

For office use only

T1 _____

T2 _____

T3 _____

T4 _____

Team Control Number

1916709

Problem Chosen

C

For office use only

F1 _____

F2 _____

F3 _____

F4 _____

2019
MCM/ICM
Summary Sheet

The Opioid Crisis: Combined System for Analysis, Prediction and Solution

Summary

The trend of opioids abuse is spreading across the U.S., causing a serious crisis in society. In this paper, we first analyze the spread and characteristics of drug use in five states, then construct a mathematical model of drug spread and distribution, and provide a spatiotemporal prediction model as well as crisis warning mechanisms. Moreover, the correlation between socio-economic factors and drug use is analyzed, according to the socio-economic factors of the U.S. census data set. As a consequence, the advanced model is derived combining the influence of social affairs. Finally, based on the previous model results, we propose and evaluate some feasible policies, and its effectiveness is verified.

Firstly, we propose a space-time distribution model based on Artificial Potential Method (APM) to express the spread and characteristics of opioids. Considering the heatmap spatiotemporal distribution of drug use, we select the Analytic Hierarchy Process (AHP) to measure the distance metrics of the clustering algorithm. Then, the attractive exclusive points used for APM are clustered. Therefore, we perform it to obtain the propagation trends of synthetic and non-synthetic opioids in and between states. In order to estimate the extent of drug abuse in the counties, the thresholds of specific concern are obtained through an iterative algorithm for later prediction. As to reflect the influencing factors of spatial distance in the time sequence prediction model, we propose a two-dimensional Gaussian distribution weighted average method based on spatial neighbors (GDWM) to map the coupling relationship of the spatial neighbors to the cluster centroids. Then, the ARMA algorithm is used to predict the number of drug cases around centroids. The short-term prediction and the concern threshold were combined to provide the crisis warning methods and early warning analysis.

Secondly, considering the influence of social-economic factors, the model has been improved and supplemented. We divide the U.S. Census data into 4 major categories (relationship, education, society and family history), and the principal component analysis is carried out for various factors in each large category, and the characteristics of each major category is described. In order to establish a comprehensive index of socio-economic factors and better describe its correlation with drug use, the PSO and Euclidean distance optimization process is proposed. Therefore, a set of optimal weight vectors is obtained through optimization, and the comprehensive index is obtained by linear weighting. Through correlation analysis, the result shows that there is a certain correlation between drug use and the evaluation index. Besides, in order to better describe the drug use situation and environment, comprehensive consideration of drug use indicators and comprehensive indicators, the best description was obtained by least squares method.

Finally, on the basis of summarizing the previous models, the policies of directly suppressing drug abuse and indirectly affecting social factors are proposed. According to the PMC index, we evaluate the consequence of directly and indirectly suppress policies, and then the iterative prediction method is used to verify the effectiveness of the policies.

Content

1	Introduction	2
1.1	Problem Statement.....	2
1.2	Our Goals.....	2
2	Assumptions and Notations.....	2
2.1	Assumptions	2
2.2	Notations.....	2
3	APF-Based Spread Distribution Prediction	3
3.1	An overview of the modeling scheme	3
3.2	Exploration: the general distribution of drug cases	3
3.3	Inspiration: the generation of artificial potential field	4
3.4	Discovering: AHP based K-Means clustering.....	5
3.5	Construction: spread and characteristics description	7
4	Trend Forecasting of “Concern” Thresholds	8
4.1	The overall scheme of trend forecasting	8
4.2	Determination of “concern” thresholds by iterating.....	9
4.3	GDWM: Gaussian distribution weighted method.....	9
4.4	ARMA prediction results for “Where and When”	10
5	Modified Model: social-economic factors	13
5.1	An overview of the modeling scheme	13
5.2	Principal Components Selection.....	14
5.3	PSO: Particle Swarm Optimization	14
5.4	Establishment of Comprehensive Evaluating Index	16
6	Strategy Formulation and Evaluation.....	17
6.1	Strategy Formulation.....	17
6.2	Policy scoring mechanism based on PMC index	17
6.3	Iterative prediction model.....	18
6.4	Strategy Evaluation.....	18
6	Strengths and Weaknesses	19
6.1	Strengths	19
6.2	Weaknesses	20
7	Conclusions	20
	MEMO	20
	References.....	22

The Opioid Crisis: Combined System for Analysis, Prediction and Solution

1 Introduction

1.1 Problem Statement

The United States is experiencing a nationwide crisis due to the abuse of opioids. In order to solve this drug abuse crisis, try to use the “drug identification results and socio-economic factors information from drug cases analyzed by federal, state, and local forensic laboratories” and census data of five states (Ohio, Kentucky, West Virginia, Virginia and Pennsylvania) in recent years to analyze and forecast, give possible policy solutions and quantify the policy plan to eliminate or alleviate the opioid crisis.

1.2 Our Goals

Based on our understanding of the problem, we set our goals as follows:

- Use the given data to establish a model to describe the spread and characteristics of the opioid cases in and between counties and states.
- Use the given data to model and predict the use of opioid if the patterns keep unchanged.
- Establish a set of early warning systems for government about the drug abuse situation.
- Based on the socio-economic data, we need to consider the impact of both the number of drug reports and the socio-economic factors on drug abuse.
- Test and find the relationship between socio-economic factors and the number of drug reports
- Give a possible strategy that government can use to eliminate the crisis, and need quantitative analysis and prediction of the impact of the strategy.

2 Assumptions and Notations

2.1 Assumptions

Assumption 1. The degree of drug abuse in a certain area is determined by the number of “DrugReports”.

Assumption 2. Ignore the impact of the areas besides the five states provided.

Assumption 3. In short-term forecasts, ignore the impact of policies on related factors over time.

Assumption 4. The provided data is realistic and accurate to a certain degree.

2.2 Notations

Items	Definition
NSO	synthetic opioids
SO	non-synthetic opioids
SEEI	Society-economic Evaluation Index
SEEIM	Society-economic Evaluation Index Matrix
CEI	Comprehensive Evaluation Index
DR	Drug Reports
GDWM	Two-dimensional Gaussian distribution weighted average method based on spatial neighbors

3 APF-Based Spread Distribution Prediction

3.1 An overview of the modeling scheme

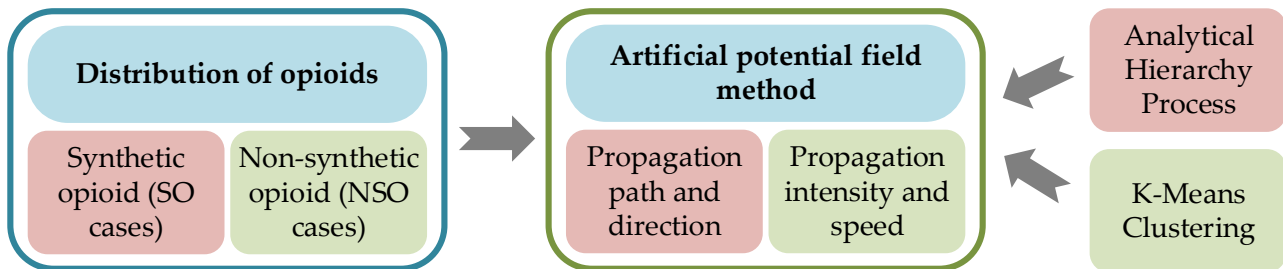


Fig. 1 Modeling scheme overview of drug spread

Before model construction process, an overview of the modeling scheme is provided, as shown in Fig. 7. The model of opioids spread distribution mainly uses the artificial potential field method to characterize the trend direction of drug transmission. At the very beginning, the overall trend is analyzed through the heatmap. In order to determine the attractive and exclusive points of the artificial potential field, the analytic hierarchy process and K-Means clustering algorithm are also needed to classify the data points into three categories, then define two of them as the source and the end points of spread.

3.2 Exploration: the general distribution of drug cases

The general distribution analysis of a data set is the first step in Spatiotemporal modeling, from this perspective, we adopt the Heatmap of different years to exhibit the “DrugReports” distribution among the states and counties.

To distinguish the differences between the spread characteristics of synthetic opioid and heroin incidents, we separately consider the synthetic opioids (SO) and the non-synthetic opioids (NSO) “DrugReports”, as shown in Fig.2 and Fig.3^①.

