Team Control Number

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

**1922154**

Problem Chosen

**C**

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

**2019**
**MCM/ICM**
**Summary Sheet**

# The Current Status, Future and Strategy of Opioid

Summary

The United States is experiencing a crisis regarding the abuse of opioids which poses a great threat to the development prospects of the United States. Based on the idea of cellular automata, we not only describe the spread and characteristics of reported synthetic opioid and heroin cases in Ohio, Kentucky, West Virginia, Virginia, and Pennsylvania, but also develop a possible strategy countering the opioid crisis.

We define a county and the nearest *k* counties around it as an “environment”. Based on the idea of KNN, we determine the *m* “environments” that are most similar to the “environment” of the county, and then use the cellular automata to predict the number of cases in the county next year with the growth rate of *m* “environments”. At the same time, we define the opioid incidents concentration index (CI) to characterize the degree of aggregation of cases by reference to the HHI index. Finally, we obtain the distribution of synthetic opioids and heroin incidents in five states. Cases are still concentrated in transportation hubs and there is a tendency to spread. Heroin spread to the southwest in Kentucky with Lexington as the center and has a tendency to spread throughout Pennsylvania and Virginia. Based on historical data and prediction, we determine the drug identification threshold levels for each state. In 2026, Ohio will reach its threshold of 120,000, making it difficult for the government to control the amount of opioid use and the speed of spread.

In order to determine whether certain socio-economic factors have a significant impact on the trend of opioid usage, we select the first 25% and the last 25% of the data for all state cases in 2010-2016 for analysis of variance if the data passes the test of variance homogeneity. Correlation analysis is performed on data that do not pass the test for variance homogeneity to determine the correlation between socio-economic factors and trends in opioids usage. The final selection of significant factors is marital status, educational attainment, ancestry, and language spoken at home. Adding the important factors selected above to the “environment” similarity considerations, we have obtained a modified model that considers socio-economic factors.

Based on the above analysis, we develop a strategy contains two actions for dealing with opioid crisis. The first one is giving couples a discount on tax and mortgage rates to encourage people to marry at legal age. The other is opening a low-cost English language training institution to improve the English proficiency of non-native English speakers.

**Key words:** Opioid; Cellular automata; Concentration index; Spread; Characteristics

# MEMO

**From: Team # 1922154**
**To: Chief Administrator**
**Data: January 27, 2019**
**Subject: How to deal with the opioid crisis**

Dear chief administrator, we are honored to inform you our achievement after performing data analysis and modeling.

First, we introduce the spread and characteristics of synthetic opioid and heroin usage between the five states and their counties from 2010 to 2017. Combining the provided data with the collected latitude and longitude data, we notice that the aggregation point of opioid incidents is mainly in the areas with developed traffic, and there is a tendency to spread around. The distribution of synthetic opioid in Virginia is the most extensive, moreover, the distribution in Pennsylvania is the most concentrated. Heroin once had a tendency to spread. However, perhaps for some reason, this trend has been arrested. Nowadays Heroin is spreading again in some states, such as Virginia.

Then, we forecast the synthetic opioid and heroin usage in each county from 2017 to 2026. According to the prediction, synthetic opioids will spread throughout Kentucky in the future. And the synthetic opioids usage of counties around Washington is growing from the forecasting.

Based on our observation on provided data and Calculated data, we think about the U.S. government is concern about two points:

- The opioid usage should be restricted to a certain level.
- The spread of the opioid should be controlled within a certain range.

According to the historical data and prediction, we can identify the drug identification threshold levels to predict when and where the government’s concern will occur. For example, the threshold level of Ohio is 120,000. government’s concern has occurred in Ohio in 2026.

By analyzing the Census socio-economic data, we notice that some important variables such as marital status, educational attainment, ancestry and the language spoken will impact on the use of the opioid in each county.

Based on the above analysis, we propose a strategy that includes two actions.

- Give couples a discount on tax and mortgage rates to encourage people to marry at legal age.
- Open a low-cost English language training institution to improve the English proficiency of non-native English speakers.

Our strategy can effectively reduce opioid use.

- By taking the action 1, the opioid cases will reduce from 257496 to 231073
- By taking the action 2, the opioid cases will reduce from 257496 to 225873.

The above is the summary of our study. We sincerely hope that it will provide you with useful information.

Thanks!

# Contents

# 1 Introduction

## 1.1 Background

At present, the phenomenon of addiction and abuse of opioids in the United States is serious. The abuse of opioids not only imposes a heavy economic burden on the US government, but also affects the quantity and quality of the US workforce and the prospects for the US economy.

The DEA/National Forensic Laboratory Information System (NFLIS) of the Drug Enforcement Administration's (DEA) Office publishes an annual report on drug identification results and associated information from drug cases. Specifically, they need:

- a description of the spread and characteristics of synthetic opioids and heroin events reported between five states and their counties over time and identify possible locations where specific opioids may have begun to be used in five states.
- an analysis of important factors affecting the use or use of opioids in socio-economic data from the US Census.
- a possible strategy for countering the opioid crisis.

A large amount of literature tracks the abuse of opioids in the United States: for example, Cicero, Inciardi, and Muñoz [1] specifically described the trend of abuse of opioids in the United States from 2002 to 2004 based on the Researched Abuse, Diversion and Addiction-Related Surveillance (RADARS®) system; Volkow, Jones, Einstein, and Wargo [2] analyzed the factors that triggered the opioid crisis and its further evolution, as well as interventions to manage and prevent opioid use disorders.

However, most of the literature does not scientifically summarize its propagation patterns and distribution characteristics over time based on the data of the opioid drug identification cases in various counties, so that the future predictions cannot accurately indicate the time and place where a drug identification may be transmitted. In addition, past work has not been able to propose effective strategies to deal with the opioid crisis.

## 1.2 Planned Approach

Based on the above analysis, we propose the framework model shown in Figure 1, which can be summarized as the following steps:

- **Characteristics and Spread**

Draw heat maps and other visualizations using NFLIS data and geographic data (latitude and longitude), and analyze the spread and characteristics of the reported synthetic opioid and heroin incidents in and between the five states and their counties over time.

- **Cellular Automata Model**

With the idea of cellular automata which is the state of the next moment is determined by the surrounding and its own state, a new cellular automata model is constructed by combining the ideas of clustering and KNN. This model will fully exploit the information of historical data to achieve a more accurate simulation.

- **Analyze Socio-economic Factors**

  We plan to use statistical one-way ANOVA and correlation analysis to find socioeconomic factors that have a significant impact on the model and to correct the model.

- **Identify a Possible Strategies**

  We will consider the results of the cellular automata model and the influential socio-economic factors of the analysis, and then develop a possible strategy for countering the opioid crisis. The model will also be used to verify the effectiveness of the strategy.

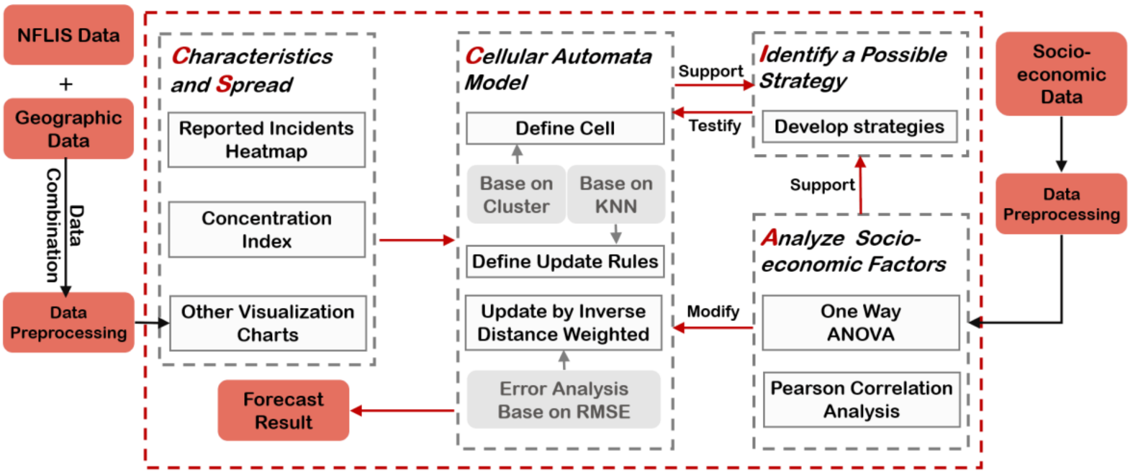


**Figure 1:** Model framework

## 2 Terminology, Symbols and Assumptions

### 2.1 Terms

- **Opioids[3]:** medically they are primarily used for pain relief, including anesthesia and they are also frequently used non-medically for their euphoric effects or to prevent withdrawal.
- **Heroin[4]:** an opioid most commonly used as a recreational drug for its euphoric effects. It is generally illegal to make, possess, or sell heroin without a license.
- **HHI[5]:** the Herfindahl-Hirschman index is a statistical measure of concentration. For example, it can be used to measure the market concentration. It is calculated by squaring the market shares of all firms in a market and then summing the squares.

