment, Fertility, Households by Type, World Region of Birth, Relationship, Place of Birth, U.S. Citizenship Status. Computers and Internet Use has many missing values, which would negatively affect the accuracy of our model, so we deleted it. To make our calculations more efficient, we decreased the number of columns in the following way:

1. Extract all data values from a state.
2. Analyze data values individually from the 16 major factors. We will take marital status as our example in the following explanation.
3. Analyze data values in 7 individual years from 2010-2016.
4. Marital Status has many sub-features, including never married, now married, except separated, separated, widowed, divorced, etc. With each sub-feature containing male and female, we have more than 10 sub-features. Each sub-feature is performed with PCA, and has a weight determined.
5. PCA calculations of the weights of the sub-features are combined to represent the weight value of the Marital Status factor.
6. Seven years of weight is averaged.
7. Repeat steps 2-6. We received 16 weight average.
8. Calculate drug ratio for each county (drug type detected amount/total drug).
9. Combine drug ratio and the 6 factors. Performing Correlation Analysis, we have the factors and drug ratios of the 16 factors (Table 1).

Table 2: Partial values for VA factor correlation to drug use. Higher value indicate higher correlation.

| Rank | Factor | Drug Ratio |
|---|---|---|
| 1 | U.S. CITIZENSHIP STATUS | 0.184463 |
| 2 | PLACE OF BIRTH | 0.12149 |
| 3 | RELATIONSHIP | 0.081251 |

After knowing the comparative relationships between the major factors, we delve in deeper analysis to find how the 151 features relate to drug consumption. Here we do not separate states in our calculations. The method of calculation is similar to the method of decreasing columns, only we do not use PCA on factors. Features are used instead. Performing the correlation analysis, we combined the 151 drug feature and drug ratio (Table 3).

## 5.2 Vector Autoregression (VAR)

Vector Autoregression (VAR)[4] is a great multivariate time series prediction model which captures the linear interdependencies of multiple time series. We applied VAR to this problem to combine the time series of possible locations and the socio-economic factors determined in section 4.2 and then predict total drug

Table 3: Partial values for VA factor correlation to drug use. Higher value indicate higher correlation.

| Factor | Correlation Coefficient |
|---|---|
| 18 to 64 years - With a disability | 0.359508 |
| 65 years and over - With a disability | 0.357106 |
| Total Civilian Noninstitutionalized Population - With a disability | 0.330771 |
| Number of grandparents living with own grandchildren under 18 years - Grandparents responsible for grandchildren | 0.292931 |
| Females 15 years and over - Divorced | 0.264919 |
| Males 15 years and over - Widowed | 0.253088 |
| Females 15 years and over | 0.244958 |
| Males 15 years and over | 0.232723 |
| PLACE OF BIRTH - Total population - Native | 0.231523 |
| Number of grandparents living with own grandchildren under 18 years - Years responsible for grandchildren - 3 or 4 years | 0.220305 |
| Percent; RESIDENCE 1 YEAR AGO - Population 1 year and over - Same house | 0.219719 |

consumption for the following years. VAR describes the change of several variables in the same time interval as a linear function. A p-th order is defined as follows[5]:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + e_t \tag{8}$$

where $y_{t-1}$ is called the i-th lag of y, c is a k×1 vector of constants, $A_i$ is a time-invariant k×k matrix and $e_t$ is a k×1 vector of error terms. For a multivariate time series, $e_t$ should be a continuous random vector that satisfies the following conditions:

$$E(e_t) = 0 \tag{9}$$

That means the expected value of the error vector is zero.

## 5.3 Solutions and Results of VAR

As shown in Figure 19, VAR predictions show more realistic details when taking into account of the socio-economic factors. The trend behave with wave-like qualities, but the overall change increases regardlessly.