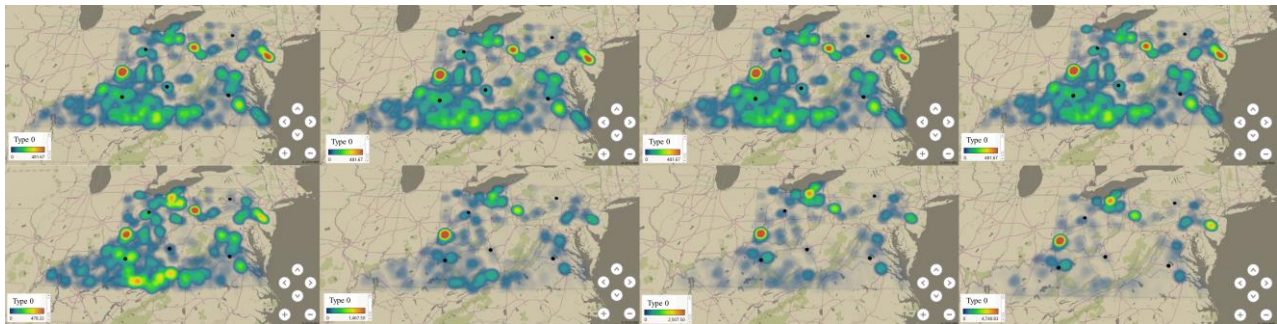


Fig. 2 Heatmap of SO “DrugReports” (2010-2017 respectively from left to right, top to bottom)

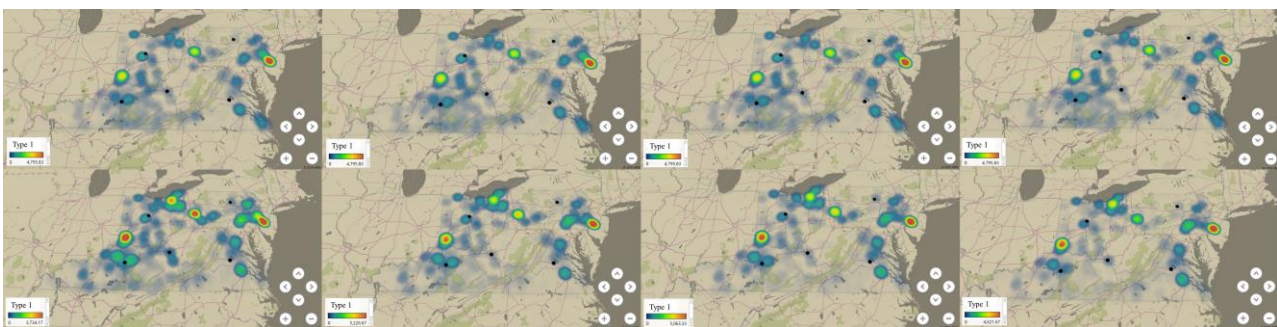


Fig. 3 Heatmap of NSO “DrugReports” (2010-2017 respectively from left to right, top to bottom)

^① The latitude and longitude data of are cited from: https://geonames.usgs.gov/domestic/download_data.htm

According to the dynamic trend of the heatmap distribution, it can be seen that the number of cases of opioid drugs has a certain temporal and spatial distribution pattern. Intuitively, the source of drugs is distributed in two states of PA and OH, while WV has almost no distribution of opioid cases.

In addition, the two opioid cases were analyzed separately. The results indicate that the spread of SO and NSO “DrugReports” cases is different. It is worth noting that the spread of NSO is more severe than that of SO, which can be seen from the severity of the gradation changes in the heatmap. Therefore, the artificial potential field method is used to simulate the spatial characteristics of drug transmission.

3.3 Inspiration: the generation of artificial potential field

Due to the need to obtain the transmission characteristics of opioids in various states and counties, one possible way is to construct a spatial distribution field describing the trend of drug transmission, using the optimization goal of minimizing energy, based on the relevant rules of attraction and exclusion, making the drug transmission path towards the direction of potential energy declines.

The main idea of artificial potential field method is to discover a field function representing the energy of the system, and the object in the field is applied a force either repulsive or attractive, minimizing the energy value of the entire system with the object.

The first step is to build the artificial potential field, which can be divided by two potential energy: the attractive potential and the repulsive potential.

1) the attractive potential

$$U_{att}(\mathbf{x}) = \begin{cases} K_a |\mathbf{x} - \mathbf{x}_d|^2, & |\mathbf{x} - \mathbf{x}_d| \leq d_a \\ K_a (2d_a |\mathbf{x} - \mathbf{x}_d| - d_a^2)^2, & |\mathbf{x} - \mathbf{x}_d| > d_a \end{cases}$$

Where K_a represents the attractive gain, \mathbf{x} is the evaluated point, \mathbf{x}_d is the target point, and d_a is the distance threshold.

2) the repulsive potential

$$U_{rep}(\mathbf{x}) = \begin{cases} \frac{1}{2} K_r \left(\frac{1}{\rho} - \frac{1}{\rho_0} \right)^2, & \rho \leq \rho_0 \\ 0, & \rho > \rho_0 \end{cases}$$

Where K_r represents the repulsive gain, $\rho = |\mathbf{x} - \mathbf{x}_d|$ represents the distance between the evaluated point, and ρ_0 represents the predefined distance threshold.

The second step is to get the derivative of the potential field functions, which represent the force applied to the object in the field.

1) the attractive force

$$F_{att}(\mathbf{x}) = -\nabla U_{att}(\mathbf{x}) = \begin{cases} -2K_a(\mathbf{x} - \mathbf{x}_d), & |\mathbf{x} - \mathbf{x}_d| \leq d_a \\ -2K_a d_a \frac{\mathbf{x} - \mathbf{x}_d}{|\mathbf{x} - \mathbf{x}_d|}, & |\mathbf{x} - \mathbf{x}_d| > d_a \end{cases}$$

2) the repulsive force

$$F_{rep}(\mathbf{x}) = -\nabla U_{rep}(\mathbf{x}) = \begin{cases} K_r \left(\frac{1}{\rho} - \frac{1}{\rho_0} \right) \frac{1}{\rho^2} \frac{\partial \rho}{\partial \mathbf{x}}, & \rho \leq \rho_0 \\ 0, & \rho > \rho_0 \end{cases}$$

Where $\frac{\partial \rho}{\partial \mathbf{x}} = \left(\frac{\partial \rho}{\partial x} \frac{\partial \rho}{\partial y} \right)^T = \frac{\mathbf{x} - \mathbf{x}_0}{\rho}$, and \mathbf{x}_0 is the coordinate vector of the repulsive point nearest to the point \mathbf{x} .

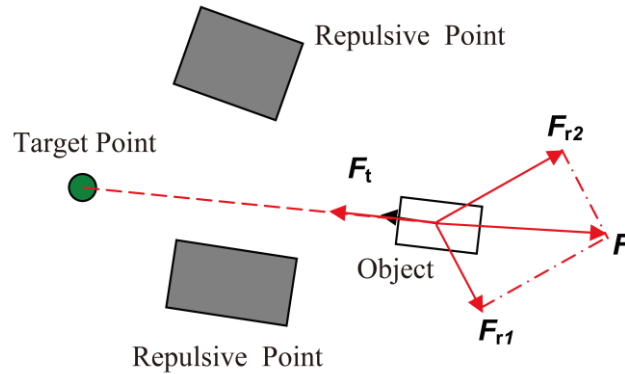


Fig. 4 Schematic diagram of artificial potential field method

The last step is to calculate the resultant force and find the motion direction of the object by combining two coordinate forces, as shown in Fig. 1.

$$\begin{aligned} F(\mathbf{x}) &= -\nabla U(\mathbf{x}) \\ &= -\nabla U_{att}(\mathbf{x}) - \nabla U_{rep}(\mathbf{x}) \\ &= F_{att}(\mathbf{x}) + F_{rep}(\mathbf{x}) \end{aligned}$$


3.4 Discovering: AHP based K-Means clustering

AHP (Analytical Hierarchy Process) is a hierarchical weighted decision analysis method, which is suitable for **assigning homologous weights to different indicators according to their importance**, so that all the indicators can be comprehensively considered for holistic decision-making or analysis.

Through AHP process, we could evaluate and quantify the extent to which these factors affect the comprehensive indicators.

Firstly, we need to construct a **Pairwise Comparison Matrix** based on the degree of influence of these factors on the composite indicator. Since judging the importance in pairs is relatively easy, the **Pairwise Comparison Matrix** $A_{n \times n}$ is constructed, where a_{ij} represents the comparison result of the i -th factor with respect to the j -th factor, and the comparison takes a scale of 1-9 (referred to as **the comparison scale**).

Tab. 1 The Comparison Scale and its Meaning

SCALE	MEANING	
1	the i-th factor is as important as the j-th factor	
3	the i-th factor is a little more important than the j-th factor	
5	the i-th factor is more important than the j-th factor	
7	the i-th factor is significantly more important than the j-th factor	
9	the i-th factor is completely more important than the j-th factor	

Secondly, we need to use the newly established **Pairwise Comparison Matrix** $A_{n \times n}$ for **Consistency Check**. For the Consistency Check, We denote the maximum eigenvalue of $A_{n \times n}$ as λ and define the Consistency Index $CI = \frac{\lambda - n}{n - 1}$, define the Consistency Ratio $CR = \frac{CI}{RI}$ (RI is Random consistency index which we should look it up in the RI Table), when $CR < 0.1$, we regard $A_{n \times n}$ as a consistent matrix.

Finally, if $A_{n \times n}$ is a consistent matrix, then we can use the Normalized Eigenvector corresponding to the maximum eigenvalue λ as the weight vector, which can assign weights to each factor.

