For office use only	Team Control Number	For office use only
T1 _____	1918889	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	C	F4 _____

2019
MCM/ICM
Summary Sheet

An Opioid-Combating Mechanism Based On Data Insight

Drug addiction and opioid abuse are now ravaging America. In 2017, more people died in opioid crisis than in either car accidents or gun violence. Moreover, This multi-faceted epidemic imposes destructive effects not only on health of several generations but also on other aspects of the society, such as economic development. In this paper, we construct a **CA-SIR** model to imitate the spreading process of the reported synthetic opioid and heroin incidents, and associate them with socio-economic data provided. We finally succeed in determining an optimal strategy for fighting against opioid crisis.

First, we operate on the data. We filter data based on the integrity and redundancy of the information, delete data with information less than the threshold, and combine similar attributes using **PCA**. For socio-economic data, we do data imputation to fill in missing data based on **K-means clustering**. Then we normalize all the data so that they are comparable in the following analysis.

As for the part I, we use **SIR** model and categorize people into three healthy conditions. Based on the given data of counties in Ohio, Kentucky, West Virginia, Virginia, and Pennsylvania states, we then construct a **Cellular Automaton** model to imitate the spreading process of the reported synthetic opioid and heroin incidents. In this model we set up three methods for every individual to travel randomly within the five states. Next, we apply **Genetic Algorithm** into the refining procedure to get the accurate parameters. By doing this, we find the most possible locations and time nodes that specific opioid use might have started.

In the second part, as a factor extraction, we choose a selection method called **post-LASSO** to acquire the most significant socio-economic factors. In this case, five factors are selected. We then adopt a linear model to take the factors into consideration and revise our **Cellular Automation** model. Next, we rerun the **Genetic Algorithm** to refine our parameters.

In the third part, combining outputs from models above and relevant researches, we set up a comprehensive strategy to combat the crisis including views from different perspectives. Moreover, we test the effectiveness of the strategy using our model. After that, we point out the constraints conditions of different parameters. We further discuss how to devise the strategy with our methods and put up insightful suggestions for doctors, addiction treatment providers, law enforcement, insurers, the medical industry. And we send the insights to the Chief Administrator of DEA/NFLIS Database.

Finally, we examine the sensitivity of our model, changing the amount of our parameters in CA imitation progress. The result shows that our model is robust.

Keywords: Opioid Crisis, SIR, K-Means Clustering, CA, Genetic Algorithm

Contents

1 Introduction **1**

1.1 Problem Background 1

1.2 Overview of our work 1

1.3 Notations 2

1.4 Assumptions 3

2 Data analysis **3**

2.1 Data imputation 3

2.2 Data Normalization 4

3 Model Construction **4**

3.1 SIR Model 4

3.2 Imitation with CA 5

3.3 Genetic Algorithm 7

3.4 Results 8

3.5 Finding the Origins 9

3.6 Specific concerns and relative prediction 10

4 Extension of Model **11**

4.1 Selecting the socio-economic parameters 11

4.2 Refitting 14

4.3 Results Analysis 14

5 Strategy Making **14**

5.1 Strategy Elaboration 14

5.2 Testing on effectiveness 15

6 Sensitivity Analysis **16**

6.1 Strengths 17

6.2 Weaknesses 17

7 Memo to the Chief Administrator of DEA/NFLIS Database **19**

References **21**

Appendix **22**

1 Introduction

1.1 Problem Background

Dated back to the mid- to late-19th century, the first national opioid crisis occurred and rose dramatically sine then, which was fueled by physicians' unrestrained opioid prescriptions (including morphine, laudanum, paregoric, codeine, and heroin) for pain or other ailments, and by liberal use of opioid-based treatments for injuries and diseases.

Now the situation is becoming even worse. In 2017, More than 175 lives lost every day due to opioid misuse[1]. If a terrorist organization was killing 175 Americans a day on American soil, what would people do to stop them? They would do anything and everything. Similarly, Things must be done to combat opioid crisis as it won't get better by itself. Specifically speaking, comprehensive study on origins, spreading process and prediction of the crisis are in urgent demand. This nation-wide crisis has extended across socio-economic statuses so relative factors are of significant importance to explain current opioid use. We are tasked with creating model to look deep into this crisis. The strategy proposed within this paper will offer an insight to counter the opioid crisis.

1.2 Overview of our work

First,we clarify some key points for solving the problem :

- The volume of data is large and of different types. How to do the normalization of the data.
- Among the massive data, there are many missing data in the files. How to fill in the data.
- How to classify the people into subgroups and define their different moving paths.
- How to extract socio-economic factors that effect the opioid use most from the data set.

On the basis of above discussion, to built up matching model and develop effective strategy finally, we will break our work into four following steps :

- Using the data provided by NFLIS, we build a mathematical model to to imitate the spreading process of the reported synthetic opioid and heroin incidents within five states and extract significant characteristics.
- we use post-LASSO model to select the most significant socio-economic factors.With the factors we revise our cellular automation model and use that to refine our parameters.

- Combining the outputs above,we set up a comprehensive strategy to cope the crisis.After that we test the effectiveness and point out the constraint conditions.
- Futhermore,we put up insightful suggestions for doctors, addiction treatment providers, law enforcement, insurers, the medical industry. And we send the insights to the Chief Administrator of DEA/NFLIS Database.

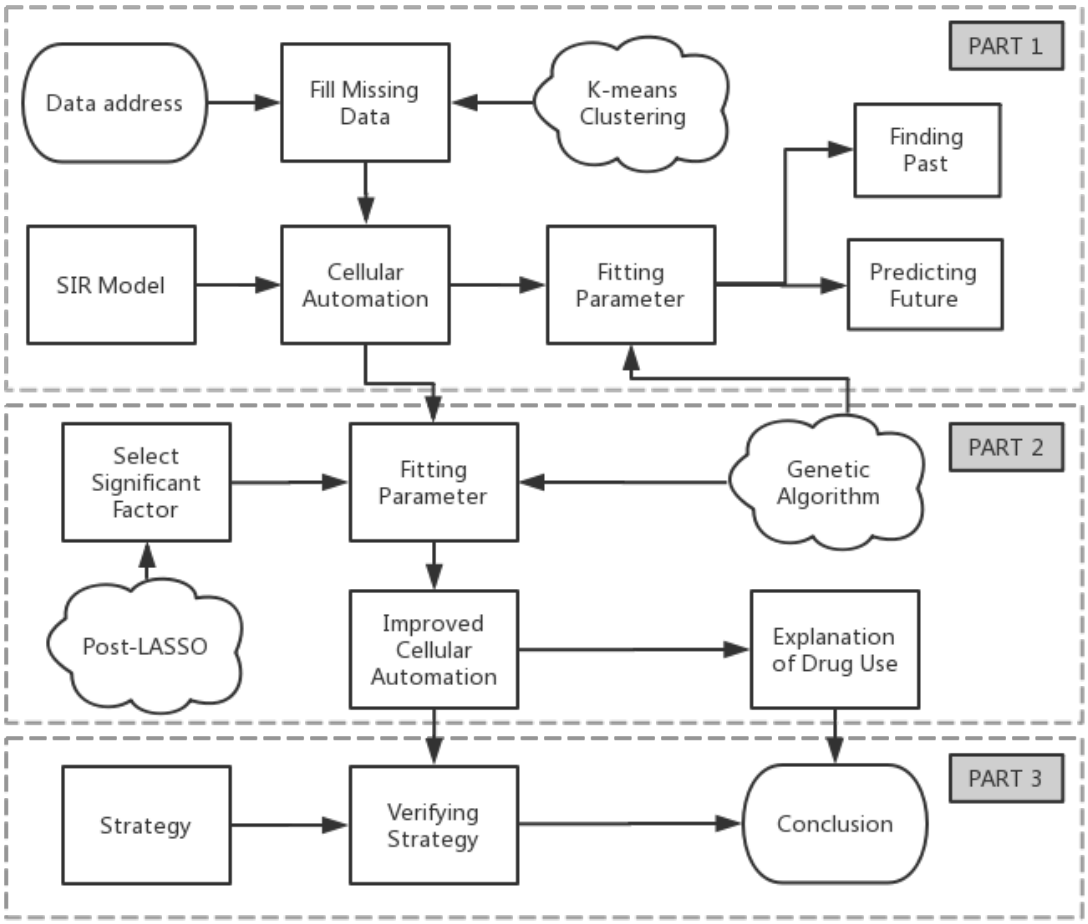

Figure 1: The process of our work

1.3 Notations

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

Table 1: Notations

Variables	Definition
P_{if}, P_{rc}, P_{ri}	the probability of infection, recovery and re-infection
P_{wk}, P_{cm}, P_{tv}	the probability of walking, commuting and traveling
F_1, F_2, F_3, F_4	the socio-economic variable associate to opioid abuse