### 2.2 Symbols

**Table 1:** Variable description

| Symbol | Definition |
|---|---|
| $C_i$ | The $i^{\text{th}}$ county |
| $\overrightarrow{C_i^n}$ | The environment(vector) related to $i^{th}$ county in the $n^{th}$ year |
| $r_i^n$ | The growth rate of opioid usage in the $i^{th}$ county in the $n^{th}$ year |

| | |
|---|---|
| $RMSE^{y}_{(k,m)}$ | The error in $y^{th}$ year when the value of parameters are k and m |
| $CI$ | The concentration index |
| $MS$ | Marital status |
| $EA$ | Educational attainment |
| $AN$ | Ancestry |
| $LS$ | Language spoken at home |

## 2.3 General Assumptions

- **Assumption 1:** The change in the number of a county opioid incidents is greatly affected by the surrounding counties, and the historical data can reflect the development of opioids to a certain extent.
  Reason: This assumption was made to ensure the validity of the cellular automata model we constructed.

- **Assumption 2:** The government will not have excessive rectification of opioids from now until 2026, and the changes in the opioids of each county will follow the historical law of 2010-2017.
  Reason: The reason for this assumption is to ensure the validity of the results predicted by the model to some extent.

- **Assumption 3:** The data used in this paper is realistic and accurate to a certain degree.
  Reason: Although the data is incomplete and there are some tolerable errors in the statistics, we make this assumption to ensure an effective solution.

- **Assumption 4:** Counties, which are not involved in NFLIS Data is not considered in our model.
  Reason: We believe the counties not involved in NFLIS Data are of little significance to the problem studied

# 3 Spread and Characteristics of Opioid Incidents

## 3.1 Preprocess Data

### 3.1.1 Missing value processing

The first file (MCM_NFLIS_Data.xlsx) contains most of the county's drug identification counts in year 2010-2017, but some counties still have missing data for a certain year or even years. We suspect that the county has a drug identification count in the absence of these years, but the name of the drug identified cannot be determined. Therefore, we fill in the missing value of variable *County total count of all substances identified* as follows:

$$X_i = \frac{X_{i-1} + X_{i+1}}{2} \tag{1}$$

where $X_i$ indicates the missing total count of all substances identified in the county in the $i^{th}$ year, $X_{i-1}$ indicates the total count of all substances identified in the previous year, $X_{i+1}$ indicates the total count of all substances identified in the following year.

**Notes:** If the county has more than three missing data, we believe that the county data is invalid. We give up filling in missing values and abandon them.

### 3.1.2 Geographic coordinate acquisition

In order to see the geographical distribution of the submitted cases, we obtained five states: the latitude and longitude data of all counties in Ohio, Kentucky, West Virginia, Virginia, and Pennsylvania from the United States Cities Database website [6]. And then we calculate the distance between each county using the *Haversine* formula [7]. Suppose the latitude and longitude of the two counties are ($\varphi_1$, $\lambda_1$) and ($\varphi_2$, $\lambda_2$), respectively.

$$d = 2r \cdot arcsin(\sqrt{sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + cos\,\varphi_1 \cdot cos\varphi_2 \cdot sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}) \tag{2}$$

where *d* is the distance between the two counties, *r* is the radius of the earth.

### 3.1.3 Overview of drug cases distribution

We draw heat maps using the latitude and longitude data of each state and the data of drug identification counts for narcotic analgesics (synthetic opioids) and heroin in each of the five states, to have a general understanding of the distribution of reported cases.

Take the case reported in 2010 as an example (shown in Figure 2). It can be roughly seen from the heat map that the proportion in the transportation hubs and along the lake and coastal areas is high.

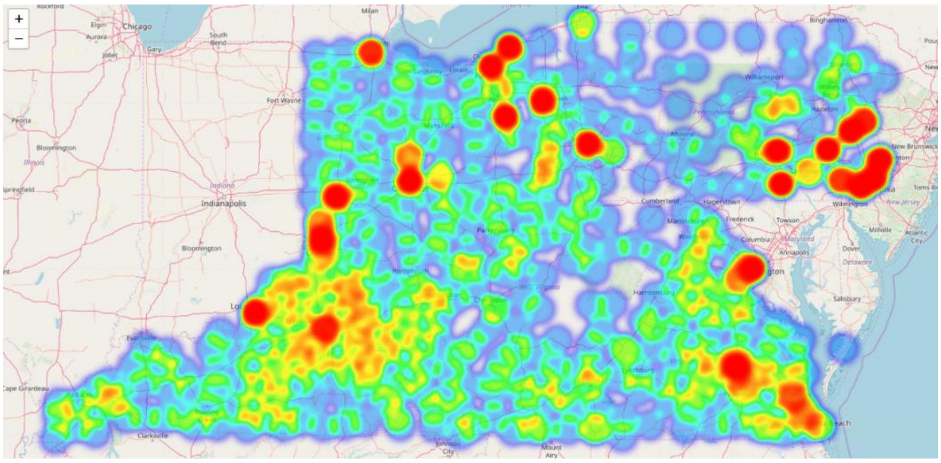


**Figure 2:** Distribution of drug identification cases in 2010

## 3.2 Spread of Opioid Incidents base on CA Model

### 3.2.1 Introduction to the idea of method

To help us understand the spread of opioid usage between the five states and their counties in the past, we propose a model to simulate the use of opioids over the past eight years in various regions. The simulation results of the model are then used to identify any possible locations in five states that may have begun to use a specific opioid.

Based on the analysis of the problem and data, we summarize the following challenges:

- The model should be able to reflect the interaction between the use of opioid each county.
- The model should be able to reflect the impact of the historical development of each county on its future.
- Models must be able to simulate changes in the number of opioid cases in all counties.

In view of these challenges, we adopt the Cellular Automata [8] (CA), a grid dynamics model, in which time, space, and state are all discrete and have the ability to simulate the evolution process of complex systems. Cellular Automata is a widely used model to analyze the spread problems [9]. In this case, the map is divided into cells, and each county occupies a separate cell. A cell records the total drug reports of the county. The state of the cell update based on its current state and the current state of the surrounding cells. We apply the self-defined update rules to simulate the evolution of each county's opioid usage in our model.

### 3.2.2 Attributes of a Cell

In our model, each cell can only represent at most one county. A cell has 3 attributes:

- An integer $c$ (0 or 1) to represent the cell status, 0 for no county, 1 for one county
- Current number of opioid cases in the county $C_i$.
- The rate of change in the number of opioid cases in the county that year $r_i$

The other two properties only make sense when $c$ is non-zero. In each step of the simulation, $C_i$ is updated by the $r_i$. Self-defined rules are introduced in following sections.

### 3.2.3 Self-defined Rules

The key to the cellular automata that can be used to describe the use of opioids in each county is to develop rules that are close to reality. In this case, we develop update rules based on given historical data.

First we calculate the distance between them by the latitude and longitude of each county. Then we take each county and its nearest $k$ counties as a set of vectors what we called "environment " based on the idea of cluster. The mathematical expression for each set of vectors is as follows:

$$\overrightarrow{C_i^n} = (r_i^n, C_i^n(0), C_i^n(1), C_i^n(2), C_i^n(3), \cdots, C_i^n(k))$$

where $\overrightarrow{C_i^n}$ is the environment(vector) related to $i^{th}$ county in the $n^{th}$ year, $r_i^n$ is the growth rate of opioid usage in the $i^{th}$ county in the $n^{th}$ year (growth rate can be negative). $C_i^n(0)$ is the opioid usage of itself $C_i^n(1)$ represents the opioid usage of the county with the shortest distance from $i^{th}$ county in the $n^{th}$ year. A set of vectors can be generated based on the annual data of each county.