For the problem in Part 1, in order to quantify the characteristics of the reported SO and NSO incidents in each county, we use a comprehensive indicator to measure. We assume that the sum of the SO "DrugReports" or NSO "DrugReports" in each county plays a leading role in the comprehensive indicator. But at the same time, its "TotalDrugReportsCounty" will also affect the comprehensive indicator to a certain extent (reflecting the atmosphere of the drug abuse of this county), and "TotalDrugReportsState" will also have an impact on the comprehensive index of the county (reflecting the drug abuse of the state where the county is located). The Pairwise Comparison Matrix $A_{3 \times 3}$ is

$$A_{3 \times 3} = \begin{pmatrix} 1 & 7 & 9 \\ 1/7 & 1 & 3 \\ 1/9 & 1/3 & 1 \end{pmatrix}$$

The result of its corresponding consistency is $CI = 0.0401, CR = 0.0692 < 0.1$, thus we could use the Normalized Eigenvector corresponding to the maximum eigenvalue λ as the weight vector, and finally we get the weight vector as

$$Weight = [0.7854 \quad 0.1488 \quad 0.0658]$$

Based on the weight vector obtained above, we could establish an evaluation index for distance measurement. Specifically speaking, we assign different weights to the distance of different dimensions, and the evaluation index is shown as follow

$$d(x, c_i) = \sqrt{\sum_{j=1}^{len(W)} W_j(x) (x_j^2 - c_{ij}^2)}$$

In order to define the repulsive points and attractive points in artificial potential field, we use the three-dimensional K-Means clustering algorithm to cluster the data points among the three dimensions of "DrugReports", "TotalDrugReportsCounty" and "TotalDrugReportsState" in the original data set. Among them, the DrugReports dimension is divided into SO "DrugReports" and NSO "DrugReports", as described in the previous section.

Our goal is to divide all counties into three categories in three-dimensional space. Below is the basic principle of the K-Means clustering algorithm.

Firstly, randomly select k points as the centroid, group the data points into the same cluster as the nearest centroid, calculate the centroid of each cluster as the new centroid, and repeat the above operations until the centroid position does not change.

The distance measurement adopts the evaluation index obtained by AHP, namely the attribute values of each dimension have different weight coefficients, and the minimum square error based on the evaluation index needs to be obtained. Therefore, the objective function of our clustering is

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} d^2(x, c_i)$$

Where, k is the number of clusters, c_i represents the i -th ($i \in 1, 2, \dots, k$) centroid, and $d(x, c_i)$ is the distance metric.

According to the weight coefficients we previously obtained, the distance metric evaluation index is

$$d(x, c_i) = \sqrt{0.7854 \times (x - DR_i)^2 + 0.1488 \times (x - TDRC_i)^2 + 0.0658 \times (x - TDRS_i)^2}$$

We need to choose the clustering situation corresponding to the minimum SSE. Then, calculate the

k -th centroid DR_k , to find the minimizing SSE value, we should obtain the partial derivative of the SSE evaluation function.

$$\begin{aligned}\frac{\partial}{\partial c_k} \text{SSE} &= \frac{\partial}{\partial c_k} \sum_{i=1}^K \sum_{x \in C_i} d^2(x, c_i) \\ &= \sum_{i=1}^K \sum_{x \in C_i} \frac{\partial}{\partial c_k} d^2(x, c_i) \\ &= \sum_{x \in C_k} 2 \times \begin{bmatrix} 0.7854 \times |x - DR_i| \\ 0.1488 \times |x - TDRC_i| \\ 0.0658 \times |x - TDRS_i| \end{bmatrix}\end{aligned}$$

In the end, the minimizing SSE can be found, along with the changed centroid c_k .

$$m_k c_k = \sum_{x \in C_k} x_k \Rightarrow c_k = \frac{1}{m_k} \sum_{x \in C_k} x_k$$

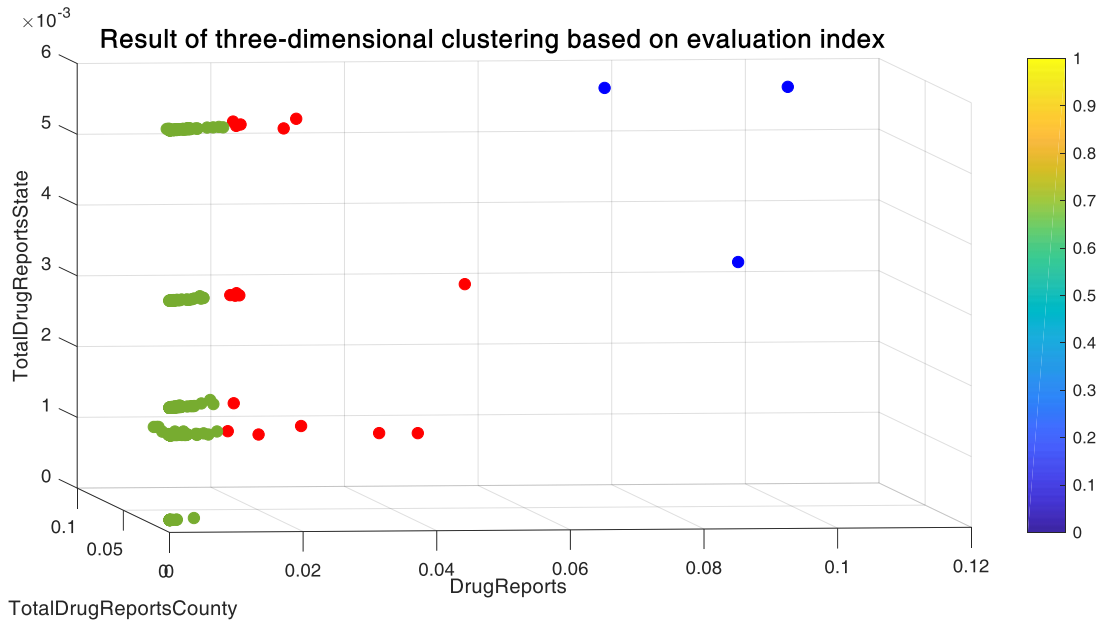


Fig. 5 Result of three-dimensional clustering based on evaluation index

Analysis of results: The division of the three categories is basically consistent with the geographical division between the five states. Combined with the heatmap of drug cases, it can be seen from the clustering results (Fig. 5) that PA and OH are the sources of opioid drug transmission, which can be used as the repulsive points of artificial potential field; VA is the end point of opioid drug transmission, which can be used as the attractive point of the artificial potential field; while WV and KY are between the above two categories, which do not affect the artificial potential field distribution.

3.5 Construction: spread and characteristics description

Through the three-dimensional clustering result, two types can be determined as the attractive exclusive points considering the actual Heatmap distribution, and then the artificial potential field space model of drug distribution will be constructed. The direction of potential field represents the "force" direction of drug transmission. The results of SO and NSO "DrugReports" are shown in Fig. 6 and Fig. 7.

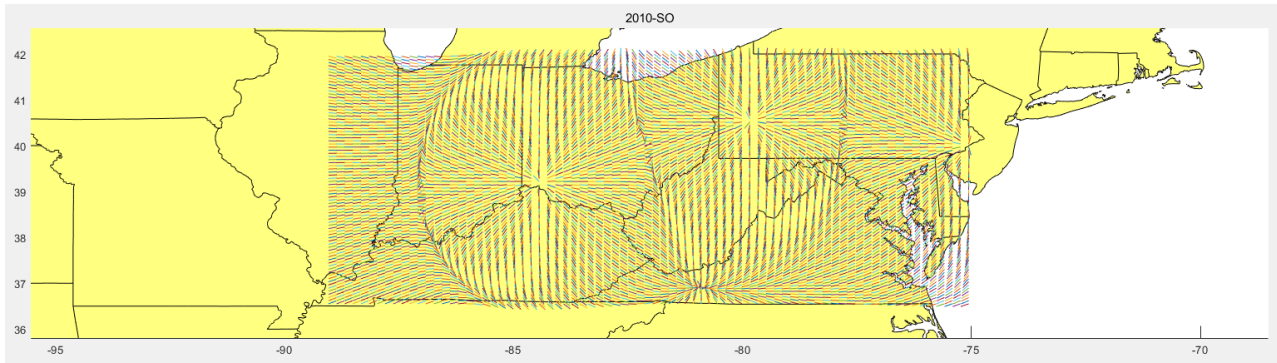


Fig. 6 The spread potential field of SO "DrugReports" in 2010

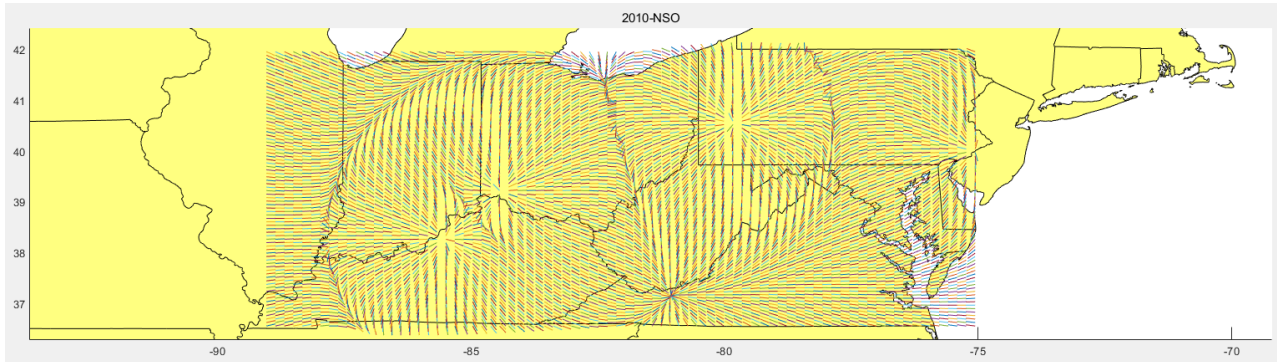


Fig. 7 The spread potential field of NSO "DrugReports" in 2010

We have obtained SO and NSO "DrugReports" of each year (16 results in total), due to the limited space, more for details of all the distribution, please refer to **Appendix**.