1.4 Assumptions

1. In the initial situation, the infected people in each county are distributed randomly.

2. The spreading mechanism of opioid is similar to that of infected diseases, assuming that people can get infected by daily contacts.

3. Each entire state can be divided into 1000*1000 cells and people in same cells will act similarly.

2 Data analysis

Since this is a big data issue, our work begins with organizing the data. The data provided by the question has some problems ranging from data missing, unnatural variables, trivial variables to duplication of names. We need to deal with these issues as rectification on given data is a prerequisite for model fitting and analysis.

2.1 Data imputation

As for data in the the attached Excel document MCM_NFLIS_Data, we need not to do further imputation, since their integrity are satisfying enough. But for the others, they are incomplete to different extent, and some of them are under same name but with different interpretations.

To first clean these data, we delete the incomplete series and pair those with same interpretations to improve their integrity. But there are still a lot of similar data and some data missing.

For data with missing rate over 40%, almost any current prevailing method would fall victim to under- or over-randomization. As a result, we omit this kind of data for simplicity.

For variables with missing rate under 40%, we address K-Means clustering[4] to put them into groups, then use the average value of each group to replace the missing value.

K-Means clustering is an algorithm to classify or to group objects based on attributes, into K number of groups. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster center. The most common implementation of the K-Means algorithm is the Lloyd's algorithm using iterative refinement heuristics.

Hereby we let K=20 and successfully cluster our data into 20 groups. Then we replace the missing data with the average value of their groups. By doing this, our missing data is refilled.

Moreover, to deal with the high-dimension issue, we apply PCA method in each group to decrease the number of variables to 20.

2.2 Data Normalization

In the multi-index evaluation system, due to different nature of each evaluation index, it usually has different dimensions and orders of magnitude. When levels between the indicators vary dramatically, if the analysis is performed directly with the original indicator values, the role of the higher-value indicators in the comprehensive analysis will be highlighted, therefore the effects of the lower-level indicators will be relatively weakened.

Therefore, in order to ensure the reliability of the results, the original indicator data needs to be standardized. For socio-economic data, we use Min-max normalization, to do linear transformation of the raw data and map data set

$$x = \{x_1, x_2, ..., x_n\}$$

into [0,1] interval. Let x' be the normalized value, the formula is :

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

After that, all scio-economic data is normalized. As for drug data, we don't need to perform normalization because we need their absolute value.

3 Model Construction

Basically, we use concept of infected diseases to fit the opioid-addiction cases. Initially, we use developed SIR model to depict human behaviour. After that we use Cellular Automata Model to exert simulation calculations. Finally we use genetic algorithm to manage the fitting for optimal parameters.

3.1 SIR Model

SIR model is a universally used epidemic model, which had been established by Kermack and McKendrick in 1927 and further developed by exploring Threshold Theory from epidemic dynamics[2]. In the model, S,I,R represent the susceptible, the infected and the removed respectively. The infected can infect the susceptible with possibilities. The removal refers to the group of people who are isolated or obtain immunity from disease. We apply SIR model to this issue because of certain similarities between addiction to opioid and infection.People in the same area can exchange diseases or opioid using habits through interaction.

However, discrepancy still exists between model imitating addicts' behaviour and traditional SIR model. For example, traditional SIR model ignores interactions between different areas. Also probability of relapse after detoxification treatment is much higher than that of initial addiction. To solve the potential problems, we revise the structure of the model in three perspectives:

1.Introduce mobility of people to make remote contacts possible. We will further explain this part later on when setting up CA model.

2.Bring birth and death events to the model to produce a more realistic outcome.

We suppose that people could be addicted, be cured, and relapse, which represent the infection, recovery, and re-infection in the SIR model. Using the model, we can solve the problem more conveniently.

3.2 Imitation with CA

In order to simulate the spread and develop of the drug, we decide to use Cellular Automaton model.

Cellular automaton is a discrete model studied in computer science, mathematics, physics, complexity science, theoretical biology and micro-structure modeling[8]. A cellular automaton consists of a regular grid of cells, each in one of a finite number of its information. The cells may change their information and influence others with time going by.

We download the geography data of all the states from Arcgis[3]. Then we create a map with 1 million (1000*1000) cells, and project the real map of a state to it. We use PNPoly algorithm(W. Randolph Franklin, 2006) to determine which county every cell belongs to, and there may be some cells that not belong to any county in this state, which will be ignored in the following calculating.

Taking Ohio as an example, the cell map of it is as follows.Areas with different depth of colors refer to different counties. Note that we have not uploaded the data about the infection to the map, so the color depth in certain county does not change.

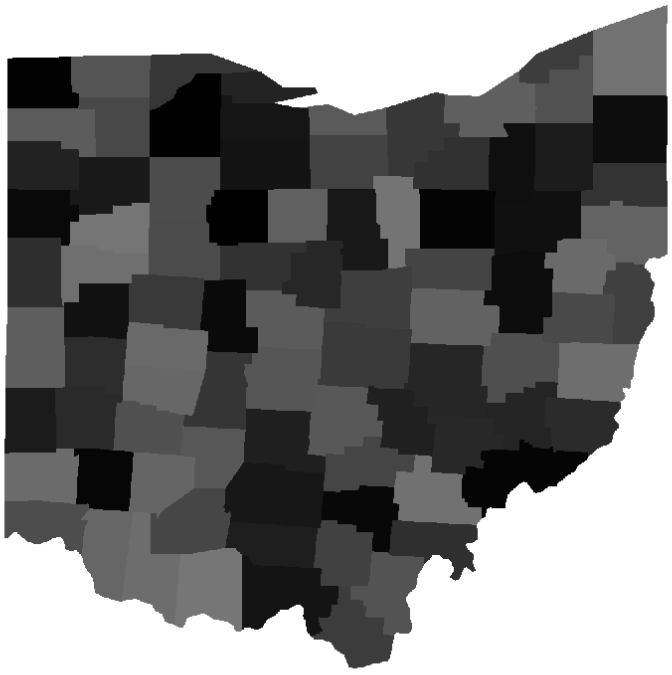

Figure 2: The cell map of Ohio

Using this method, we can get a map with quantities of cells for every certain state. Actually, there may be about 10 persons in a cell. But here, we consider them as one individual. A cell may be healthy or infected.

In the real world, a person may move, and contact with others, and he may get to be infected when an infected person influence him. So, in order to imitate the reality, a cell in the map can move and influence other cells that it may contact with.

We build the imitation procedure on a weekly basis. At the beginning of each week, each cell starts its movement. Hereby, coming to the end of the week we define each cell as infected or healthy.

Do this operation repeatedly, Then we can get a forecast cell map after several weeks.