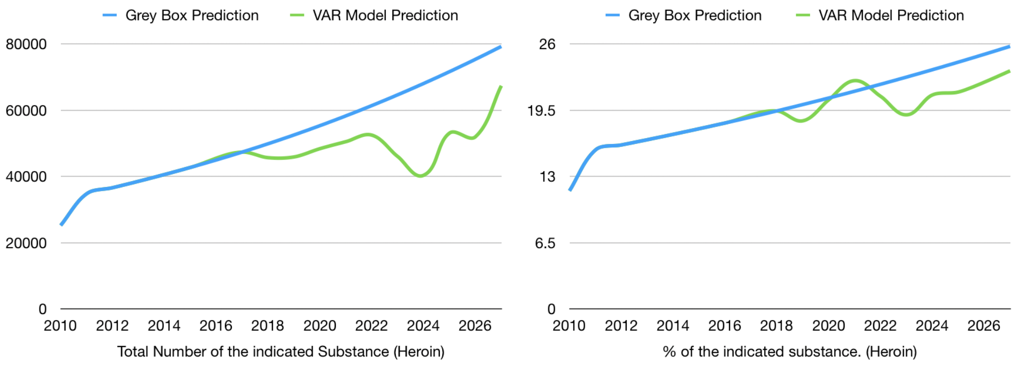



Figure 19: VAR predictions of Heroin consumption amount (left) and percentage (right)

# 6 Strategy of Countering the Opioid Crisis

## 6.1 Hypothesis and Analysis

Disability status has the highest impact on opioids abuse (Table 3). We inferred that it is the most important factor in determining the opioids consumption trends. We therefore start with the hypothesis that changing the status for the disabled will affect opioids abuse. To test this hypothesis, we start by lowering the value for the disabled 1% each year. The multiplication factor is:

$$0.99^n \tag{10}$$

where n is the number of years passed since 2018. By decreasing 1% each year, we believe the hope of such decrease has been kept within bound, and it is realistic to achieve such goals.

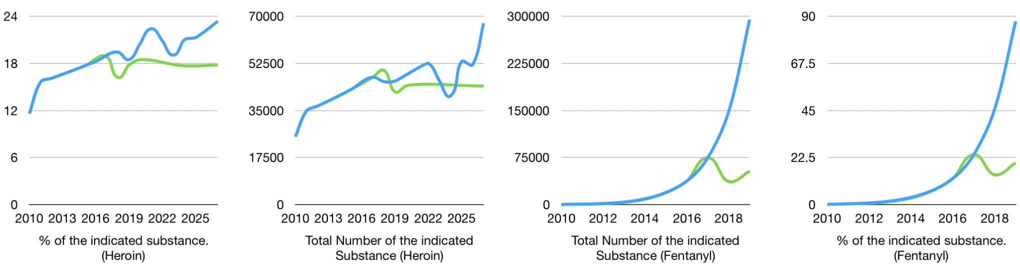



Figure 20: The left two graphs depict trends of Heroin consumption amount and percentage. The right two graphs depict trends of Fentanyl consumption amount and percentage.

where the blue lines are VAR predictions with socio-economic factors. Green lines are VAR predictions with the decrease in the disabled opioids consumption. We decreased the number of disabled opioids consumption starting from 2010 by 1%. This illustrates that there is a trend of steady decrease in Heroin consumption, but Fentanyl consumption gradually gains the trend to increase, likely because it is a highly addictive substance. Although the 1% decrease has been modeled since 2010, it does illustrate the impact of the decrease in the disabled consumption. We therefore expect similar decrease in overall opioids consumption if the 1% decrease has been modeled in the future.

### 6.2 Financial Support for Institutions Helping the Disabled

We propose the 1% decrease in disabled opioids consumption with government funding for institutions that help the disabled. The institutions may help the disabled to rehabilitation. Factors that will inhibit such successes include the lack of funding for such institutions. For, if opioids abuses are not beneficial to high-level industries, neither would rehabilitation centers for the disabled. Such funding may be seen as unbeneficial. However, opioids abuse is a phenomenon that spreads across counties within a state, and possibly across the nation. Therefore, though seemingly not beneficial in the short-run, fundings for the stop of opioids spread can benefit the U.S. in the long-run. First, by restoring social stability and increasing work forces. Second, by reducing monies spent on health care.

## 7 Evaluation of Our Models

### 7.1 Strengths of Our Models

- We utilize the Bayesian network to model the drug transmission process within counties, and when building the Bayesian network, we take the distance between the counties in to consideration, which will be a more reasonable and accuracy way to describe the drug transmission process.
- For the forecasting model, at first we use the grey box model, which can complete irregular data into patterns with stronger correlation. And compare to other forecasting model, it suits the lack of enough historical data and its low reliability.
- We divide the features in the given dataset into several factors, and utilize Principle Component Analysis (PCA) method and weight approach to reduce the dimensions of data and get reasonable result.
- After we obtain the features and factors that have strong correlation with the drug usage, we then upgrade our forecast model by combined those feature with Vector autoregression (VAR) model. Compare to other models, it supports multiple variable forecasting by capture the linear interdependencies among multiple time series.
- We provide a strategy that has been proved effective by our forecast model.