The starting point of the drug transmission trend is the county corresponding to the center position obtained by the three-dimensional clustering, and also the position of the exclusion point in the artificial potential field method. In the results of the artificial potential field method, the arrow represents the direction of the potential field gradient at the point, which is also the trend of the opioids spread.

4 Trend Forecasting of "Concern" Thresholds

4.1 The overall scheme of trend forecasting

Firstly, according to the clustering results in previous chapter, there are two thresholds for the three categories. Therefore, we are supposed to find these two thresholds, one possible way is to use the iterative threshold method to determine their exact value.

Secondly, in order to reflect the concept of space-time prediction, it is necessary to combine the influencing factors of spatial neighboring points, thus, we have proposed **Two-dimensional Gaussian distribution weighted average method based on spatial neighbors**.

Finally, by applying the ARMA algorithm, the developing trend of SO and NSO cases can be predicted, in forms of data points sequence. In addition, we can use the interpolation method to interpolate discrete-time sequence points to obtain the trend curve of drug crime situation. Therefore, prediction results of "Where and When" the threshold exceeded, as well as its corresponding spatial position (i.e. latitude and longitude) can be analyzed.

4.2 Determination of “concern” thresholds by iterating

In order to determine the specific concerns, the thresholds need to be defined first. We assume that the US government should have two concerns, one is to prevent specific county from turning from opioid drug transmission to starting to accept drug transmission, and the second is to prevent drug-trafficking areas from turning into a source of drug transmission. Therefore, we need to set two thresholds, predicting “Where and When” the two thresholds will be exceeded, and give the corresponding notice (less severe) and warning (more severe) to help U.S. government take measures.

The determination of the two thresholds uses the iterative threshold method to obtain two thresholds in three types of three-dimensional spatial clustering (class 1 represents the source of propagation, class 2 represents the propagation endpoint, and class 3 represents the vacuum area). **Fig. 8** shows the threshold iterating scheme, in which the two boundaries are exactly what we seek for.

You can split class 1 and class 2, class 2 and class 3 in two. First, select an initial threshold T_1 , divide the population of class 1 and class 2 into two categories; secondly, calculate the mean value μ_1, μ_2 between each category; then, obtain the new threshold $T_2 = (\mu_1 + \mu_2) / 2$; finally, if $|T_2 - T_1| \leq \varepsilon$ or the number of iterations reaches the upper limit, terminate the iteration, otherwise continue iterating.

Because our understanding of concern is the threshold of Drug Identification. Finally, the thresholds are determined to be **1462.3** (from class 3 to class 2) and **10118.6** (from class 2 to class 1) by iteration.

4.3 GDWM: Gaussian distribution weighted method

4.3.1 Involving spatial influence by K-Means clustering

To involve the spatial influence into the evaluation index, we have come up with the idea of, to some extent, measuring spatial neighbors' contribution to the centroids. For this concern, we gathered 100 cluster points, each with some spatial neighbors around. **Fig.9** shows the result of cluster assignments and centroids using K-Means algorithm, and the Euclidean distance in the two-dimensional space is used as the cluster distance index.

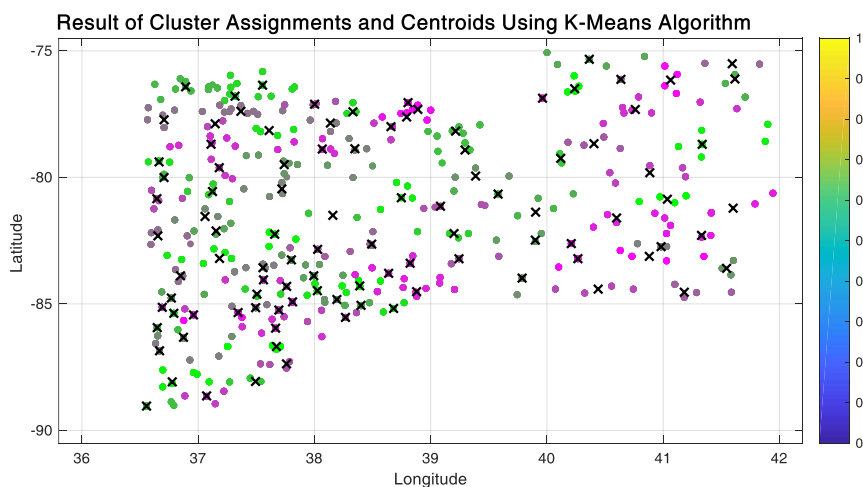


Fig. 9 Result of cluster assignments and centroids using K-Means algorithm

Therefore, we consider the 100 points above as our concern, attributing the “Where and When” question to the time sequence prediction problem of the cluster centroids.

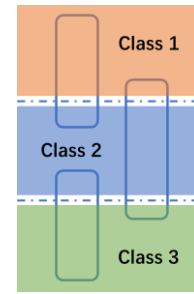


Fig. 8 Schematic diagram of iterative threshold method

4.3.2 Measuring the metric of spatial influence through GDWM

Once the clustering is finished, the influence of the neighborhood of the surrounding space on the center point needs to be considered to reflect the correlation of time and space in the prediction model. As a consequence, we propose “Two-dimensional Gaussian distribution weighted average method based on spatial neighbors (GDWM)”. Using the probability density function of the two-dimensional Gaussian distribution, the influence index of the spatial neighbors is mapped to the centroids, which means the interaction between “TotalDrugCounty” and the spatial neighborhood of the location is characterized.

The general idea of the GDWM method is to pass the attribute value around a certain point through the law of two-dimensional Gaussian distribution, and influence the attribute value of the point according to a series of certain weight coefficients. That is to say, GDWM method is equivalent to smoothing the attribute values of the spatial neighbors, so that the influence of the spatial distance is included in the prediction of the time sequence model, which can also be used to construct the prediction model combining time-spatial factors.

4.3.3 Gaussian distribution probability density function

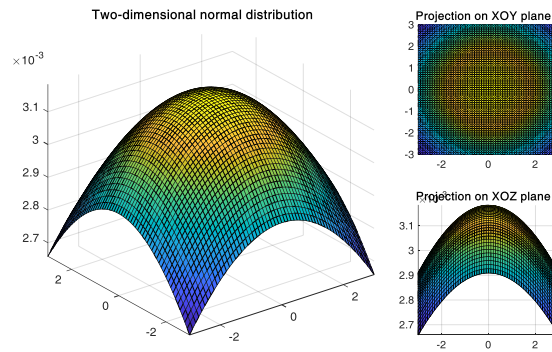


Fig. 10 Two-dimensional Gaussian distribution probability density function

From the Gaussian distribution probability density map (Fig. 10), it can be seen that the height of the center point is the highest, indicating that the main influencing factor is the attribute value of the cluster center point, and the Gaussian distribution is used to calculate the weight coefficients of each point in the neighborhood, which estimates to what extent the nearby points affect centroids.

Therefore, the reason why we consider Gaussian distribution is that Normal distribution is a regular pattern generally satisfying the realistic conditions, and that Gaussian distribution can reasonably reflect spatial distance correction of the centroids and their neighbors.

4.4 ARMA prediction results for “Where and When”

ARMA (Auto-Regressive and Moving Average Model) is a time series analysis model synthesized by AR (Auto-Regressive model) and MA(Moving Average Model). It combines the advantages of AR and MA models. **In ARMA model, AR is responsible for quantifying the relationship between current data and previous data. The MA is responsible for solving the problem of random variables, and ARMA is often used to analyze and predict a relatively stable time series.**

For the p-order **AR model**, its autoregressive performance is to establish a linear regression model between its pre-p data $\{Y_{t-p} \cdots Y_{t-1}\}$ and the current data Y_t , which is expressed as follows:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \epsilon_t$$

For the q-order **MA model**, the moving average is expressed as a linear regression model between the white noise $\{u_{t-q} \cdots u_t\}$ and the current time point data X_t of the previous q time points, which are expressed as follows:

$$X_t = u_t + \phi_1 u_{t-1} + \phi_2 u_{t-2} + \cdots + \phi_q u_{t-q}$$

ARMA(p,q) can supplement the representation of random noise in AR(p) by MA(q), which is expressed as

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + u_t + \phi_1 u_{t-1} + \phi_2 u_{t-2} + \cdots + \phi_q u_{t-q}$$

To reflect the drug abuse situation in each county, consider using the “Total Drug Reports County” data to describe. When the drug abuse situation continues to change according to the previous characteristics, after analysis and testing, consider using the ARMA (3,1) model to predict the drug abuse situation in each county for the next three years, and give the predicted value and the 95% confidence interval. (the upper and lower bounds)

According to the threshold obtained from section 5.2, we may divide counties into three classes, which are the source of propagation, the propagation endpoint and the vacuum area. To better predict the trend of drug-spread distribution over time, we build four standards to illustrate the trend. When the upper confidence limit meets the threshold, we regard it as a notice signal, and when the predict value confidence meets the threshold, we regard it as a warning signal. We have two thresholds to separate the points into three classes, thus, we may have two kinds of situations need to be warning and noticed, which are vacuum points transform to propagation endpoints (warning 3 to 2, notice 3 to 2) and propagation endpoints transform to source of propagation (warning 2 to 1, notice 2 to 1). The results are shown in Fig. 11 ~ Fig. 14 and Tab. 3 ~ Tab. 6.