Firstly, we define the motion rule. Each cell may move with one of the ways in follow with certain probability (Of course they can stay):

1.Walk: Exchange position with adjacent cells, which imitates normal contacts in walking distance. Its probability P_{wk} is of medium level.

2.Commute: Exchange position with any cells within 10-grids distance, which imitates moving by vehicles in a county. Its probability P_{cm} is smaller than that of exchange and decreases with the distance.

3.Travel: Exchange position with a random cell in the whole map, which imitates travelling by inter-county transportation. It happens at the lowest probability P_{tv}.

Then, we define the contagion rule as follows.

1.Infection: If the cell is healthy and it have never been infected, and there are some adjacent cells that are infected, it may be infected with probability P_{if}.

2.Recovery: If the cell is infected, it may recovers with probability P_{rc}.

3.Re-infection: If the cell is healthy but it have been infected, and there are some adjacent cells that are infected, it may be infected with probability P_{ri}.

The probability of the walking, commuting, and travelling is set up according to the actual situation, and we need to get proper probability of the infection, recovery, and re-infection through our algorithm.

Table 2: Probability of motion rule

Parameters	Values
P_{wk}	0.0200
P_{cm}	0.0040
P_{tv}	0.0010

3.3 Genetic Algorithm

In order to find the optimal P_{if}, P_{rc}, and P_{ri}, we choose genetic algorithm to calculate. The genetic algorithm is a method for solving both constrained and unconstrained optimization problems. It is based on natural selection and biological evolution. The genetic algorithm repeatedly modifies a population of individual solutions.

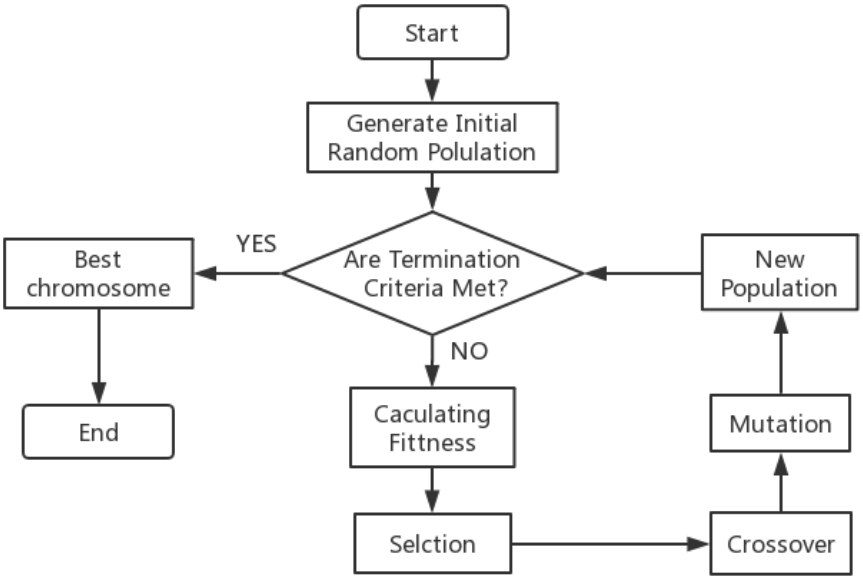

Figure 3: the Interpretation of GA

Here, we consider the vector of these three probability, i.e. [P_{if}, P_{rc},P_{ri}]', as a individual. We set the vector to the CA model, based on the situation of a specific time node(i.e. the number of infected people of every county in a state). Then We imitate the situation after one year or several years. After that, we compare the forecast situation with the real situation. The more similar the forecast is to the real situation, the more fit-value the individual get. The formula of the fit-value is:

$$FV = \frac{1}{distance(NI_{forecast}, NI_{real})}$$

where $NI_{forecast}$ is a vector. This vector saves the number of the infected people of every county we forecast. NI_{real} is also a vector, which saves the real number of the infected people of every county. Distance indicates the vector-distance between them.

If the fit-value of a individual is large, the situation we forecast is similar to the real situation. After the calculation of the genetic algorithm, we can get a strongest individual, who can tell us the optimal value of P_{if}, P_{rc}, and P_{ri}.

The result of GA is as follows.

Table 3: Result of GA

Variables	Fitting outputs
P_{if}	0.0011
P_{rc}	0.0052
P_{ri}	0.0406
Generation	109
Minimal MSE	5.32

Then we use the three value to operate the CA model, and get the forecast infected situation of all the states.

3.4 Results

Using the models above, we get the forecast situation of all the 5 states, then we compare them with the real forecast, and find that the results we forecast is similar to the reality.

We visualize the forecast and the reality by coloring the cell maps. Here are some examples. More details can be found in the appendix.

The red cells represent the infected cells while the gray cells represent the healthy cells.

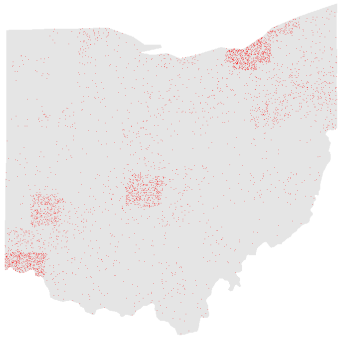

Figure 4: Reality,Ohio,2014

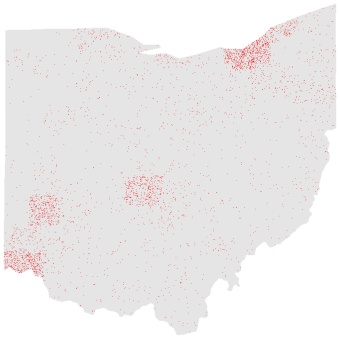

Figure 5: Forecast,Ohio,2014

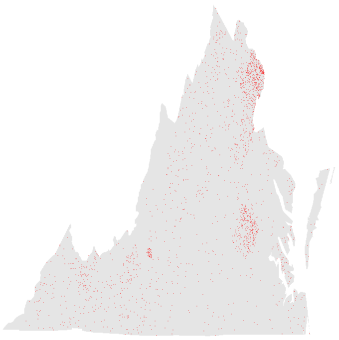

Figure 6: Reality,Virginia,2016

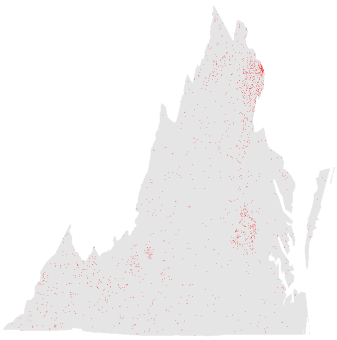

Figure 7: Forecast,Virginia,2016

These maps confirm that our model are correct and available.

3.5 Finding the Origins

Now we can use our models to forecast the future. However, in order to find the origin of the opioid abusing, we need to estimate the past first.

The SIR model classifies people to three types: the susceptible people (healthy), the infected people and the removed people (also healthy). Healthy people may become infected, also infected people may become healthy.

Consider the model the other way round. If time turns back, the healthy people might be infected in the past, with probability same as the probability of recovery in the actual time flow. Also, the infected people might be healthy at that time.

So, we can re-work our model after adjusting P_{if}, P_{rc}, and P_{ri}. And we can get the estimated situation 10 years before 2010. The results are as follows.