- Our results and conclusions are supported with concrete data, detailed analysis and a variety of splendid charts.
- Repeat steps 2-6. We received 16 weight average.
- Calculate drug ratio for each county (drug type detected amount/total drug).
- Combine drug ratio and the 6 factors. Performing Correlation Analysis, we have the factors and drug ratios of the 16 factors (Table 1).

### 7.2 Weakness of Our Models

- Grey Model is only applicable to short- and medium-term forecasts. If we want a long-term forecasts, the result might be less accuracy.
- There are hundreds of Socio-Economic features in the given dataset. And the 16 factors we divided from the features may be too general.

## 8 Error and Precision

We took account of the error. After taking the average for each factor, we found our errors to be approximately 1% -2%. We infer that our predictions will also contain 1% error. For example, the green line in Figure 21 shows the prediction of the total Heroin usage for the following years after lowering the values for the disabled opioids consumption by 1% each year. The dash lines shows the margin of error for this prediction.

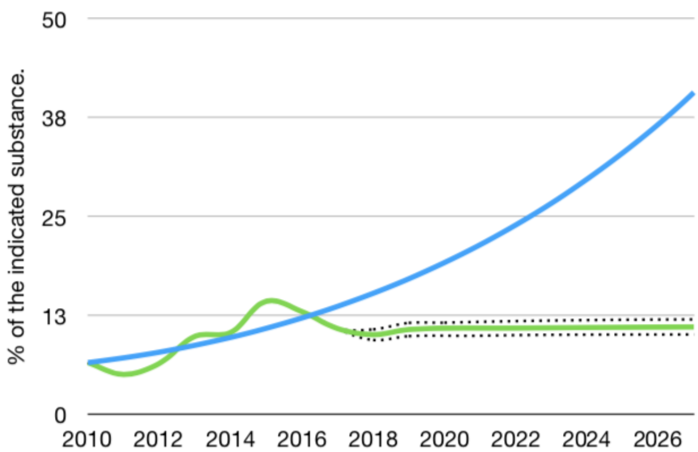



Figure 21: The prediction and margin of error of the total Heroin usage in five states.

# 9 Conclusion

Analyzing and modeling, we successfully determined the precise answers to the questions from Part 1-3, namely the locations in which certain opioids consumption take place, the socio-economic factors that relate to opioids consumption, and countermeasures.

First, we performed the Bayesian Networks modeling and mapped location-specific opioids consumption rate. Then, we used Grey Box model to complete the data in predictions up to 2027. Predictions suggest that Heroin and Fentanyl are addictive opioids that will impact future consumptions. Our model predictions suggest that in one to two years, Heroin and Fentanyl consumption will reach to roughly 1.5 times the current consumption amount.

Second, we performed the Principal Component Analysis to help us determine the impact of socio-economic factors in opioids consumption. Then, we performed the Vector Autoregression to remodel our Grey Box model and take account of the socio-economic factors in opioids consumption.

Last, we proposed that reducing the number of disabled substance abuse users will decrease the overall trend of opioid consumption. We carried out a Vector Autoregression test with a 1% decrease each year and confirmed our hypothesis. We propose funding for Institutions with opioids rehabilitation centers to aid the problem.

# References

[1] Toda, Hiro Y., and Peter C. B. Phillips. "Vector Autoregressions and Causality." Econometrica, vol. 61, no. 6, 1993, pp. 1367_1393.
www.jstor.org_stable_2951647.