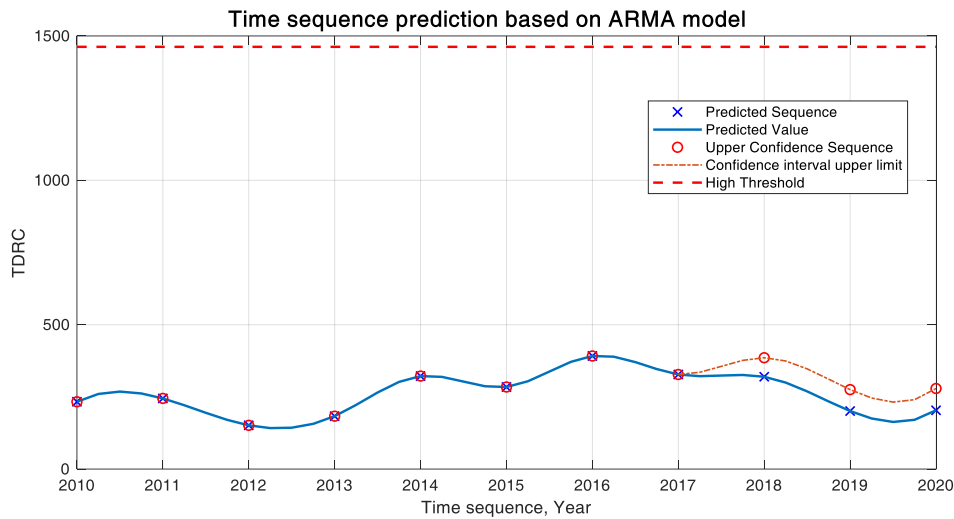


Fig. 11 Time sequence prediction based on ARMA model for data1

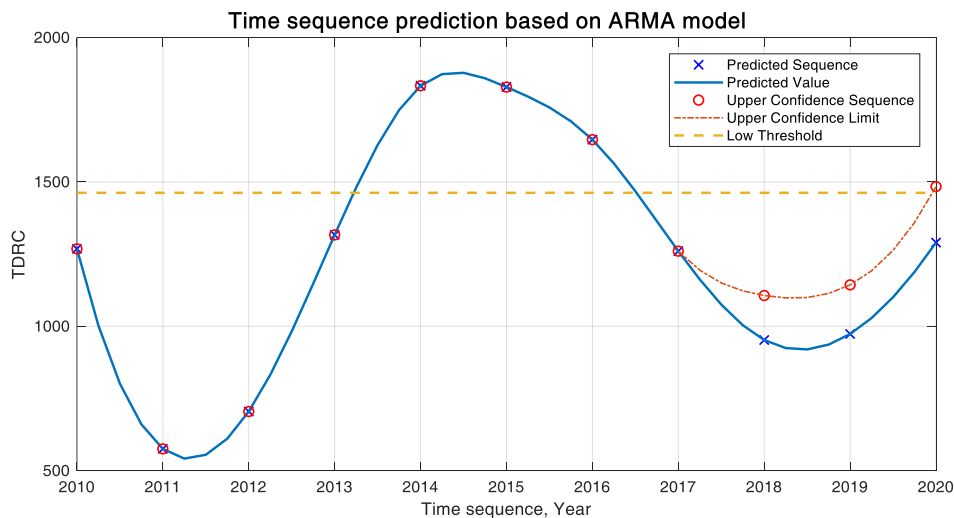


Fig. 12 Time sequence prediction based on ARMA model for data2

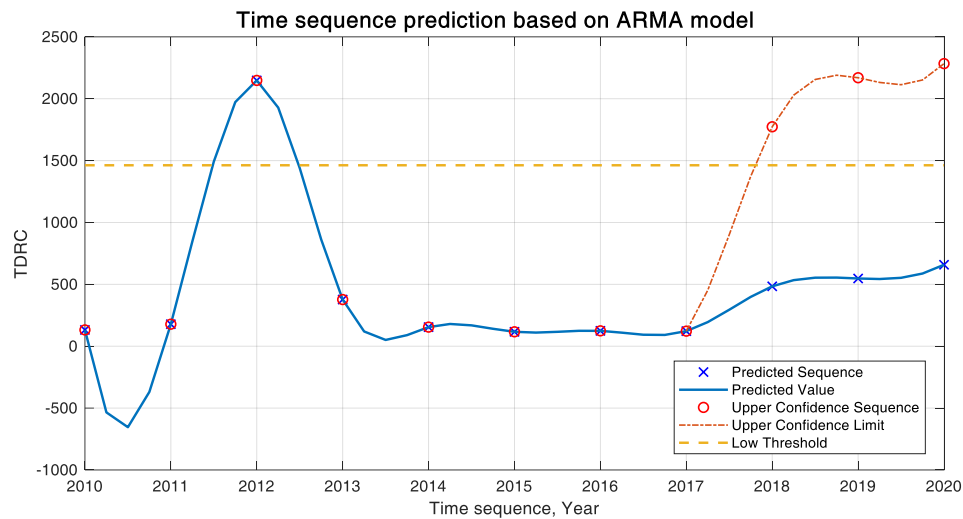


Fig. 13 Time sequence prediction based on ARMA model for data3

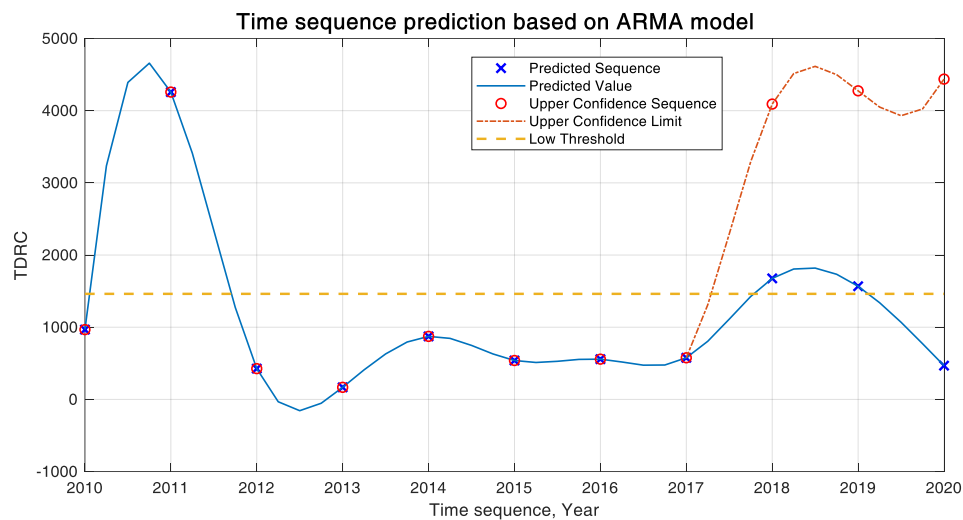


Fig. 14 Time sequence prediction based on ARMA model for data4

Tab. 3 Notice 3 to 2

class	latitude	longitude	Time
2	41.03575	-80.8633	2020
7	41.33562	-78.6931	2018
14	38.33124	-77.408	2018
18	37.31681	-76.7927	2017.5
21	37.4931	-88.0642	2017.75
23	38.1583	-81.5003	2017.5
42	41.54519	-83.5941	2018.75
43	39.38648	-79.9478	2017.75
47	37.71957	-80.4548	2017.75
49	40.982	-82.7378	2017.75
52	38.13736	-77.8544	2017.75
53	37.49424	-85.1488	2017.5
62	41.18144	-84.5372	2017.75
65	39.2434	-83.2169	2018

class	latitude	longitude	Time
80	40.76022	-77.3173	2018.5
82	37.11043	-78.6913	2017.75
86	38.63813	-83.7899	2017.75
91	41.33005	-82.294	2018
96	38.88956	-77.315	2017.5

Tab. 4 Warning 3 to 2

class	latitude	longitude	Time
23	38.1583	-81.5003	2018
53	37.49424	-85.1488	2018.75
62	41.18144	-84.5372	2018.5
96	38.88956	-77.315	2017.75

Tab. 5 Notice 2 to 1

class	latitude	longitude	Time
34	40.36427	-75.3343	2020
88	41.60186	-81.2176	2018

Tab. 6 Warning 2 to 1

None

5 Modified Model: social-economic factors

5.1 An overview of the modeling scheme

In this section, we firstly analyze the attributes of the provided data set and divided them into four categories C_i . According to relevant literatures, some irrelevant attributes are excluded. By applying Principal Components Analysis (PCA), principal components are selected as evaluation index for different categories.