So we have 3003 groups environment(vector). We believe that the growth in the number of opioid cases in a county is determinable. It can be found similar environment to determine the growth rate of the county's opioid cases in historical data. Before we look for the similar environment, we're going to do something with environment(vector). We sort $C_i^n(1), C_i^n(2), C_i^n(3), \cdots, C_i^n(k)$ in ascending order. If we do not make ascending order ranking to the $C_i^n(k)$, we will amplify the error. For example, we think $(r_i^n, 1,2,3,4,5)$ should equal to the $(r_j^n, 1,5,4,3,2,1)$. But when we calculate the Euclidean distance between them, the result shows that they are different.

We can calculate the Euclidean distance between each environment(vector) $\overrightarrow{C_i^n}$. The Euclidean distance calculation formula is as follows:

$$d_{ij} = \sqrt{\sum_{i=0,j=0}^{k} \left(C_i^n - C_j^n\right)^2} \tag{3}$$

Based on the idea of KNN, we use the formula to find the *m* vectors closest to $\overrightarrow{C_i^n}$. Based on the $r_i^n$ in the *m* vectors, we can determine the annual growth rate of the number of opioid cases in $i$ county in 2010 by inverse distance weighting calculation. The inverse distance weighting formula is as follows:

$$w_j = \frac{\frac{1}{d_{ij}}}{\sum_{j=1}^{m} \frac{1}{d_{ij}}} \tag{4}$$

where $w_j$ is the weight of each vector's influence on the growth rate in $i$ county.

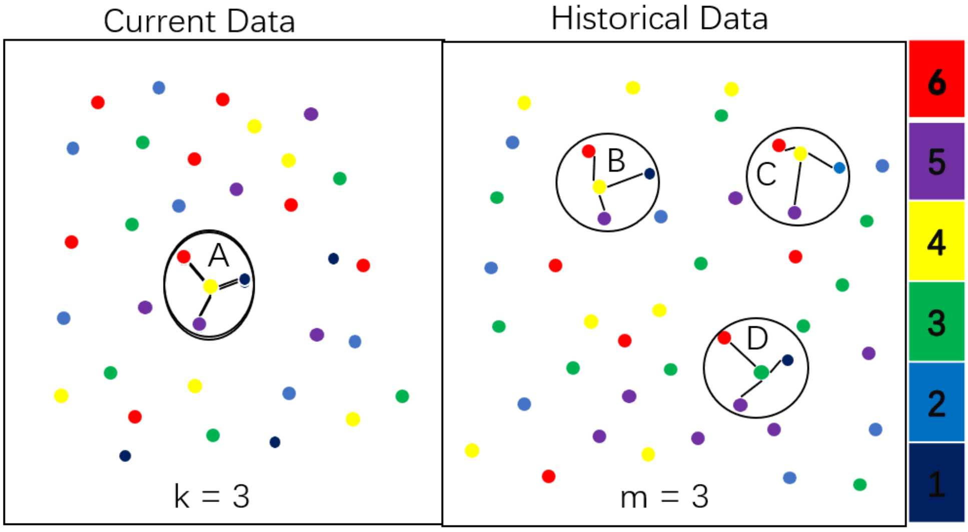


**Figure 3**: Updating Rule

Figure 3 shows how the rule works in practice. The points indicate counties, and the colors of the point indicate different values. In this case, we choose *k*=3 and *m*=3. So we can find environment $\overrightarrow{A^n} = [r_A^n, 4,1,5,6]$. From the historical data, the three most similar historical environments are $\overrightarrow{B^n} = \left[r_B^{n_b}, 4,1,5,6\right]$, $\overrightarrow{C^n} = \left[r_C^{n_c}, 4,2,5,6\right]$, $\overrightarrow{D^n} = [r_D^{n_d}, 3,1,5,6]$. Then we can calculate the Euclidean distance between each environment.

Finally, we can determine the growth rate $r_i$ of the county in the current year to calculate the number of specific opioid cases in the next year. The formula for calculating the growth rate is as follows:

$$r_i = \sum_{j=1}^{m} w_j r_j^n \tag{5}$$

Through the above method, we can determine the evolution rules of each cell in the cellular automata.

### 3.2.4 Concentration index (CI)

In economics, the industrial concentration is generally measured by the Hirschman index [5] which is not affected by the number and size of the company. It can better measure the changes in the concentration of the industry. Here we want to portray the concentration of opioids, so we define the opioid incidents concentration index (CI) by reference to the HHI index.

At first, we determined to take the state as the research unit. The general idea is to select all counties in the certain state as a whole to calculate the CI to represent the concentration of drug identification cases in the certain state. The formula is as follows:

$$\mathrm{CI} = \sum_{i=1}^{k} \left( \frac{C_i}{\sum_{i=1}^{k} C_i} \right)^2 \tag{7}$$

where $C_i$ is the drug identification count in the $i^{th}$ county, $k$ is the number of counties in a state.

## 3.3 Results & analysis

In section 3.3.1. We will draw conclusions of spread and characteristics of the reported synthetic opioid and heroin incidents by analyzing the result of the model to at first. We then consider heroin as a specific opioid, we will use our model to find any possible county where heroin use in five states.

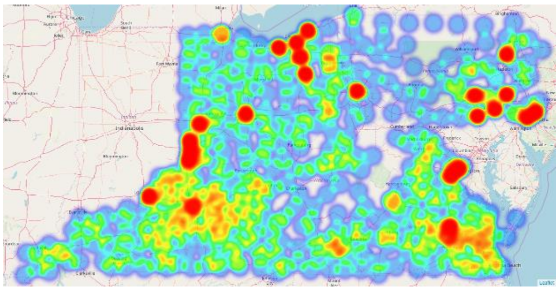
In section 3.3.2. We will talk about the government’s specific concerns. And then we will find the threshold levels to determine when and where the government’s concerns will occur in the future by using our model to predict the change of opioid usage.

### 3.3.1 Spread and Characteristics

To help us have a good grasp of opioid usage, identifying the way of opioid spread is very important. We have already built CA model and have defined our rule to change the status of cell. Based on the historical data, we use our model to predict the spread of opioids from 2017 to 2026.

- **Spread**

The result of synthetic opioid and heroin spread in the future is solved by the CA model. The spread of synthetic opioid in 2017 and 2020 are shown in Figure 4:



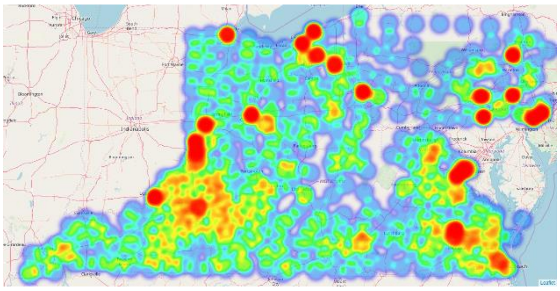
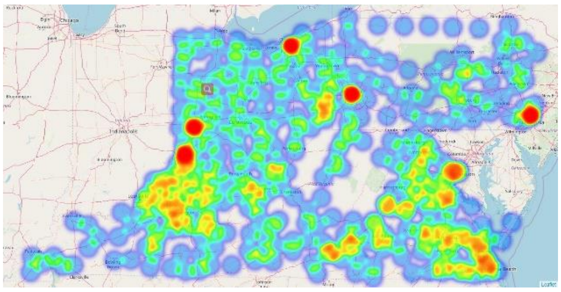


**Figure 4** (a) 2017 synthetic opioid **Figure 4** (b) 2020 synthetic opioid

By comparing Figure 3(a) and Figure 3(b), We find that some areas have begun to form new aggregation points. At the same time, some of the old aggregation points are transferred or disappeared. The aggregation point is still mainly in the areas with developed traffic, and there is a tendency to spread around. According to the prediction, synthetic opioids will spread throughout Kentucky in the future. And the synthetic opioids usage of counties around Washington is growing from the forecasting.