[2] Nian Sao. Guide to prediction and modeling of multivariate time series. 2018.12.20
https://www.ziiai.com_blog_665

[3] Makridakis, S. , Andersen, A. , Carbone, R. , Fildes, R. , Hibon, M. , Lewandowski, R. , Newton, J. , Parzen, E. and Winkler, R. (1982), The accuracy of extrapolation (time series) methods: Results of a forecasting competition.
J. Forecast., 1: 111–153. doi:10.1002/for.3980010202

[4] Jing Yu, V. Anne Smith, Paul P. Wang, Alexander J. Hartemink, Erich D. Jarvis; Advances to Bayesian network inference for generating causal networks from observational biological data, Bioinformatics, Volume 20, Issue 18, 12 December 2004, Pages 3594âĂŞ3603,
https://doi.org/10.1093/bioinformatics/bth448

[5] Ma, W., & Wang, R. (2015). Traffic flow forecasting research based on Bayesian normalized Elman neural network. 2015 IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE).
doi:10.1109/dsp-spe.2015.7369592

[6] Zeng, B., & Luo, C. (2017). Forecasting the total energy consumption in China using a new-structure grey system model. Grey Systems: Theory and Application, 7(2), 194-217.
doi:10.1109/dsp-spe.2015.7369592

[7] Zeng, B., & Luo, C. (2017). Forecasting the total energy consumption in China using a new-structure grey system model. Grey Systems: Theory and Application, 7(2), 194_217.
doi:10.1109_dsp-spe.2015.7369592

[8] Gatignon, H. (2009). Reliability Alpha, Principle Component Analysis, and Exploratory Factor Analysis. Statistical Analysis of Management Data, 29-57.
doi:10.1007_978_1_4419_1270-1_3

[9] Song, H., & Witt, S. F. (2000). Vector autoregression (VAR) and cointegration. Tourism Demand Modelling and Forecasting, 91-122.
doi:10.1016/b978-0-08-043673-9.50009-9

[10] Orme, J. G., & Combs-Orme, T. (2009). Regression with a Polytomous Dependent Variable Regression with a Polytomous Dependent Variable. Multiple Regression with Discrete Dependent Variables, 91-122.
doi:10.1093/acprof:oso/9780195329452.003.0003

# Appendices

Table 4: Partial values for VA factor correlation to drug use. Higher value indicate higher correlation.

| Rank | Factor | Drug Ratio |
|---|---|---|
| 1 | U.S. CITIZENSHIP STATUS | 0.184463 |
| 2 | PLACE OF BIRTH | 0.12149 |
| 3 | RELATIONSHIP | 0.081251 |
| 4 | WORLD REGION OF BIRTH OF FOREIGN BORN | 0.07785392 |
| 5 | HOUSEHOLDS BY TYPE | 0.07168105 |

Table 5: Partial values for OH factor correlation to drug use. Higher value indicate higher correlation.

| Rank | Factor | Drug Ratio |
|---|---|---|
| 1 | MARITAL STATUS | 0.37674187 |
| 2 | RELATIONSHIP | 0.10035144 |
| 3 | RESIDENCE | 0.09209107 |
| 4 | HOUSEHOLDS BY TYPE | 0.04332892 |
| 5 | WORLD REGION OF BIRTH OF FOREIGN BORN | -0.0119751 |

Table 6: Partial values for PA factor correlation to drug use. Higher value indicate higher correlation.

| Rank | Factor | Drug Ratio |
|---|---|---|
| 1 | ANCESTRY | 0.223277497 |
| 2 | GRANDPARENTS | 0.188308211 |
| 3 | HOUSEHOLDS BY TYPE | 0.178967324 |
| 4 | DISABILITY STATUS | 0.152852919 |
| 5 | PLACE OF BIRTH | 0.057299206 |

Table 7: Partial values for KY factor correlation to drug use. Higher value indicate higher correlation.

| Rank | Factor | Drug Ratio |
|---|---|---|
| 1 | SCHOOL ENROLLMENT | 0.17447842 |
| 2 | HOUSEHOLDS BY TYPE | 0.058423833 |
| 3 | EDUCATIONAL ATTAINMENT | 0.021234597 |
| 4 | ANCESTRY | 0.011816553 |
| 5 | VETERAN STATUS | -0.000298374 |

Table 8: Partial values for WV factor correlation to drug use. Higher value indicate higher correlation.

| Rank | Factor | Drug Ratio |
|---|---|---|
| 1 | ANCESTRY | 0.3250101 |
| 2 | EDUCATIONAL ATTAINMENT | 0.2994007 |
| 3 | YEAR OF ENTRY | 0.2934181 |
| 2 | SCHOOL ENROLLMENT | 0.2781926 |
| 3 | VETERAN STATUS | 0.254952 |