Afterwards, we attempt to figure out a linear combination of principal components as Social-Economic Evaluation Index (SEEI). We propose Particle Swarm Optimization method, through which we could heuristically obtain the best weight W_i for the linear combination of different category C_i . By calculating the Pearson Coefficient between the Social-economic Evaluation Index Matrix (SEEIM) and the "DrugReports" matrix, we are supposed to evaluate the correlation between Social-Economic Evaluation Index and "DrugReports" quantitatively.

Finally, we attempt to combine Social-Economic Evaluation Index and drug spread level, which is represented by "DrugReports", in order to reach a Comprehensive Evaluation Index (CEI). Comprehensive Evaluation Index is a linear combination of "DrugReports" and Social-Economic Evaluation Index, in other words, $CEI = a \cdot DR + b \cdot SEEI$. Here, we apply the Least Squares Regression Method to obtain the optimal weights a, b for linear combination.

5.2 Principal Components Selection

Because the U.S. Census Bureau involves nearly 150 attribute types, each attribute type corresponds to four description indicators; we first preprocess the data to obtain the principal components that best characterize the data set, thereby reducing the influence of the arithmetic data and irrelevant variables.

To start with, we divide the attributes in the data set into four categories, where C_1 represents the relationship issues, C_2 represents the education issues, C_3 represents the society issues, C_4 represents the family history issues.

After that, some of the irrelevant attribute values are eliminated by consulting the relevant documents. After that, the data is normalized according to the attribute value, and the PCA dimensionality reduction operation is performed to obtain the principal component eigenvalue λ_i , the factor load matrix A_{ij} and the corresponding factor property P_{ij} corresponding to different categories C_i . By the relation $U_{ij} = A_{ij} / \sqrt{\lambda_i}$, the corresponding principal component load matrix U_{ij} can be obtained, so that the principal components F_i of different classes C_i are obtained as follows

$$F_i = \sum_j U_{ij} P_{ij}$$

Owing to the fact that nearly 150 kinds of attributes are involved, dimension reduction process is essential. Through data preprocessing, we could obtain the principal components that best characterize the data, which could not only reduce the impact of unrelated variables, but also reduce the data size for further processing.

5.3 PSO: Particle Swarm Optimization

In order to establish the relationship between the four socio-economic descriptive indicators $\{C_1 C_2 C_3 C_4\}$ extracted in 5.2 and the drug abuse situation, **consider using a weight vector $\omega = [\omega_1 \omega_2 \omega_3 \omega_4]$ to linearly construct a comprehensive socio-economic factor evaluation index(SSEI) through four socio-economic description indicators**

$$SSEI = [\omega_1 \omega_2 \omega_3 \omega_4][C_1 C_2 C_3 C_4]^T$$

In order to better establish the relationship between the SSEI and the drug abuse situation, consider constructing a **NumberOfCounty×NumberOfYear dimension** global SSEI matrix (SSEIM) and a DrugReport matrix (DRM) by the SSEI of each county and its number of DrugReports per year (**Row represents each county, column represents each year**). And find an optimal weight vector $[\omega_1 \omega_2 \omega_3 \omega_4]$, so that the Euclidean distance between the DRM and the SSEIM is the smallest. Its visual description is as follows

$$DRM = \begin{bmatrix} dr_{11} & \cdots & dr_{1m} \\ \vdots & \ddots & \vdots \\ dr_{n1} & \cdots & dr_{nm} \end{bmatrix}_{n \times m} \quad SSEIM = \begin{bmatrix} ss_{11} & \cdots & ss_{1m} \\ \vdots & \ddots & \vdots \\ ss_{n1} & \cdots & ss_{nm} \end{bmatrix}_{n \times m}$$

The problem can be translated into the following mathematical description

$$\min u = f(\omega_1, \omega_2, \omega_3, \omega_4) = \|DRM - SSEIM\|_2$$

In order to obtain the weight vector, which minimizes u , PSO is adopted. **PSO (Particle Swarm Optimization)** is an evolutionary algorithm proposed by J. Kennedy and R.C. Eberhart et al. in 1995. Similar to the Simulated Annealing, **it starts from a random solution and finally finds the best solution through numerical iteration**. Because of its fast convergence, high precision and easy implementation,

PSO is suitable for numerical optimization. The algorithm absorbs the natural principles of bird flocking, its iterative optimization process and flow chart are as follows:

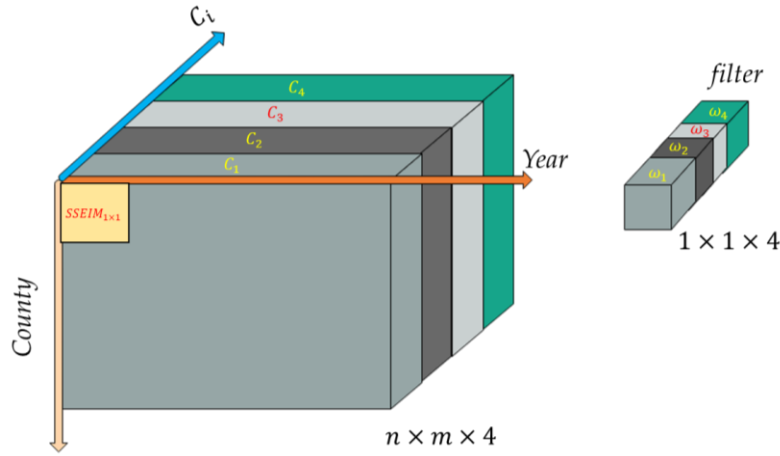


Fig. 15 Schematic diagram of calculating elements in SSEIM

The d-dimensional velocity update formula of the i -th particle

$$V_{id}^{k+1} = wV_{id}^k + c_1 \text{rand}(P_{id}^k - X_{id}^k) + c_2 \text{rand}(P_{gd}^k - X_{id}^k)$$

The d-dimensional position update formula of the i -th particle:

$$X_{id}^{k+1} = X_{id}^k + V_{id}^{k+1}$$

V_{id}^k : The d-th component of the i -th particle's velocity vector at the k -th iteration.

X_{id}^k : The d-th component of the i -th particle's position vector at the k -th iteration.

P_{id}^k : The d-th component of the historical best position vector of the i -th particle at the k -th iteration.

P_{gd}^k : The d-th component of the historical best position vector of the entire particle swarm at the k -th iteration.

w : Inertia weight; c_1, c_2 : Learning factor; **rand** : a [0,1] random number, increase search randomness

The process flow chart is shown in Fig. 16.

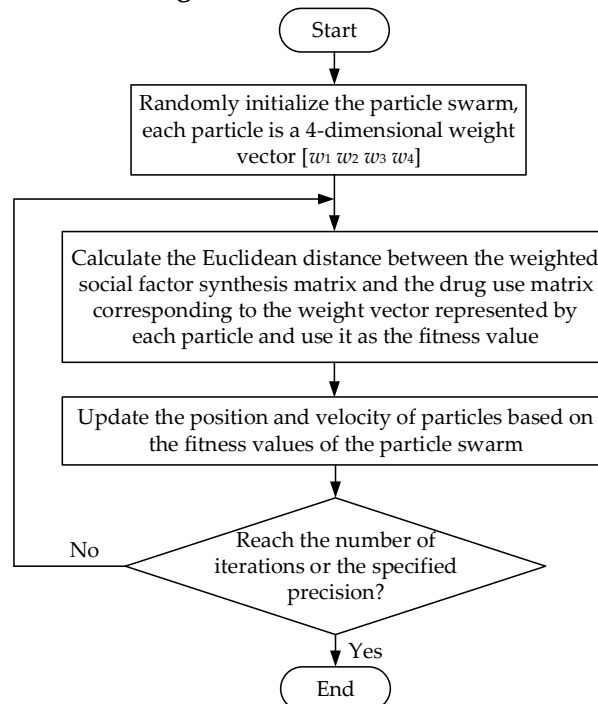


Fig. 16 Program flow chart of particle swarm algorithm

Finally, initialize 1000 particles, $w = 1.4$, $c_1 = c_2 = 2$, after iteration 40,000 times, we obtain a set of better solutions (maybe not the optimal solution) which get the Euclidean distance $u = 0.3903$) $\omega = [\omega_1 \ \omega_2 \ \omega_3 \ \omega_4] = [-0.97585, 0.934164, 0.474319, -0.56624]$. Therefore, the comprehensive SSEI can be established. And make it better describe the drug abuse situation

$$SSEI = [-0.97585, 0.934164, 0.474319, -0.56624][C_1 \ C_2 \ C_3 \ C_4]^T$$

In order to find the relationship between SSEI and drug abuse, the correlation coefficient between SSEIM and DRM is sought. **Since each county is an independent individual, the correlation coefficient between the SSEI and the sequence of the DR in different years can be obtained to describe the correlation between the SSEI and the DR of the county.** Finally, the correlation coefficient between SSEI and DR of each county is counted, and the whole is obtained

Tab. 7 Similarity Statistics

Correlation	Highly Relevant	Strongly Relevant	Moderately Relevant	Weekly Relevant	Almost Irrelevant
Number of Counties	24	92	84	92	86

So we can draw a conclusion that more than half of the counties have at least moderate similarity between SSEI and DR.