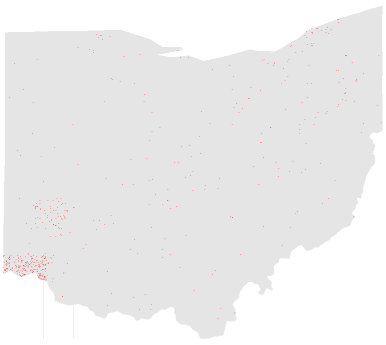

Figure 8: Estimation,Ohio,2000

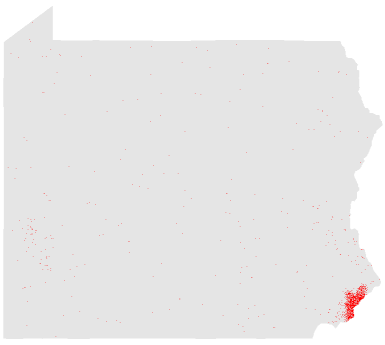

Figure 9: Estimation,Pennsylvania,2000

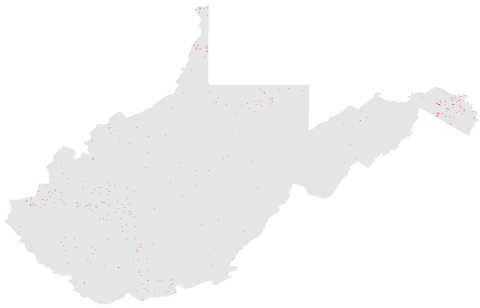

Figure 10: Estimation,West Virginia,2000

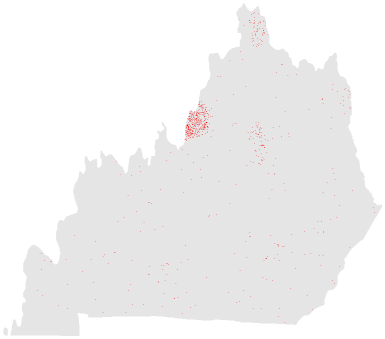

Figure 11: Estimation,Kentucky,2000

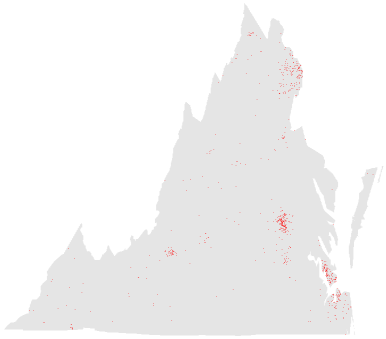

Figure 12: Estimation,Virginia,2000

From these figures, we can easily find the origin in every state.

In Ohio, the origins are Hamilton and Montgomery.

In Pennsylvania, the origin is Philadelphia.

In Kentucky, the origins are Jefferson and Woodford.

In Virginia, the origins are Fairfax and Richmond.

In West Virginia, the origins are Kanawha and Berkely.

3.6 Specific concerns and relative prediction

With more and more people getting infected, Federal government must pay more attention to following bad effects. Labor market is hit by the lower efficiency and amount of the labor. Hereby a number of high-end industry are effected. Decrease in people's wage will possibly result in more disharmonious factors in the society.

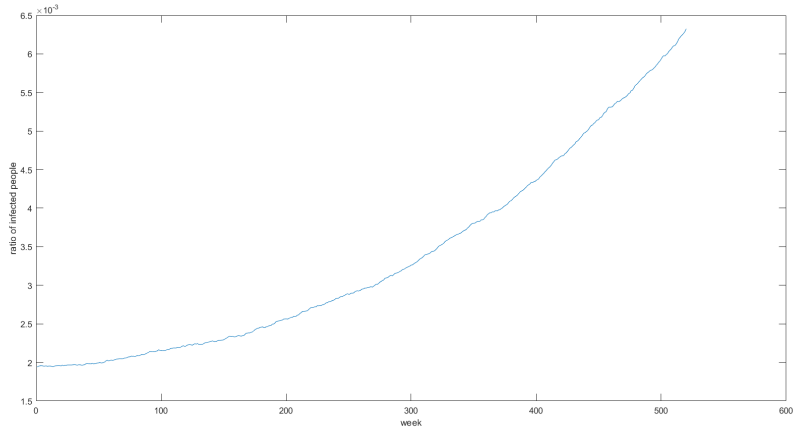

Figure 13: the ratio of infected people to all the people ,forecast,West Virginia,2018 to 2027

We take West Virginia into consideration, and forecast the ratio of infected people to all the people from 2018 to 2027. From the figure, we can find that at about week 320 (i.e.2024), its slope changed sharply, which indicates a great influence to the society and economics. The ratio at that time is $3.5 * 10^{-3}$.

So, for West Virginia, the threshold level will be reached in 2024, with the ratio of infected people $3.5 * 10^{-3}$.

Using the same method, we can get the situation of other states.

For Ohio, the threshold level will be reached in 2023, with the ratio 0.019.

For Pennsylvania, the threshold level will be reached in 2024, with the ratio 0.007.

For Kentucky, the threshold level will be reached in 2022, with the ratio 0.014.

For Virginia, the threshold level will be reached in 2024, with the ratio 0.008.

4 Extension of Model

In fact, the above model ignores the differences between counties, resulting in certain deviations. In this part, we will introduce the counties' data to study the relationship between counties' data and the use of opioids.

4.1 Selecting the socio-economic parameters

In the data imputation process above, we use the PCA method to integrate the given variables into 20 variables. In this step, we need to filter out the variables that affect the drug transmission most. As a more ordinary model, linear regression usually uses the least squares method by searching the smallest square errors between our predictive

value and the actual value.

$$\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} (Y - x_i^T \beta)^2$$

where $N = 3592$, $x_j = \{1, x_{j1}, x_{j2}, ..., x_{j7}\}$

However, this method may meet problems when the dimension of data fitting becomes higher. While the amount of the variables included in the model is increasing, the cost function will naturally decreases. And that makes it difficult to predict the future performance of the model. As there are hundreds of potential variables to choose, We need to identify variables that are truly important to performance indicators.Above all,we introduce a new method called post-LASSO[5].

Post-LASSO is to apply the normal least squares method (OLS) to the model chosen by the LASSO procedure of the first step. The square error is between the predicted value and the actual value. Penalty term is also included in the objective function:

$$\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} (Y - x_i^T \beta)^2 - \lambda ||\beta||$$

where λ is a penalty parameter. The penalty term will force the coefficients of variables to shrink to zero as long as the variable is not important. The penalty coefficient λ determines the degree of penalty for including variables in the model. After minimizing the cost function, we can find the most essential variables in the model.

We use the LASSO toolbox in MATLAB to implement our procedure. The result are as follows:

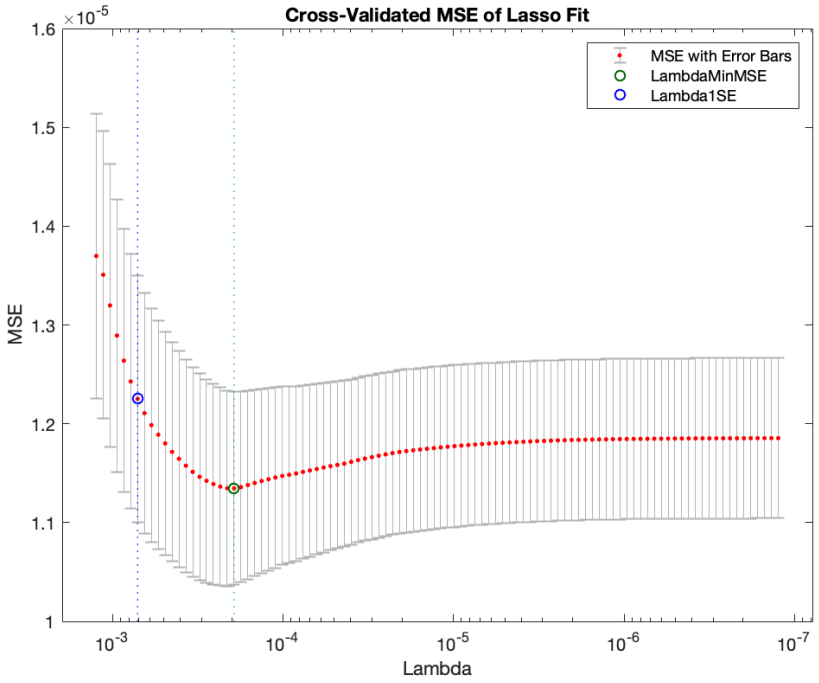

Figure 14: Cross-Validated MSE of Lasso

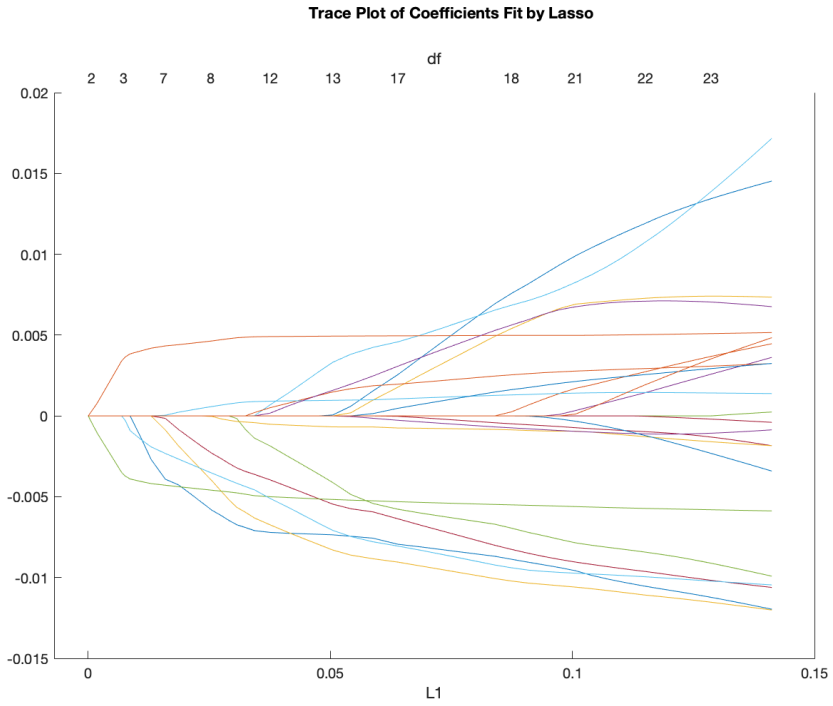

Figure 15: Trace Plot of Coefficients Fit

We use the coefficient from the least MSE situation, and then sort our value by the significance. By doing this, we chose 4 variables. Following is demonstration of these 4 variables and corresponding post-LASSO results:

Table 4: The Chosen Variables

Variables	Fitting outputs
LANGUAGE_SPOKEN_AT_HOME	0.0167
VETERAN	0.0104
HOUSEHOLDS	0.0035
MARITAL_STATUS	0.0014
Observation	3592
R^2	0.74
Adjusted R^2	0.77

where the LANGUAGE_SPOKEN_AT_HOME, VETERAN, HOUSEHOLDS,and MARITAL_STATUS present the percent of people who do not speak English at home, percent of people who live alone, percent of veteran in population, and the percent of people who have high stability of marriage, respectively. For simplicity, we denote them as F_1, F_2, F_3, and F_4 in following content.

4.2 Refitting

In this section, we are going to put our four variables chosen above into our CA model. For simplicity, we use linear model to pair our variables and parameters in our model:

$$\begin{cases} P_{if} = a_{10} + a_{11}F_1 + a_{12}F_2 + a_{13}F_3 + a_{14}F_4 \\ P_{rc} = a_{20} + a_{21}F_1 + a_{22}F_2 + a_{23}F_3 + a_{24}F_4 \\ P_{ir} = a_{30} + a_{31}F_1 + a_{32}F_2 + a_{33}F_3 + a_{34}F_4 \end{cases}$$

where F_1, F_2, F_3, F_4 represent the variables we acquired before, and $a_{10}, a_{11}, ..., a_{34}$ represent their coefficients respectively.

Then, we need to return the genetic algorithm to get those coefficients. By doing this, we can apply socio-economic data to our model and further analyze their influence on the use of drugs. The result of CA is showed below:

Table 5: Result of CA

Variables	Fitting outputs
a_{10}, a_{11}, a_{12}, a_{13}, a_{14}	0.0104, 0.0893, 0.0211, 0.0043, 0.5133
a_{20}, a_{21}, a_{22}, a_{23}, a_{24}	0.0021, 0.0324, 0.0058, 0.0370, -0.0158
a_{30}, a_{31}, a_{32}, a_{33}, a_{34}	0.1623, 0.0518, 0.0550, 0.0688, 0.0344
Generation	173
Minimal MSE	0.832

4.3 Results Analysis

Scrutinizing the fitting outputs above and the minimal MSE is 0.832(<1), thus we can conclude that selected variables truly influence the infection-related probability. For example, we notice that probability of infection is strongly and positively related marital stability. The more stable marriage one own, it is less possible for him to be infected. And the negative relation between probability of recovery and marital stability explains that in some extreme situations, such as divorce, people are more likely to feel miserable and indulge themselves.

5 Strategy Making

5.1 Strategy Elaboration

Based on the outputs of our models above and relevant researches, we set up a strategy to look deep into three dimensions of this crisis and provide suggestions to combat this issue.

1.Individual participation

With imitation about spreading process of the opioid misuse based on SIR model, we have learned that the addiction spreads across counties at a dramatically high speed. As a result, the number of people addicted or overdosed will become surprisingly high, due to the enhanced communication methods. We only consider two participators in this mechanism who are the distributor and the patient.

- Medical and ethical education on physicians should be improved as distributors' ethics as well as their specialty plays an important role in prescribing process. Unrestrained distributors may profit from selling medicine or feel irresponsible for prescriptions they made.
- Patients and their families should be fully informed whether their prescriptions are opioids, the risks of opioid addiction or overdose, control and diversion or potential side effects. Also they need aware the damaging results and the punishment of selling legitimately prescribed opioids.

2.Industrial supervision

For the federal government and all the participators in this industry, they should create an integrated data environment that brings together publicly available data with agency specific data to help address this epidemic.

- Introduce Prescription Drug Monitoring Programs (PDMP) to give prescribers and many pharmacists access to critical information regarding a patient's controlled substance prescription history. It can also help patients who may be misusing prescription opioids or other prescription drugs to examine themselves. Also this program should be under monitor and hold responsibility for uses' privacy.
- Exert health care insurers' influence to the right direction. Sales of prescription opioids in the U.S. nearly quadrupled from 1999 to 2014, most of these is paid by insurance carriers. It is estimated that 1 out of 5 patients with non-cancer pain or pain-related diagnoses are prescribed opioids in office-based settings. Insurance carriers should serve as a stop-gap to the huge influx of opioid prescriptions by lifting the threshold of reimbursement for prescription opioids.

5.2 Testing on effectiveness

We will test two situation as example: enhancing education and introducing PDMP. In the first case, people can receive better education,so their horizon will be broadened, and their knowledge of drug and opioid will be enriched. Therefore, their probability of infection will decrease. In the second case, because the strong restricts on drug and opioids, people will be afraid of abusing drug. Moreover, after people infected, they will be forced to undertake the treatments by DEA. Therefore, their probability of infection will decrease, and their probability of recovery will increase.

We adjust our parameter according to our analysis above, and rerun our model. The figure followed gives the prediction of the number of infected people in different situations form 2018 to 2020.