The spread results of heroin in 2017 and 2020 are shown in Figure 5.



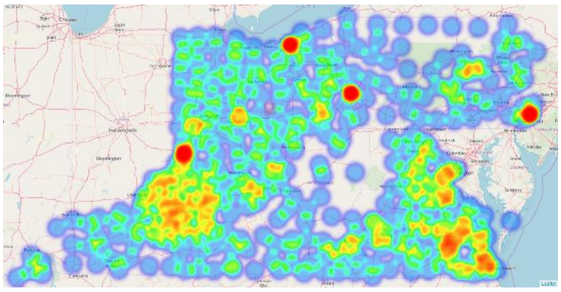


**Figure 5** (a) Heroin opioid in 2017 **Figure 5** (b) Heroin opioid in 2020

By comparing Figure 5(a) and Figure 5(b), we find that heroin spread to the southwest in Kentucky with Lexington as the center according to the prediction. Based on the forecasting, it can be seen that the aggregation points gradually disappear in the Ohio. That's probably because Ohio is starting to crack down on drugs. But in Pennsylvania and Virginia, heroin still has tendency to spread across wide areas of these two states.

- **Characteristics**

In order to portray the characteristics of the synthetic opioid and heroin, we calculate the concentration index of them based on the historical data.

By calculation, we can obtain the concentration index of the synthetic opioid. The result is shown in Figure 6. We notice that the value of concentration index is periodic. According to the formula of concentration index, we know that the smaller the value, the stronger the degree of diffusion.

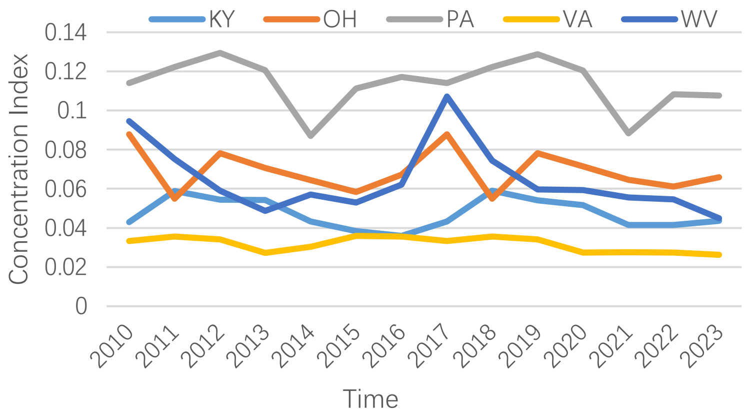


**Figure 6:** CI of the synthetic opioid

From the Figure 6, we find the distribution of synthetic opioid in Virginia is the most extensive. Moreover, the distribution in Pennsylvania is the most concentrated. From the Figure 5, we can also find the cases of opioid mainly occur in metropolis in Pennsylvania.

The result of heroin's concentration index is shown in Figure 7.

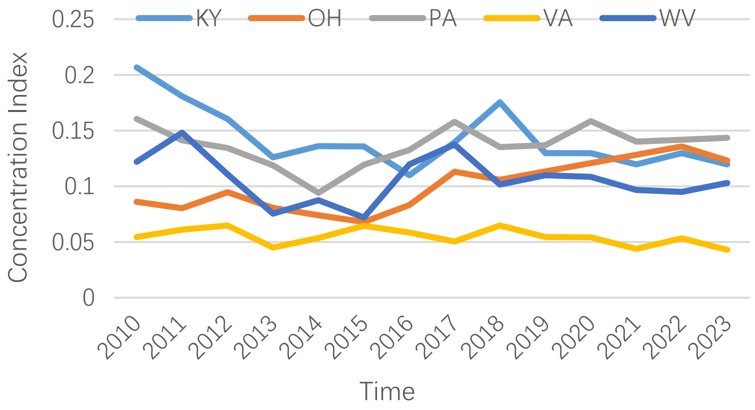


**Figure 7:** Heroin's concentration index

From the Figure 7, we notice that heroin once had a tendency to spread. However, perhaps for some reason, this trend has been arrested. According to the Figure 7, it can be seen that heroin is spreading again in some states, such as Virginia.

### 3.3.2 Concern and Occur

Based on our observation on provided data and calculated data, we think about the U.S. government is concern about two aspects:

- The opioid usage should be restricted to a certain level.
- The spread of the opioid should be controlled within a certain range.

When the number of reported cases reaches to a certain level, government will take action to limit its development.

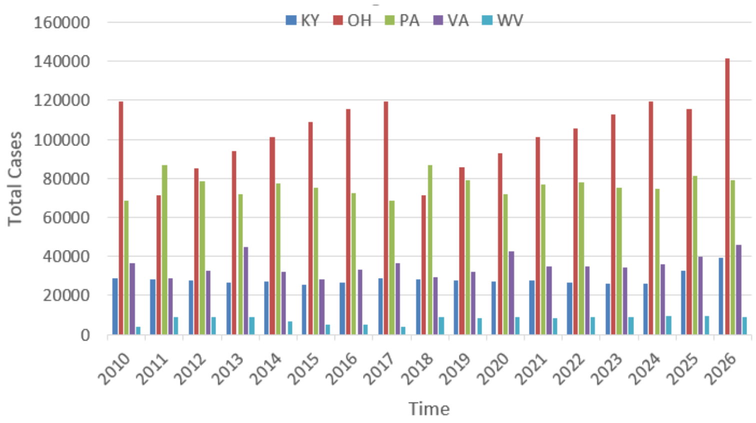


**Figure 8:** The use of opioid from 2010 to 2026

According to the historical data and prediction, we notice that the reported cases of each state is be limited to a certain. For instance, the reported cases in Ohio is be restricted under the 120000 from 2010 to 2023. Consequently, we can identify the drug identification threshold levels to predict when and where the government's concern will occur. From the Figure 8, we can determine the threshold level of Ohio is 120,000. But the number of incidents exceed the threshold level of Ohio in 2026 according to the prediction. Then we think the government's concern has occurred in 2026.

## 3.4 Sensitivity analysis of model

Based on a simple analysis, we notice that our model is sensitive to particular parameters. We run the model with different parameter values based on the provided data. Take the 2010 data as the initial state of the model.

Then we will analyze how the parameters in our model influences our result. Finally, we will determine the value of the parameter based on the results of the sensitivity analysis.

At first, we choose root-mean-square error (RMSE) to evaluate the error between the result of the CA model and the actual value under different parameter values. The formula of RMSE is:

$$RMSE_{(k,m)}^{y} = \sqrt{\frac{\sum_{i}^{n}\left(\widehat{C_\iota} - C_i\right)^2}{n}} \tag{8}$$

where $RMSE_{(k,m)}^{y}$ represents the error in $y^{th}$ year when the value of parameters are $k$ and $m$, $\widehat{C_\iota}$ is the estimated value, $C_i$ is the actual value.

In order to evaluate the accuracy of the CA model from a global perspective, we should take RMSE of every year into account. The mathematical expression that is ultimately used to estimate the accuracy of the model is:

$$RMSE_{(k,m)} = \frac{\sum_{y=1}^{a} RMSE_{(k,m)}^{y}}{a} \tag{9}$$

Then we can obtain the integrated $RMSE$ to depict the accuracy of the model.

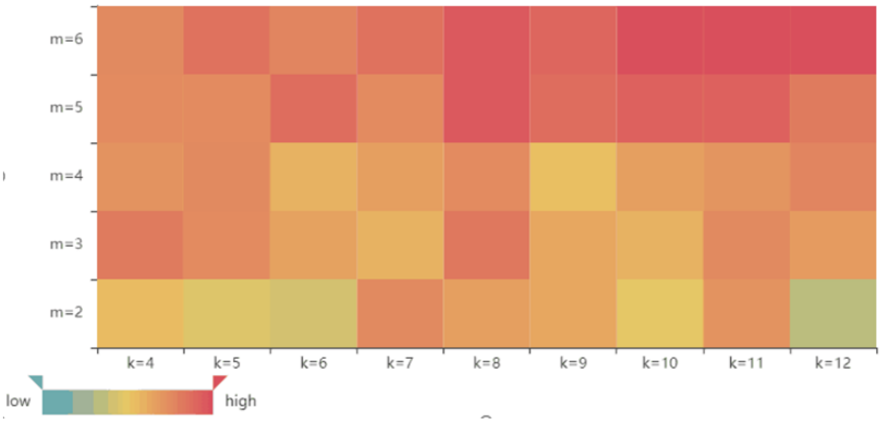


**Figure 9:** Heat map for RMSE

According to the Figure 9, we can find when *k* is 12 and *m* is 2, the RMSE is lowest, the value is 104.21. Therefore, we determine the value of *k* and *m*. The specific values are shown in Appendix I.