5.4 Establishment of Comprehensive Evaluating Index

After obtaining the Society-economic Index, we need to comprehensively consider the regional drug trafficking activeness and Society-economic Index. Therefore, we propose Comprehensive Evaluation Index (CEI) which is shown as follow

$$CEI = a \cdot DR + b \cdot SSEI$$

In order to determine the optimal weight for linear combination, we apply the Least Squares Regression. By considering $(DR_i, SSEI_i)$ of different counties in different years as independent variables, and DR_{i+1} as dependent variable, we are supposed to represent dependent variables, independent variables, weight parameters and the residual matrix by matrixes as follows

$$X = \begin{bmatrix} DR_1 & SSEI_1 \\ DR_2 & SSEI_2 \\ \vdots & \vdots \\ DR_n & SSEI_n \end{bmatrix}, Y = \begin{bmatrix} DR_2 \\ DR_3 \\ \vdots \\ DR_{n+1} \end{bmatrix}, P = \begin{bmatrix} a \\ b \end{bmatrix}, E = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Linear regression value could be obtained as $\hat{y} = XB$, and the corresponding residual matrix could be calculated as $E = \hat{Y} - Y = XB - Y$. Define Euclidean distance as $L(P) = \|XB - Y\|^2$, the goal of Least Squares Regression is to minimize the norm of the residual matrix, which could also be represented by $L(P)$. By making the partial derivative equal to zero, we could characterize the minimum value of $L(P)$

$$\frac{\partial L(P)}{\partial P} = 0 \Rightarrow X^T X P - X^T Y = 0 \Rightarrow P = (X^T X)^{-1} X^T Y$$

Based on Least Squares Regression, we could figure out the optimal weight of the CEI linear combination is $a = 1.0154, b = -0.0106$, **which means** $CEI = 1.0154DR - 0.0106SSEI$.

Above all, CEI is a Comprehensive Evaluate Index we established to better describe the drug trafficking activeness. According to the previous analysis, **four important factors, which are**

Relationship, Education, Status and Family history, should be included in order to better construct our model.

6 Strategy Formulation and Evaluation

6.1 Strategy Formulation

According to the Social-economic index and the Comprehensive evaluation index obtained in the previous chapter, we can construct an attribute vector $\mathbf{Atr}_{i,k} = [DR_{i,k} \ R_{i,k} \ E_{i,k} \ S_{i,k} \ F_{i,k}]^T$ to describe the condition of the i -th county in k -th year, therefore, we are supposed to propose corresponding policies influencing the elements in attribute vector $\mathbf{Atr}_{i,k}$. Our proposed policies mainly include the 4 aspects as follows.

- Suppress the birthplace of drug trafficking;
- Strengthen community construction and improve people's livelihood;
- Strengthen education, improve national quality of people and raise their awareness of drugs;
- Strengthen customs control over imports and exports.

6.2 Policy scoring mechanism based on PMC index

To quantitatively evaluate the effectiveness of a policy, we introduce the concept of the PMC index. According to the principal component selection in Section 5.2, the four social factor attribute values can be attributed to the description of their corresponding factors. The core idea of the PMC index is to measure the influence level of the policy's impact on its corresponding category, with a series of weighted sum of the influence on the factors contained in the category.

During the policy scoring analysis process, we use $S_{ij} \in \{-3, -2, -1, 0, 1, 2, 3\}$ to describe the influential extent of how a policy will affect the j -th factor of category i . The meaning of the S_{ij} value is shown in **Tab. 8**. Due to the limited space, the detailed results of the evaluation of the influential extent of each factor are listed in **appendix**.

Tab. 8 The meaning of the S_{ij} value

S_{ij}	Definition
-3	Strategy has strong inhibitory effects on the i -th factor of the j -th category
-2	Strategy has moderate inhibitory effects on the i -th factor of the j -th category
-1	Strategy has slight inhibitory effects on the i -th factor of the j -th category
0	Strategy has no effect on the i -th factor of the j -th category
1	Strategy has slight promoting effects on the i -th factor of the j -th category
2	Strategy has moderate promoting effects on the i -th factor of the j -th category
3	Strategy has strong promoting effects on the i -th factor of the j -th category

Then, according to the contribution evaluation index of the factor in each category (**i.e.** the linear weight value ω_{ij} after PCA dimension reduction), sum up the influential index by weighted coefficients.

Therefore, we can obtain the total influence degree of the policy on each category $S_i = \sum_j \omega_{ij} S_{ij}$, which is exactly the policy-affected PMC index.

In order to associate the policy-affected PMC index with the attribute value of categories, we propose

the concept of the impact factor, which is a number within the interval $[0.8, 1.2]$. Each category i can be mapped into the impact factor I_i by linear mapping operation, according to the PMC index. Finally, we obtained the value of impact factors of different categories, as shown in the **Tab. 9**.

Tab. 9 The value of impact factors of different categories

Categories	Influence Index
Drug Report	0.80
Personal Relationship	0.92
Education Background	1.10
Society Status	1.00
Family History	0.93

Describe all the impact factors in the form of impact factor matrix, referred to as

$$\mathbf{I} = \text{diag}[I_{DR} \quad I_R \quad I_E \quad I_S \quad I_F]$$

For the attribute vector $\mathbf{Atr}_{i,k}$ of the i -th county in k -th year, the attribute vector of the next year can be expressed as $\mathbf{Atr}_{i,k+1} = \mathbf{I} \times \mathbf{Atr}_{i,k}$. Therefore, through the impact factor matrix, we are supposed to represent influential extent of a policy into the attribute vector.

6.3 Iterative prediction model

Once the impact factors of each category are derived, we proposed an iterative prediction model to evaluate the effectiveness of the policy.

We adopt the value of “DrugReports” to characterize the activeness of drug trafficking in a specific region. As explained in the previous section, based on the established Social-economic Index and the Comprehensive Evaluation Index, using the filter $\omega = [\omega_1 \quad \omega_2 \quad \omega_3 \quad \omega_4]^T$ (consistent with the previous one) and $P = [a \quad b]^T$, we can iteratively predict the $DR_{i(k+1)}$ value of the next year. Besides, according to our hypothesis, the influential degree of the policy on the category is approximately constant in the short-term forecast. Therefore, we can iteratively update and solve the social factors $R_{ik}, E_{ik}, S_{ik}, F_{ik}$ of the next year, using the solved influence factor matrix. According to these conditions, we can continuously update the attribute vector $\mathbf{Atr}_{i,k+1}$ of the next year, realizing the expectation of iterative prediction.

6.4 Strategy Evaluation

Based on the iterative prediction model proposed in Section 6.3 and the corresponding impact factor matrix, we input the iterative prediction model based on the attribute value data from 2010 to 2016 to predict the impact of policy implementation on the degree of drug trafficking in different counts over a three-year period. In order to more intuitively display the impact of the policy, we print out two forecast images showing the average level of the “DrugReport” and Comprehensive Evaluation Index for all the counts after the policy implementation, as shown in **Fig. 17** and **Fig. 18**.