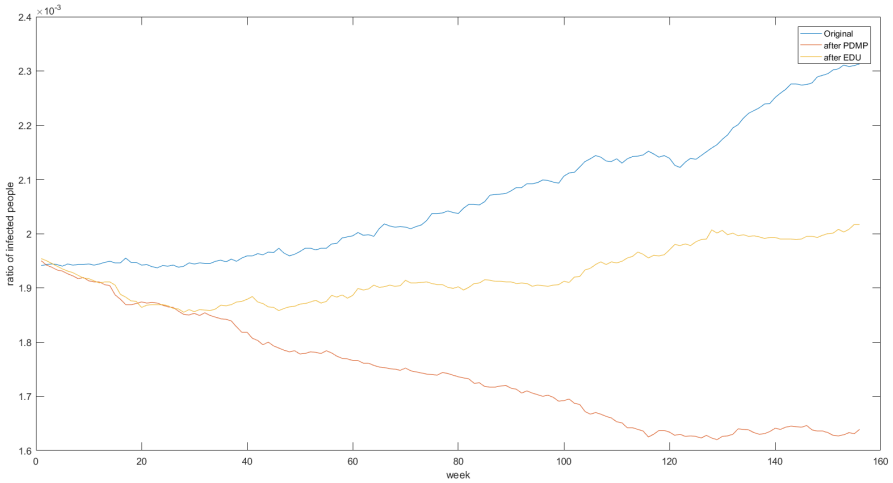

Figure 16: the ratio of infected people to all the people, forecast, Ohio, 2018 to 2020

From this Figure, we can observe that original line is increasing, while the line after education is stable, and the the line after PDMP is decreasing. Therefore, there may be a threshold or bounds of parameter where the situation will be better. So we then try different value of probability of infection, and get a range of this parameter that will make situation better. The result below show the range is $[0, 0.0008]$.

Table 6: The relation of line tendency and P_{if}

P_{if}	0.0006	0.0007	0.0008	0.0009	0.0010	0.0011
Tendency	-	-	-	+	+	+

6 Sensitivity Analysis

We probe into the sensitivity of three parameters in our CA-SIR models. We test the model with state Ohio from 2010 to 2011. As shown below, when we change P_{ri}, P_{rc} and P_{if} from 5% to 5%, the population abusing drug change just slightly. More specifily speaking, the model stay robust when P_{ri} and P_{rc} are disturbed. But the result is sensible to P_{if}. In the most severe case, the result will deviate 5%. But generally speaking, our model is robust. The result is showed in **Figure 17**.

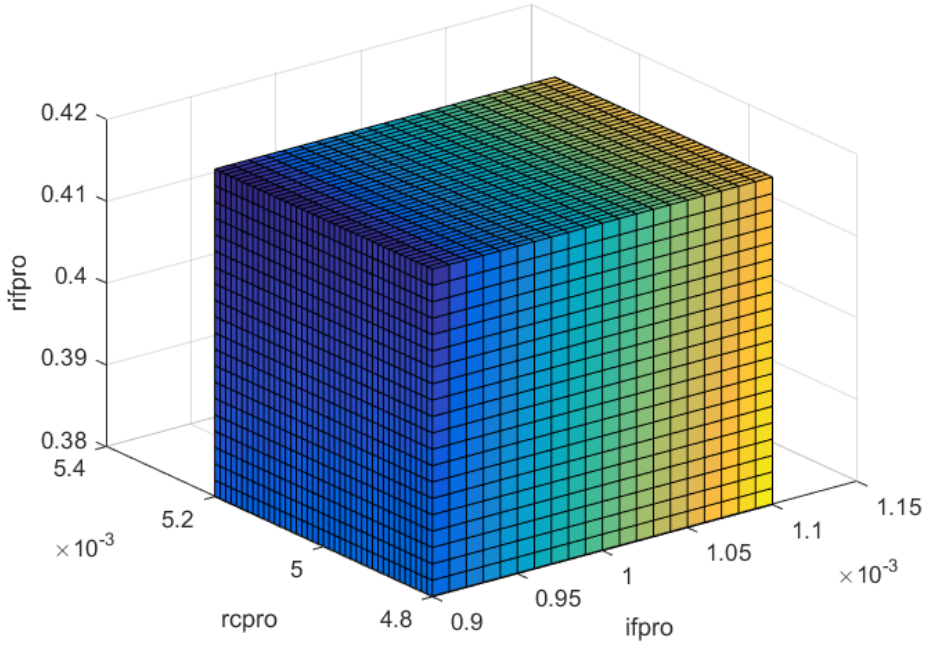

Figure 17: The percent change of drug population
(Blue present 5% and Yellow present -5%)

6.1 Strengths

In our paper, we have done lots of data processing work.We used SIR model for simulating the 5-years spreading process of the opioid overuse and extracted the characteristics. We use various theories and methods such as PCA, SIR, and other theory to complete our work. Generally speaking, we are satisfied with our strengths although weaknesses are unavoidable.Some of the major points are shown below:

- Integrity : We use all the data given comprehensively, either for model testing, or for considering exogenous factors. Thus our model is relatively credible.
- Technical Supporting: We use many theory and methods to support our work. And each one is used reasonably and properly.
- Cross Fusion:We get inspiration from neural network theory and use cellular automaton to support our mathematical model.

6.2 Weaknesses

- Algorithm Limitations. Although the Genetic Algorithm we chose is a suitable method to compute our model, it still has several limitations.The capacity of GA is not very satisfying so the accuracy of the output can't be assured.

- Subjective. There are many subjective methods in our models and some of the parameters are put up by our own experience and intuition, which are not very credible.
- Comprehensiveness. We may not make full use of all the data.

7 Memo to the Chief Administrator of DEA/NFLIS Database

To: Chief Administrator of DEA/NFLIS Database

From: Team 1918889

Date: 27 January 2019

Subject: Insights and suggestions on combating the opioid crisis

We are here to offer our informative modeling results. This memo will illustrate the origins, spreading process and prediction about the future.We also introduce socio-economic factors to make the model more credible. Specific opinions for DEA/NFLIS Database are given following the order of individual participation (physicians and patients), addicts composition, which are based on models we set up as well as relevant researches.

We first focus on the imitation about spreading process of the opioid misuse based on SIR model. It turns out that the addiction spreads out from certain counties of the states and reaches to densely-populated districts[1]. As a result, more and more people will be involved due to the enhanced communication methods. We predict that without proper regulations and necessary measures, the scale of addiction can be considerably large in the near future.

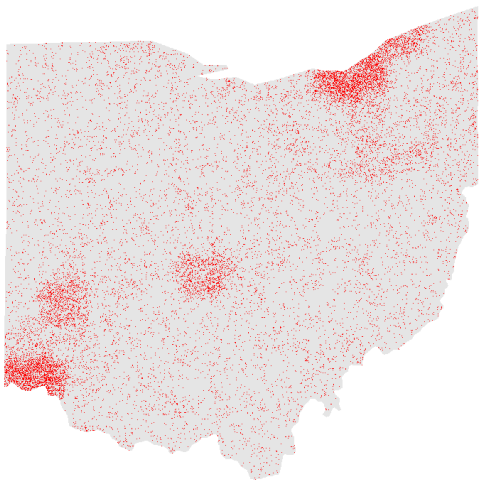

Figure 18: future prediction of Ohio

After that, Taking socio-economic factors into consideration, we obtain a comprehensive understanding about the mechanism of opioids's immoderate spreading and the impact factors behind. Most importantly, public demand for opioids continues rising rapidly, which in turn impose more and more pressure on clinicians to prescribe opioids persists. According to a recent survey from Emergency Department (ED) physicians indicated that 71% reported a perceived pressure to prescribe opioid analgesics to avoid regulatory criticism. To break the cycle, we suggestion contains two parts:

- We should make full use of our big data and help with improve the average quality of the physicians. Meanwhile we can use data supervision to regulate the distributors' behaviour.
- Patients and their families can have access to our database to have a knowledge of their prescribers and whether they are offered with wrong medicine.