# 4 Model Modification Considering Socio-economic Factors

## 4.1 Preprocess Data

### 4.1.1 Data overview

We currently have a common set of socio-economic factors collected for the counties of these five states during each of the years 2010-2016 from the U.S. Census Bureau. Except for the three special identification attributes of “GEO.id”, “GEO.id2” and “GEO.display-label”, the remaining attributes include relationship, marital status, grandparents, educational attainment, ancestors, etc. in each year’s socio-economic factors data set. And each attribute has four values (Estimate、Estimate Margin of Error、Percent、 Percent Margin of Error). The overall data frame is shown in Figure 10.

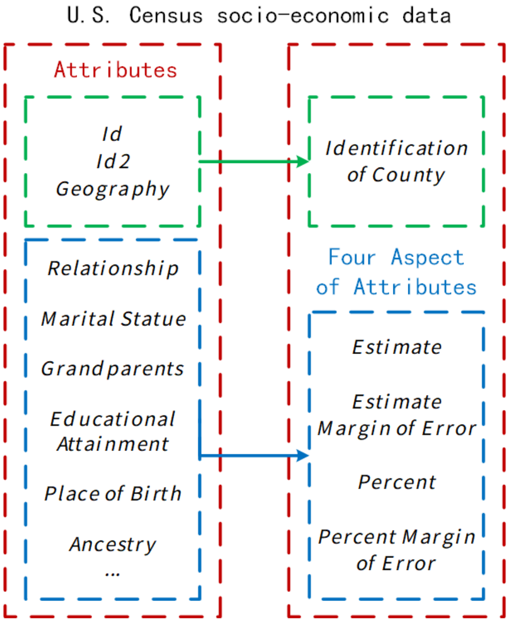


**Figure 10:** The data frame

### 4.1.2 Data selection & Analysis

Here we only study the two values of estimate and percent, and do not consider the estimation error temporarily. After selecting the attributes that have a significant impact on the model, the margin of error can be used to measure whether the attribute is valid.

In addition, we remove some of the data based on the following considerations:

- **The attribute that have an '(X)' in the margin of error column**

  We can abandon the attribute directly because an '(X)' means that the estimate is not applicable or not available.

- **Estimated data (not contain percent)**

  The issue we are considering is the impact of certain important factors in the census socio-economic data on trends in opioid usage. In general, the reason for the increase or decrease in the estimates of certain socio-economic factors is that the overall situation is changing, and the trend of use at this time will also rise or fall with the overall trend. Therefore, we believe that the estimated quantity cannot be an effective factor, while we should pay more attention to the change of its proportional structure.

After analyzing the data, we found that the data attributes of 2010-2012 are slightly different from the attributes of 2013-2017. The comparison is shown in the following table:

**Table 2:** Data comparison

| Year | Number of counties | Number of attributes | Number of attributes studied |
|---|---|---|---|
| 2010-2012 | 464 | 599 | 125 |
| 2013-2016 | 463 | 611 | 121 |

Based on the above analysis, we divided the 7-year census socio-economic data into 2010-2011 and 2012-2017 to analyze separately.

## 4.2 Important factor selection

### 4.2.1 The general idea of factors selection

Considering that we want to find out the attributes that have a significant impact on the drug cases of each county from a large number of attributes, we believe that it can be grouped according to the number of incident reports.

One-way ANOVA is used to analyze the difference of the same attribute between groups, and the attributes are initially screened out. However, the attributes selected do not necessarily have a significant impact on the number of cases. Therefore, we consider using correlation analysis to further screen and determine the final important factors.

### 4.2.2 Grouping

In order to discuss the impact of different factors on the trend of opioids' usage, we can analyze the differences in socio-economic factors in the counties with high and low frequency of opioid use. Therefore, we select two sets of extreme data in order to see the difference more clearly. The specific description is as follows:

**Table 3:** Total drug reports in all counties

| Index | 2010-2012 | 2013-2017 |
|---|---|---|

| | | |
|---|---|---|
| Mean | 501.16 | 534.45 |
| Std | 1717.67 | 1646.28 |
| Min | 0 | 0 |
| 25%(Q1) | 59 | 51 |
| 50%(Q2) | 151 | 160 |
| 75%(Q3) | 387 | 412 |
| Max | 33513 | 21761 |

We extract the first 25% (that is 0~25%) as a set of data and the last 25% (that is 75%~100%) as another set of data. Through the analysis of variance, we can analyze the difference between the two sets of data to achieve the purpose of preliminary screening.

### 4.2.3 Twofold filter

We know that the variance analysis is used to make significant differences between the two groups of sample data, and the sample data needs to satisfy the homogeneity test of variance. If the data does not pass the homogeneity test, the correlation analysis can be used to analyze the correlation.

- **ANOVA [10]**

Analysis of variance (ANOVA), also known as "coefficient of variation", was invented by the statistical expert R.A. Fisher for the significance test of the difference between two or more sample means. Its basic idea is to determine the influence of controllable factors on the research results by analyzing the contribution of variation from different sources to the total variation.

- **Correlation Analysis [11]**

Correlation analysis refers to the analysis of two or more related variable elements to measure the closeness of the two variable factors. Relevance elements need to have a certain connection or probability before they can be correlated.

Through analysis of variance and correlation analysis, we screen for socio-economic factors that have a significant impact on the trends in opioid use, as shown in Table 4:

**Table 4:** Important factors

| **Factor group** | **Explanation** |
|---|---|
| Marital status（**MS**） | Percent; Males 15 years and over - Never married（**Mn**）<br>Percent; Females 15 years and over - Never married（**Fn**） |
| Educational attainment(**EA**) | Percent; Percent bachelor's degree or higher(**Ph**) |
| Ancestry(**AN**) | Percent; Arab(**Ar**)<br>Percent; Greek(**Gr**)<br>Percent; Irish(**Ir**)<br>Percent; Italian(**It**)<br>Percent; Russian(**Ru**)<br>Percent; Ukrainian(**Uk**) |
| Language spoken at home (**LS**) | Percent; Language other than English - Speak English less than “very well”(**Lo**) |

## 4.3 Model Modification

By analyzing the Census socio-economic data, we notice that some variables will impact on the use of the opioid in each county. After the analysis in the previous section, we find the important variables are about the marital status, educational attainment, ancestry and the language spoken. These variables have an impact on the opioid usage. Accordingly, we then add them to the cellular automata model to improve the accuracy of our model. The specific variables are listed in Table 4.

In cellular automata model, the basic idea of iterative rules is to rely on a similar environment to determine the current growth rate. Hence, we take MS, EA, AN and LS into environment. However, there is a question we find is the value of the variables is very small. It will lead us to ignore the effects of these new variables in our calculations. In order to solve this problem, we redefined the formula of the variables:

$$MS = \frac{1}{Mn} + \frac{1}{Fn} \tag{10}$$

$$EA = \frac{1}{Ph} \tag{11}$$

$$AN = \frac{1}{\mathrm{Ar}} + \frac{1}{Gr} + \frac{1}{Ir} + \frac{1}{It} + \frac{1}{Ru} + \frac{1}{Uk} \tag{12}$$

$$LS = \frac{1}{Lo} \tag{13}$$

Then, we add these variables into the environment(vector). The mathematical expression for each set of new vectors is as follows:

$$\overrightarrow{C_i^n} = (r_i^n, C_i^n(0), MS_i, EA_i, AN_i, LS_i, C_i^n(1), C_i^n(2), C_i^n(3), \cdots, C_i^n(k))$$

where $MS_i, EA_i, AN_i, LS_i$ is the marital status, educational attainment, ancestry and the language spoken of $i$ county.

Finally, we followed the update rules explained in section 3.2.3 to complete the iteration of cellular automata.

## 4.4 Results & analysis

We use two methods of significance analysis, analysis of variance and correlation analysis, to examine the significant impact of census socio-economic factors on the use of opioids. The data passes the test for variance homogeneity is analyzed by ANOVA, while the other are analyzed by correlation analysis.

It is fun that when we try to correlate the factors that passed the variance significance test, the correlation is at a lower level. After careful analysis, we find that the data samples selected by the two methods are inconsistent; we extract the first 25% and the last 25% number of data sets as samples in the analysis of variance, while we use all the samples in the correlation analysis. Therefore, it can be explained that the results obtained by the two methods are different, because these attributes are localized significantly, and when the sample size increases, the significance is weakened.

We analyze the impact of significant factors on the trend of opioid usage in the following four areas:

- **Marital status**

There is a positive correlation between the use of opioids in men and women aged 15 and over who have never been married.

The reason for this situation may be that men and women who are currently unmarried are the main opioid users. Because they are still pursuing personal entertainment and enjoyment who lack of family responsibility and social responsibility. There is little motivation to resist opioids, so it is easier to abuse drugs.

- **Educational attainment**

In terms of educational attainment, we find that the results are contrary to our perception. Generally, we believe that people with higher degree are more aware of the dangers of the abuse of opioids, so they naturally contradict it.

On the contrary, the results show that people with a bachelor’s degree or higher are positively correlated with the usage of opioids. The reason for this may be that the higher the responsibility of those with high educational background, the greater the pressure of study and work. This leads them to seek exciting opioids to stay focused or to relieve stress.

- **Ancestry**

People whose ancestors are Arab, Greek, Irish, Italian, Russian, and Ukrainian have significant effects on the use of opioids, and are positively correlated.

We suspect that it is due to the mixed population. The local population has a complex source and many unstable factors, which is an excellent environment for illegal trafficking of opioids.

- **Language spoken at home**

In the United States, if English is not very good, you can’t communicate well with others. Hence, it's not easy to make friends. Therefore, in the process of dealing with people, it is easy to produce inferiority, and it is often easy to be criticized and addicted to drugs.

- **Spread**

The result of synthetic opioid spread in the future is solved by the CA model and the modified CA model. The spread result of synthetic opioid predicted by original and modified model in 2020 are shown in Figure 11.

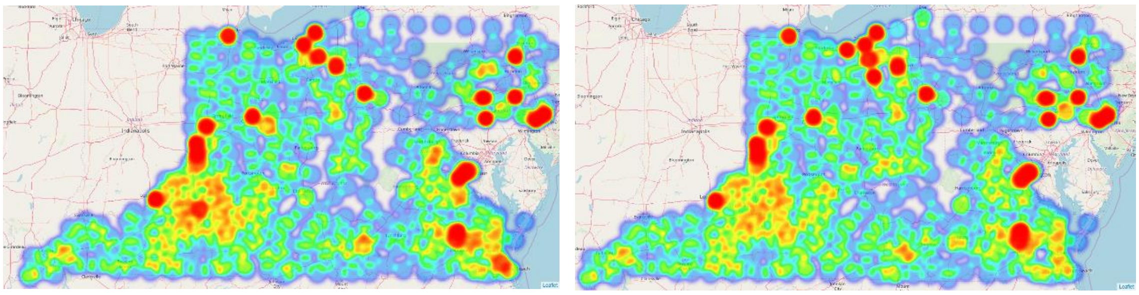


**Figure 11** (a) 2020 synthetic opioid original

**Figure 11** (b) 2020 synthetic opioid modified

By comparing Figure 11(a) and Figure 11(b), the aggregation point is still mainly in the areas with developed traffic, and there is a tendency to spread around. There are more aggregation points appear in the Ohio. It means that we have observed more propagation rules. From the Figure 11(b), we notice the Synthetic opioids usage are more concentrated in Virginia.

The spread result of heroin predicted by two model in 2020 are shown in Figure 12:

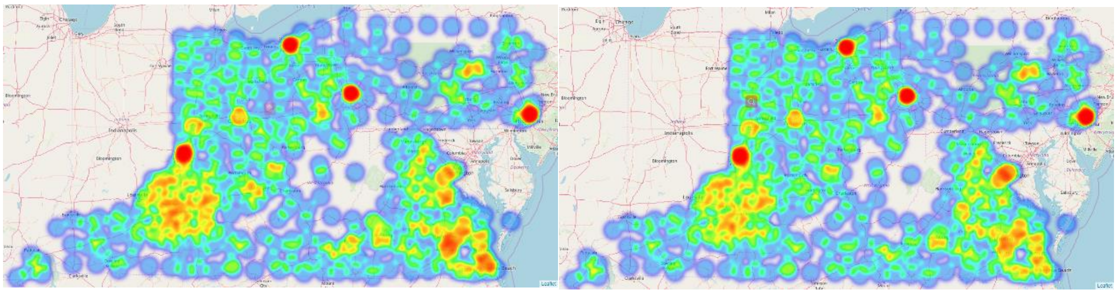


**Figure 12** (a) 2020 Heroin opioid original **Figure 12** (b) 2020 Heroin opioid modified

By comparing Figure 12(a) and Figure 12(b), we find that heroin spread to the southwest in Kentucky predicted by modified model is slower than the original one. Two model have the same prediction of the Heroin spread in the Ohio and Pennsylvania.

- **Characteristics**

By calculating relevant formulas, we can obtain the concentration index of the synthetic opioid predicted by the modified model. The result is shown in Figure 13.

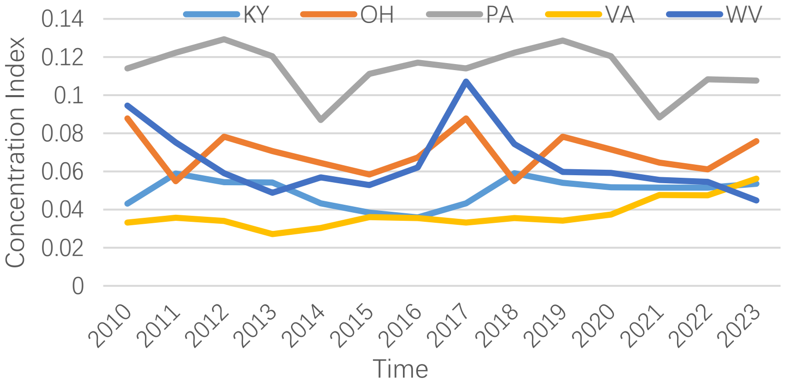


**Figure 13:** CI of the synthetic opioid modified

From the Figure 13, we find the distribution of synthetic opioid predicted by the modified model in Virginia is the more concentrated than the original one. They have the same prediction of the Concentration Index in Pennsylvania and Ohio.

## 4.5 Model Evaluation

From the modeling process of the cellular automata, we can know that the key to improving the simulation and prediction capabilities of the model is the more effective updating rules. We think simulation results of CA model will be better after we introduce new variables into updating rule.

To evaluate whether our model performs better than the unmodified one. We used RMSE as described in section 3.4 to calculate the performance of the modified model. In section 3.4, we know that when $k$ is 12 and $m$ is 2, the average RMSE of the original model is the minimum. Interestingly, $k$=12 and $m$=2 are not the optimal combination of the modified model.