Analyzing the outputs of our model, our follow-up research concentrates on categories of household types, marital status, veteran status and language spoken at home. We notice that people who don't have a good command of English are more likely to overuse opioids. They may meet difficulties in understanding the ingredients of the medicine and misuse the opioids. Also single men are morel likely to encounter the overdose as they are less responsible for the family without wife or children. For veterans, many of they passed through the war so they need medicine to keep calm for some times. Hereby our suggestion are as follows:

- Enhance information sharing on DEA/NFLIS Database to give prescribers and many pharmacists access to critical information regarding a patient's controlled substance prescription history. It is also helpful for health professionals to identify patients who may be misusing prescription opioids or other prescription drugs.
- DEA/NFLIS Database should monitor and analyse the abnormal data and help with the policy making procedure due to its big data advantage.

We sincerely thank you for your consideration of our work outputs and suggestions.

References

[1] Skolnick P. The Opioid Epidemic: Crisis and Solutions. Annu Rev Pharmacol Toxicol. 2017 Oct 2.

[2] Akbari, Ziarati (2010). "A multilevel evolutionary algorithm for optimizing numerical functions" IJIEC 2 (2011): 419-430 [1]

[3] ArcGIS, https://developers.arcgis.com

[4] Hamerly, G.; Elkan, C. (2002). "Alternatives to the k-means algorithm that find better clusterings". Proceedings of the eleventh international conference on Information and knowledge management (CIKM).

[5] Brakeman, Leo. "Better Subset Regression Using the Nonnegative Garrote". Technometrics. 37 (4): 373-384.

[6] Malek Messai,Abdeldjalil Aïssa-El-Bey,Karine Amis,Frédéric Guilloud. Iteratively reweighted two-stage LASSO for block-sparse signal recovery under finite-alphabet constraints[J]. Signal Processing,2019,157.

[7] WANG Shuangming.Threshold Dynamics of an SIR Epidemic Model with Nonlinear Incidence Rate and Non-Local Delay Effect[J].Wuhan University Journal of Natural Sciences,2018,23(06):503-513.

[8] H. Dhillon, R. Ganti, J. Andrews, "Load-aware heterogeneous cellular networks: Modeling and SIR distribution", Proc. IEEE GLOBECOM, pp. 4314-4319, Dec. 2012.

[9] M. Minelli, M. Coupechoux, J.-M. Kelif, "Average sir estimation in cellular networks with best server policy", Wireless Days (WD) 2010 IFIP, pp. 1-5, Oct 2010.

Appendix A: Some results of our CA model

Take Ohio as an example.