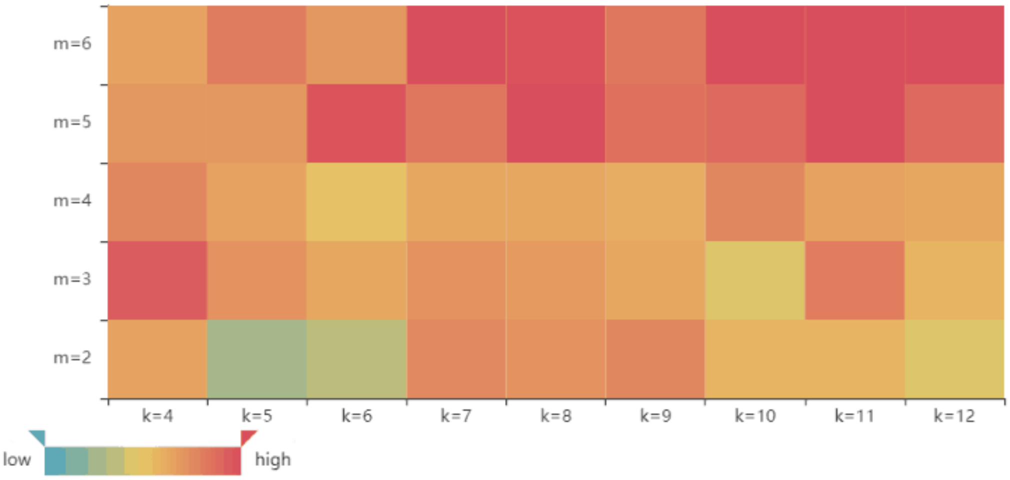


**Figure 14:** Heat map of modified model

According to the Figure 14, it is obvious that when $k$ is 5 and $m$ is 2, the modified model has the smallest average RMSE. The smallest average RMSE is 98.70.

## 4.6 Strategy

### 4.6.1 Principles of strategy

- We split the strategy into separate operations, using the control variates approach, and consider only the impact of one action on opioid use;
- The effect of the behavior will be directly reflected in the change in the value of the relevant variable;
- When the variables change, we select the combination of parameters that have been determined in Section 4.3 to simulate the use of opioid from 2010 to 2017.

Based on the above principles, we propose the following 2 actions and quantitatively interpret their promising effects. The simulation data is the simulation result of total reported opioid incidents in 2017. The real data is the total reported opioid incidents in 2017.

### 4.6.2 Action

✧ **Action One: Give couples a discount on tax and mortgage rates to encourage people to marry at legal age.**

This action will influence the marital status(MS). Assume the change rate $r$ is 5%, 10%, 15%, 20%. According to the Table 1, we notice that when the $r$ is 5%, the simulation result exceed the real data in 2017. We think that might be the cause of the error. When $r$ is 10%, 15%, 20%, it is obvious that the simulation data is lower than real data. It means the action one is effective.

**Table 5:** Comparative Results

| Data | $r$=5% | $r$=10% | $r$=15% | $r$=20% |
|---|---|---|---|---|
| Actual Data | 257496 | 257496 | 257496 | 257496 |
| Simulation Data | 263529 | 238872 | 243077 | 231073 |

✧ **Action Two: Open a low-cost English language training institution to improve the English proficiency of non-native English speakers.**

This action will influence the percent of speak English less than "very well"(Lo). Assume the change rate $r$ is -5%, -10%, -15%, -20%. We note from our simulation that when r is -10%, the simulation data exceed the actual data in 2017. When r is at other values, the simulation data is lower than the actual data. It means the action two is effective.

**Table 6:** Comparative Results

| Data | $r$=-5% | $r$=-10% | $r$=15% | $r$=20% |
|---|---|---|---|---|
| Actual Data | 257496 | 257496 | 257496 | 257496 |
| Simulation Data | 239705 | 268312 | 237580 | 225873 |

In all, we can conclude that take action one and action two as our strategy can effectively reduce the number of opioid cases.

# 5 Strength and Weakness

## 5.1 Strength

- Based on the idea of cellular automata, we skillfully combine clustering and KNN ideas to innovate the model of this paper.
- Our model fully exploits the information of historical data and simulates the changes of opioids in each county from the perspective of historical development.
- Although the model based on clustering and KNN is sensitive to $k$ values, we try multiple sets of different k values and choose the smallest k value of RMSE as the model parameter.
- Our model is extensible and can add new influence variables in the process of building an 'environment' around itself.
- We set goals in strict accordance with the optimization theory.

## 5.2 Weakness

- Our model predicts that the error of the results in the past one or two years is small, but the prediction of long-term results will be very large.
- The model results may be over-fitted, because the number of real data used to validate the model is too small.
- The model only considers the counties discussed in the original data set, but does not consider the spread of the county to the county where there is no abuse of opioids.

# 6 Conclusion

In this paper, we first collect latitude and longitude data for each county in five states. After data preprocessing, we analyze the spread and characteristics of the reported synthetic opioid and heroin incidents in and between the five states and their counties over time combined the heat maps and concentration index (CI). Next, based on the idea of cellular automata, we built a cellular automata model based on the idea of clustering and KNN, and evolved the number of cases, and analyzed the error of the model. According to the results of the model simulation, we then analyze where specific opioid use might have started in each of the five states. And then, we use ANOVA and Pearson Correlation Analysis to find some socio-economic factors that have a significant impact on the model, and to modify the model. Moreover, we propose some possible strategies to deal with the opioid crisis, which are verified by our model. Finally, the advantages and disadvantages of the model are discussed.

## 7 References

[1] Cicero, T., Inciardi, J. and Muñoz, A. (2005). Trends in Abuse of OxyContin® and Other Opioid Analgesics in the United States: 2002-2004. The Journal of Pain, 6(10), pp.662-672.

[2] Volkow, N., Jones, E., Einstein, E. and Wargo, E. (2018). Prevention and Treatment of Opioid Misuse and Addiction. JAMA Psychiatry.

[3] Opioid. (2019). Retrieved from https://en.wikipedia.org/wiki/Opioid

[4] Heroin. (2019). Retrieved from https://en.wikipedia.org/wiki/Heroin

[5] Rhoades, S. A. (1993). The herfindahl-hirschman index. Fed. Res. Bull., 79, 188.

[6] United States Cities Database. Retrieved from https://simplemaps.com/data/us-cities?tdsourcetag=s_pcqq_aiomsg

[7] Haversine formula. (2019). Retrieved from https://en.wikipedia.org/wiki/Haversine_formula

[8] Yang, J., Su, J., Chen, F., Xie, P., & Ge, Q. (2016). A Local Land Use Competition Cellular Automata Model and Its Application. ISPRS International Journal Of Geo-Information, 5(7), 106. doi: 10.3390/ijgi5070106

[9] Pfeifer, B. (2008). A Cellular Automaton Framework for Infectious Disease Spread Simulation. The Open Medical Informatics Journal, 2(1), 70-81. doi: 10.2174/1874431100802010070

[10] FARAWAY, J. (2002). Practical Regression and Anova using R. [Bath]: [University of Bath].

[11] Knapp, T. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85(2), 410-416. doi: 10.1037//0033-2909.85.2.410

# Appendix

## I. RMSE of different combinations of k and m

| | k4m2 | k4m3 | k4m4 | k4m5 | k4m6 | k5m2 | k5m3 | k5m4 | k5m5 | k5m6 | k6m2 | k6m3 | k6m4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2011 | 0.702974035 | 0.660225292 | 0.651338947 | 0.644141611 | 0.631348872 | 0.689582827 | 0.654907955 | 0.640512615 | 0.631348872 | 0.618291503 | 0.686194193 | 0.64954709 | 0.618291503 |
| 2012 | 6.382522756 | 6.227991553 | 6.993836781 | 7.969931839 | 9.086992163 | 4.858846736 | 7.678869381 | 7.768503739 | 8.055460439 | 9.061818257 | 6.812678432 | 6.943830587 | 7.560941222 |
| 2013 | 34.5552495 | 43.26408309 | 51.06440469 | 66.29665969 | 73.79199646 | 35.09461072 | 103.3068956 | 58.58633778 | 67.97781758 | 72.9797526 | 95.67006971 | 42.10246686 | 54.46951201 |
| 2014 | 118.5778098 | 133.7225235 | 173.3128395 | 205.9417903 | 231.547958 | 106.5340157 | 138.4434617 | 189.2682819 | 211.0507622 | 232.5925888 | 118.6640475 | 148.4046161 | 177.7839959 |
| 2015 | 179.7368771 | 246.459158 | 252.3333303 | 269.5203847 | 298.9185403 | 164.9199778 | 252.9398843 | 276.0128033 | 313.1720187 | 348.1102195 | 189.2682695 | 255.2801423 | 266.0224021 |
| 2016 | 451.2061458 | 573.3207221 | 440.7932323 | 473.3943688 | 336.5407286 | 191.3256417 | 565.5581953 | 547.2652012 | 499.3182182 | 396.6530281 | 221.4727564 | 438.9276082 | 317.7051638 |
| 2017 | 372.7998207 | 696.3578257 | 575.1554832 | 550.9506512 | 624.6561366 | 447.7870139 | 504.9906321 | 507.5659705 | 468.3214235 | 744.0179523 | 251.1995325 | 489.1281931 | 422.8964478 |
| | k4m2 | k4m3 | k4m4 | k4m5 | k4m6 | k5m2 | k5m3 | k5m4 | k5m5 | k5m6 | k6m2 | k6m3 | k6m4 |
| mean | 166.2801999 | 242.8589327 | 214.3292094 | 224.959704 | 225.0248144 | 135.8870985 | 224.7961209 | 226.7296587 | 224.0752928 | 257.719093 | 126.253364 | 197.3480577 | 178.1509649 |
| std | 181.0268378 | 283.2203141 | 223.5731777 | 220.9499227 | 222.671942 | 156.8786141 | 229.1343946 | 228.3187527 | 209.9527292 | 266.6617715 | 99.72903551 | 203.4261458 | 164.82979 |

| | k6m5 | k6m6 | k7m2 | k7m3 | k7m4 | k7m5 | k7m6 | k8m2 | k8m3 | k8m4 | k8m5 | k8m6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2011 | 0.610704817 | 0.618291503 | 0.69295489 | 0.67936622 | 0.675926369 | 0.654907955 | 0.647750276 | 0.699650262 | 0.668993608 | 0.656685184 | 0.64954709 | 0.651338947 |
| 2012 | 8.630325456 | 9.028703258 | 4.464049611 | 5.923561267 | 7.193463752 | 8.454182011 | 9.475769615 | 5.501218344 | 6.570223144 | 7.409909575 | 8.553819741 | 9.530468867 |
| 2013 | 71.43347436 | 92.56316435 | 27.78748859 | 42.73791218 | 54.88778551 | 78.72786624 | 86.23140656 | 45.83951431 | 90.11657821 | 92.82732045 | 107.6172966 | 116.9397171 |
| 2014 | 194.4008247 | 206.4638508 | 81.23659653 | 160.0603252 | 190.9938429 | 230.9795297 | 251.8435245 | 144.6112464 | 219.3248115 | 233.5453915 | 247.5924823 | 260.5668601 |
| 2015 | 444.4392213 | 335.0241163 | 149.0443763 | 231.4022724 | 261.9961342 | 326.3630838 | 359.1618643 | 192.6916749 | 299.5087353 | 312.2190037 | 360.4894788 | 367.3190213 |
| 2016 | 494.8957744 | 423.0661336 | 404.1983404 | 420.4118638 | 327.4526888 | 395.2867066 | 431.6021948 | 552.3179129 | 490.9140089 | 389.7772809 | 677.3048365 | 625.2052567 |
| 2017 | 616.9209449 | 545.2829038 | 909.2514493 | 376.2977204 | 549.1471879 | 533.0649449 | 625.5960272 | 465.5550964 | 616.8661153 | 522.1177849 | 604.8729952 | 604.6975433 |
| | k6m5 | k6m6 | k7m2 | k7m3 | k7m4 | k7m5 | k7m6 | k8m2 | k8m3 | k8m4 | k8m5 | k8m6 |
| mean | 261.6187528 | 230.2924519 | 225.2393222 | 176.7875745 | 198.9067185 | 224.7901744 | 252.079791 | 201.0309019 | 246.2813523 | 222.6504823 | 286.7257795 | 283.5586009 |
| std | 253.9350887 | 211.5791515 | 332.8141107 | 173.4923546 | 200.0109218 | 205.2761728 | 235.5100024 | 223.2626137 | 239.0649512 | 199.2557212 | 274.4817078 | 261.4978886 |

| | k9m2 | k9m3 | k9m4 | k9m5 | k9m6 | k10m2 | k10m3 | k10m4 | k10m5 | k10m6 | k11m2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2011 | 0.684493585 | 0.686194193 | 0.689582827 | 0.656685184 | 0.656685184 | 0.702974035 | 0.682788742 | 0.674199862 | 0.658457617 | 0.656685184 | 0.69295489 |
| 2012 | 7.120835183 | 14.1057445 | 14.0904504 | 14.20979996 | 14.54604457 | 8.450734786 | 14.08879599 | 12.28165953 | 14.34319604 | 14.52487618 | 8.444664251 |
| 2013 | 59.07880636 | 84.64556571 | 99.20734802 | 111.3241032 | 122.6053612 | 61.89817942 | 75.34132123 | 84.18581679 | 98.70752162 | 137.9641266 | 85.14030652 |
| 2014 | 127.0279598 | 166.5318102 | 182.0729381 | 208.7606899 | 251.9515669 | 103.818786 | 159.3326116 | 181.8095843 | 211.9986421 | 243.8577292 | 128.3012735 |
| 2015 | 174.7381724 | 234.855488 | 232.0894852 | 327.9651895 | 389.6765379 | 213.2507143 | 231.2633289 | 300.8321249 | 307.1571809 | 356.7766207 | 213.927171 |
| 2016 | 201.8843457 | 282.6825627 | 263.3484844 | 534.9816763 | 526.850235 | 179.1816154 | 239.8553128 | 315.1385569 | 660.3541396 | 648.3914074 | 575.7651077 |
| 2017 | 757.2435429 | 558.9201627 | 338.6156102 | 622.0920565 | 572.8990827 | 450.9889935 | 500.2835536 | 511.3414439 | 626.2145011 | 691.0289548 | 498.8250951 |
| | k9m2 | k9m3 | k9m4 | k9m5 | k9m6 | k10m2 | k10m3 | k10m4 | k10m5 | k10m6 | k11m2 |
| mean | 189.6825937 | 191.7753611 | 161.4448427 | 259.9986001 | 268.4550734 | 145.4702853 | 174.4068161 | 200.8947694 | 274.2048056 | 299.0286286 | 215.870939 |
| std | 262.2116446 | 193.6968913 | 128.2049746 | 246.2202988 | 235.1875612 | 156.803166 | 173.0293453 | 186.6657714 | 274.2056883 | 282.2549797 | 232.2630743 |

| | k11m3 | k11m4 | k11m5 | k11m6 | k12m2 | k12m3 | k12m4 | k12m5 | k12m6 |
|---|---|---|---|---|---|---|---|---|---|
| 2011 | 0.668993608 | 0.668993608 | 0.653125889 | 0.644141611 | 0.702974035 | 0.684493585 | 0.672468923 | 0.667249163 | 0.658457617 |
| 2012 | 7.456790838 | 8.110829044 | 9.927710741 | 10.80781466 | 3.19564097 | 4.891837071 | 6.155900007 | 6.82344785 | 8.610585993 |
| 2013 | 68.2665829 | 79.52509097 | 129.7653522 | 145.0459054 | 31.67462788 | 60.44919268 | 74.86483547 | 83.11639829 | 107.6360851 |
| 2014 | 196.4820792 | 184.6359614 | 207.0583006 | 240.8498163 | 83.36576525 | 169.4449323 | 200.8082445 | 226.0490437 | 255.054727 |
| 2015 | 300.2621271 | 259.1984282 | 330.2023749 | 385.1362336 | 137.3245437 | 250.5763008 | 331.2169656 | 354.8736011 | 355.0737538 |
| 2016 | 509.1044035 | 449.8267136 | 436.4830278 | 811.8192583 | 208.9219191 | 458.2980677 | 480.8881546 | 450.7891992 | 665.1452319 |
| 2017 | 506.6644407 | 505.8198681 | 831.256539 | 760.9899331 | 264.3423647 | 490.0159053 | 535.5806926 | 591.6666851 | 681.5557088 |
| | k11m3 | k11m4 | k11m5 | k11m6 | k12m2 | k12m3 | k12m4 | k12m5 | k12m6 |
| mean | 226.9864883 | 212.5408407 | 277.906633 | 336.4704433 | 104.2182622 | 204.9086756 | 232.8838945 | 244.8550892 | 296.2477929 |
| std | 219.3608208 | 203.9800991 | 291.4689724 | 334.9536692 | 103.5350184 | 204.702353 | 221.4074139 | 230.0692101 | 287.340526 |