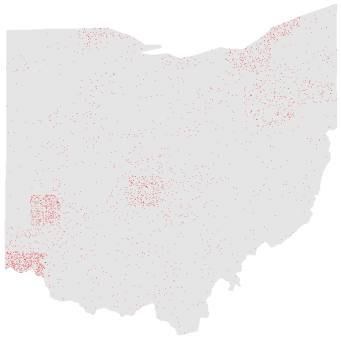

Figure 19: Reality,Ohio,2011

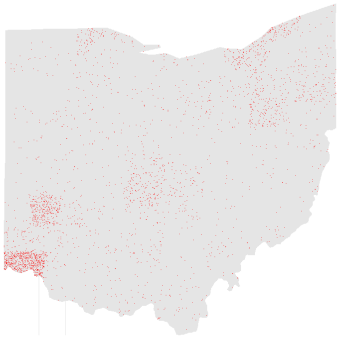

Figure 20: Forecast,Ohio,2011

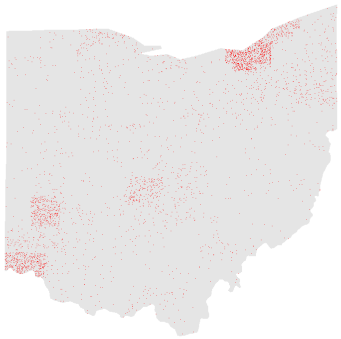

Figure 21: Reality,Ohio,2012

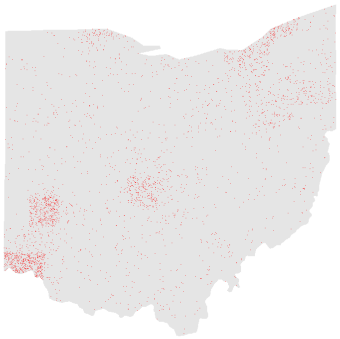

Figure 22: Forecast,Ohio,2012

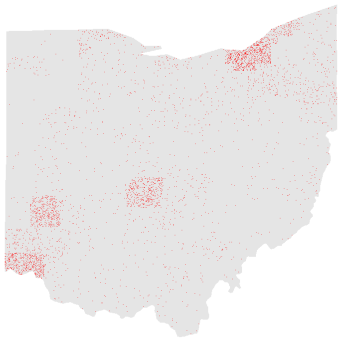

Figure 23: Reality,Ohio,2013

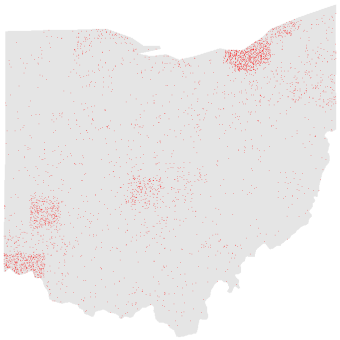

Figure 24: Forecast,Ohio,2013

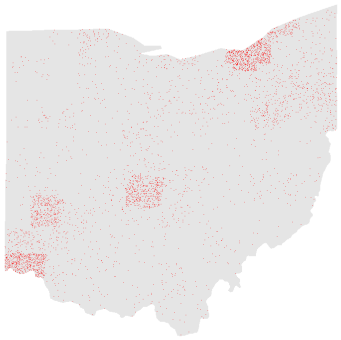

Figure 25: Reality,Ohio,2011

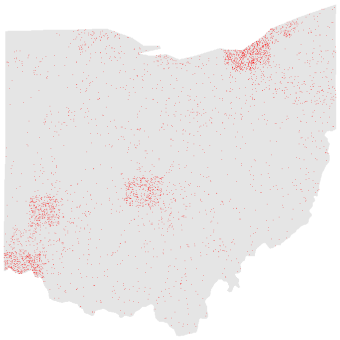

Figure 26: Forecast,Ohio,2011

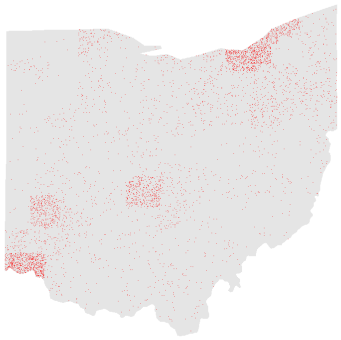

Figure 27: Reality,Ohio,2015

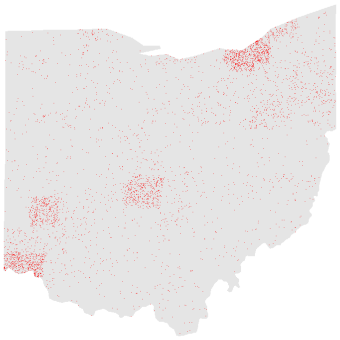

Figure 28: Forecast,Ohio,2015

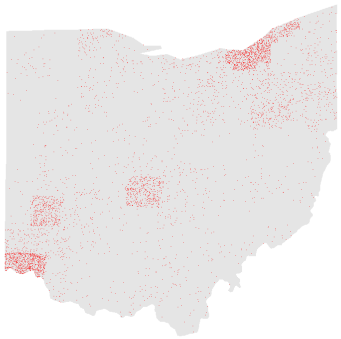

Figure 29: Reality,Ohio,2016

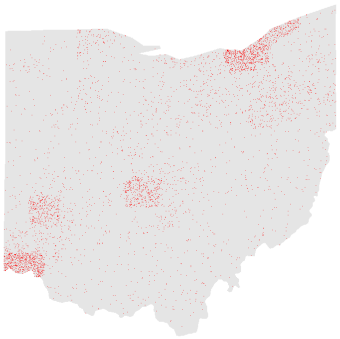

Figure 30: Forecast,Ohio,2016

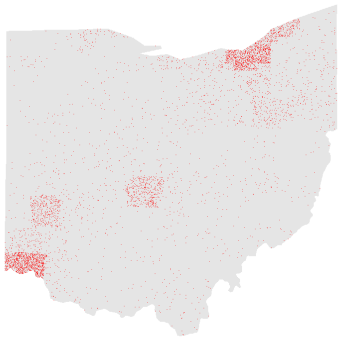

Figure 31: Reality,Ohio,2017

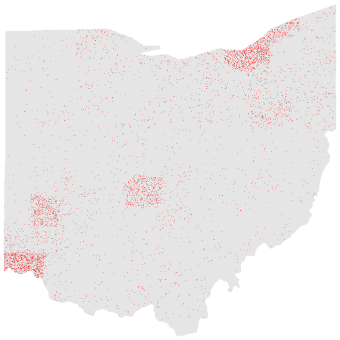

Figure 32: Forecast,Ohio,2017

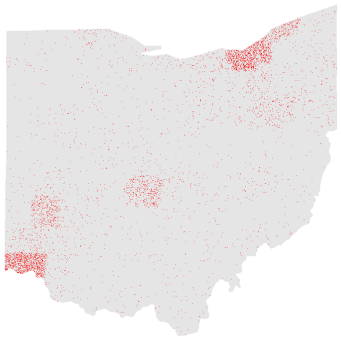

Figure 33: Forecast,Ohio,2018

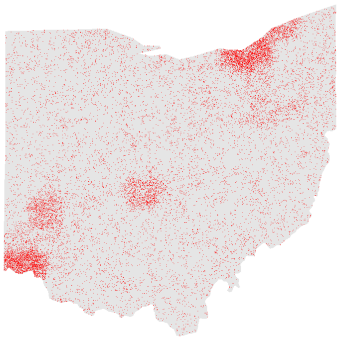

Figure 34: Forecast,Ohio,2027

Appendix B: The source codes

We use the following program to draw the cell map of all the states.

Program 1: `cellplot.m`

```
function [cellmap] = cellplot(state,n)
wholemap = shaperead('UScounties.shp');
%We have already downloaded the whole map of America with the data format '*.shp'
cellmap = zeros(n,n);

cntynum = 0;
for i=1:3141
    if (length(wholemap(i).STATE_NAME)==length(state)) &&...
    all((wholemap(i).STATE_NAME==state)==1)
        cntynum = cntynum+1;
        statemap(cntynum) = wholemap(i);
    end
end

BB = statemap(1).BoundingBox;
XMIN = BB(1,1);YMIN = BB(1,2);
XMAX = BB(2,1);YMAX = BB(2,2);
for i=2:cntynum
    BB = statemap(i).BoundingBox;
    if XMIN > BB(1,1) XMIN = BB(1,1);
    end
    if YMIN > BB(1,2) YMIN = BB(1,2);
    end
    if XMAX < BB(2,1) XMAX = BB(2,1);
    end
    if YMAX < BB(2,2) YMAX = BB(2,2);
    end
end

xlen = (XMAX-XMIN)/nnn;
ylen = (YMAX-YMIN)/nnn;
for i=1:n
for j=1:n
    xlocal = XMIN + (i-1/2)*xlen;
    ylocal = YMIN + (j-1/2)*ylen;
    %PNpoly,determine which county the county belongs to
    for k=1:cntynum
        BB=statemap(k).BoundingBox;
        XX=statemap(k).X;
        YY=statemap(k).Y;
        FIPS=str2double(statemap(k).FIPS);
        if xlocal>=BB(1,1) && xlocal<=BB(2,1) && ylocal>=BB(1,2) && ylocal<=BB(2,2)
        if inpolygon(xlocal,ylocal,XX,YY) == 1
            cellmap(i,j) = FIPS;
        end
        end
    end
end
end

cellmap = rot90(cellmap,1);

end
```

Program 2: ca.m

```
function [xx] = ca(m,n,pwk,pcm,ptv,pif,prc,pri)
load('allx.mat');
%Using the cell map we get from the above program...
%with the data provided by the problem...
%we draw a new cell map and save as allx.
%allx(:,:,m,n) means the mth state's initial situation in nth year.
x=allx(:,:,m,n);
weekn=52;
for time = 1:weekn
    x=move(x,pwk,pcm,ptv);
    x=infect(x,pif,prc,pri);
    disp('WEEK');
    disp(time);
    disp('NUMBER');
    disp(sum(x(:)==0));
    disp(x(:)==1);
    disp(x(:)==2);
end
xx=x;
end
```

Program 3: move.m

```
function [xx]=move(x,pwk,pcm,ptv)

n=1000;
for i = 1:n
for j = 1:n
    if x(i,j)~=3
        u = x(i,j);
        if rand < pwk && i~=1 && x(i-1,j)~=3
            x(i,j) = x(i-1,j);
            x(i-1,j) = u;
        else if rand < pwk && j~=1 && x(i,j-1)~=3
            x(i,j) = x(i,j-1);
            x(i,j-1) = u;
        else if rand < pwk && i~=n && x(i+1,j)~=3
            x(i,j) = x(i+1,j);
            x(i+1,j) = u;
        else if rand < pwk && j~=n && x(i,j+1)~=3
            x(i,j) = x(i,j+1);
            x(i,j+1) = u;
        end
        end
        end
        end

        u = x(i,j);
        if rand < pcm && i>=11 && i<=(n-10) && j>=11 && j<=(n-10)
            mvdt = ceil(10*rand);
            dist = ceil(4*rand);
            if dist == 1 && x(i-mvdt,j)~=3
                x(i,j) = x(i-mvdt,j);
                x(i-mvdt,j) = u;
            else if dist == 2 && x(i,j-mvdt)~=3
                x(i,j) = x(i,j-mvdt);
                x(i,j-mvdt) = u;
            else if dist == 3 && x(i+mvdt,j)~=3
                x(i,j) = x(i+mvdt,j);
```

```
                x(i+mvdt,j) = u;
            else if dist == 4 && x(i,j+mvdt)~=3
                x(i,j) = x(i,j+mvdt);
                x(i,j+mvdt) = u;
            end
            end
            end
            end
        end

        u = x(i,j);
        if rand < ptv
            u1 = ceil(rand*n);
            u2 = ceil(rand*n);
            if x(u1,u2)~=3
                x(i,j) = x(u1,u2);
                x(u1,u2) = u;
            end
        end

    end
end
end

xx=x;
end
```

Program 4: `infect.m`

```
function [xx] = infect(x,pif,prc,pri)

n=1000;
for i = 1:n
for j = 1:n
    if x(i,j) == 0
        u = (i~=1 && (x(i-1,j)==1)) + (j~=1 && (x(i,j-1)==1)) + (i~=n && (x(i+1,j)==1))
        if rand < pif * u
            x(i,j) = 1;
        end
    else if x(i,j) == 1
        if rand < prc
            x(i,j) = 2;
        end
    else if x(i,j) == 2
        u = (i~=1 && (x(i-1,j)==1)) + (j~=1 && (x(i,j-1)==1)) + (i~=n && (x(i+1,j)==1))
        if rand < pri * u
            x(i,j) = 1;
        end
    end
    end
    end

end
end

xx=x;

end
```