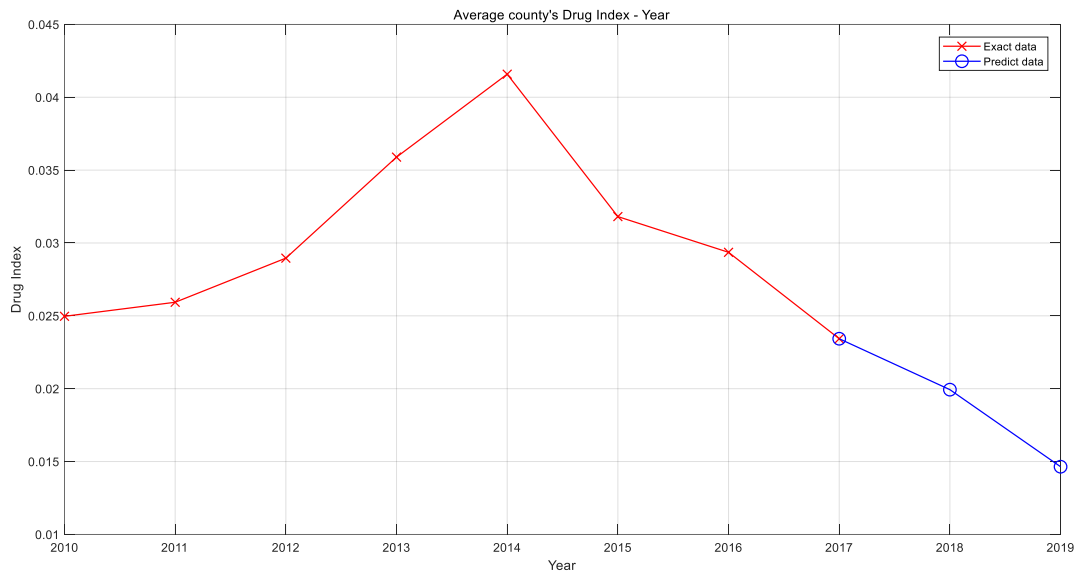


Fig. 17 Average county's Drug Index ~ Year

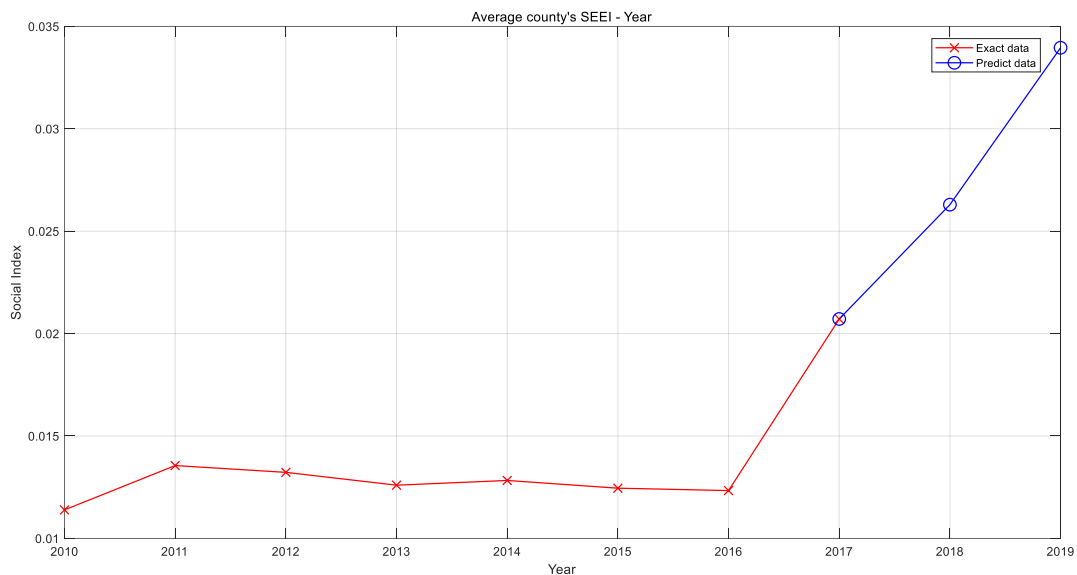


Fig. 18 Average county's SEEI ~ Year

According to Fig. 17, using our hypothesis, we can infer that the value of "DrugReports" of all counties will show a downward trend, and the Social-economic Index showed an upward trend, after the implementation of the policy (2017 or later). The results indicate that the drug trafficking situation in the United States will be better, and that the policies we proposed will have a positive effect.

6 Strengths and Weaknesses

6.1 Strengths

- **Data preprocessing:** PCA is applied to extract main variables, which transmits 150 attributes into 4 main attributes. Therefore, the influence of the arithmetic data and irrelevant variables are reduced.
- **Intuitive drug-spread field distribution:** Based on Artificial potential field method, drug-spread distribution of the whole area is intuitively characterized.

- **PSO algorithm:** It's another implementation of BP neural network based on heuristic algorithm in solving optimization problems. Compared with BP neural network, PSO algorithm could sharply reduce training time and obtain a relatively accurate solution in a shorter time. By applying Pearson coefficient, the correlation between Social-economic Index and drug trafficking activeness could be quantitatively judged.
- **High generalizability:** The model could be applied to many similar types of problems, such as infectious disease transmission, economic radiation, and talent attraction problems.

6.2 Weaknesses

- **Subjective factors involved:** AHP and PMC Index involve subjective factors to some degree, anyway, they are also an objective way to quantitatively evaluate subjective factors.
- **Iterative Prediction Model hypothesis:** In the Iterative Prediction Model, it is assumed that, in short-term forecasts, the impact of policies on various factors stay the same over time, but this hypothesis may need further consideration.

7 Conclusions

This paper analyzes the cases of opioid drug crimes in five states, and obtains the drug propagation model based on artificial potential field and the space-time prediction model based on ARMA time sequence prediction algorithm. In addition, taking into account the impact of socio-economic factors on drug transmission, through the analysis and other means, the relationship between socio-economic factors and the number of drug cases is analyzed, and the space-time prediction model is improved, with more reasonable results obtained. Finally, on the basis of summarizing the previous models, policies are proposed to directly inhibit drug abuse and indirectly affect social factors. Based on the PMC index, we evaluate the consequences of direct and indirect suppression strategies and then use iterative prediction methods to verify the effectiveness of the strategy.

MEMO

From: Team 1916709, MCM 2019

To: The Chief Administrator, DEA/NFLIS Database

Date: Jan 28, 2019

Subject: The prediction and analysis of the opioid crisis, and how to deal with the opioid abuse crisis.

Dear chief administrator,

we are honored to inform you our achievement after performing data analysis and modeling.

At present, the government should suppress the drug abuse situation in five states, mainly on how to formulate a targeted suppression policy. Our team has studied the historical data from 2010 to 2017, and obtained a heat map of the degree of drug abuse, and has predicted the trend of drug transmission and abuse from 2018 to 2020. Mainly as follows

The possible locations where specific opioid use might have started : PHILADELPHIA, MONTGOMERY, HAMILTON, BUCKS, ALLEGHENY. 60% sources of drugs are in Pennsylvania, so more strict rules should be passed in Pennsylvania in order to limit drug abuse.

If the patterns and characteristics keep unchanged, we have predicted the future drug abuse situations in all the counties. We figure out that, more attention should be paid to some counties such as (we get their locations)

LAT	LON	YEAR/MONTH
38.1583	-81.5003	2018/1
37.49424	-85.1488	2018/9
41.18144	-84.5372	2018/6
38.88956	-77.315	2017/9

When reaches the time we list in the table above, the predict value may meet the threshold, which means the drug abuse situation may become much worse.

Based on the data we have analyzed, we make some specific policies in order to influence the drug abuse situations directly or indirectly. Firstly and the most importantly, US government should suppress the birthplace of drug trafficking in order to reduce the drug trafficking activeness directly, which is also the most important way. Secondly, strengthen community construction and improve people's livelihood. Since people who live alone tend to be more likely to abuse drugs. So through this strategy, we may form a more harmonious atmosphere which may help people to develop a sense of belonging. In return, they may be less likely to abuse drugs. Thirdly, governments are supposed to strengthen education, improve national quality of people and raise their awareness of drugs. Since people attain higher education level and develop a higher drug abuse awareness, they will get to know more about the danger of drug addiction. Finally, customs should strengthen control over imports and exports, in order to prevent the spread of drugs to some degree.

We sincerely hope that you can pay attention to the opioid crisis and adapt some suggestions we have proposed above.

Thanks!

Please contact us if you have any problems.

References

- [1] Thomas L. Saaty. (2008): "Decision making with the analytic hierarchy process. Int. J. Services Sciences", Vol. 1, No. 1, 2008.
- [2] F. Janabi-Sharifi, D. Vinke, "Integration of the artificial potential field approach with simulated annealing for robot path planning", Proceedings of the 1993 IEEE International Symposium on Intelligent Control, pp. 536-541, 1993.
- [3] Bowden, N, Merino, R., Katamneni, S., Coustasse, A. (2018, April). "The cost of the opioid epidemic in West Virginia". Paper presented at the 54th Annual MBAA Conference, Chicago, IL
- [4] Saaty, T.L. (2008). "Decision making with the analytic hierarchy process", Int. J. Services Sciences, Vol. 1, No. 1, pp.83–98.
- [5] Wang, j. s. Yu, "Fault feature selection based on modified binary pso with mutation and its application in chemical process fault diagnosis," Lecture Notes in Computer Science, 2005, v3612, pp. 832-840.

Appendices

Appendix 1: Artifical Potential Field && K-mean

filename: "ArtificalPotentialField.m"

```
filename: "ArtificalPotentialField.m"clear;
clc;
close all
%% Initialization
%Range of the Map
point_num=60;
Xrang =[-89.04,-75.06]; %lonitude represents the X
Yrang =[36.55,41.95]; %latitude represents the Y
X_point = linspace(Xrang(1),Xrang(2),point_num); %discrete points' X
Y_point = linspace(Yrang(1),Yrang(2),point_num); %discrete points' Y
for i=1:point_num %Point is the array of discrete point
    for j=1:point_num
        Point((i-1)*point_num+j,1) = X_point(i);
        Point((i-1)*point_num+j,2) = Y_point(j);
    end
end
Analyze_Matrix(:,1) = xlsread('kmeans_2018.xlsx',1,'C:C'); % When read the
excel, it should be in descending order
Analyze_Matrix(:,2) = xlsread('kmeans_2018.xlsx',1,'B:B'); % When read the
excel, it should be in descending order
Analyze_Matrix(:,3) = xlsread('kmeans_2018.xlsx',1,'M:M');
Analyze_Matrix(:,4) = xlsread('kmeans_2018.xlsx',1,'I:I');
%% Get the Resistant Matrix according to threshold_high
[row,~] = size(Analyze_Matrix); %Get row
j=1;
```

```

for i=1:row %traverse the Analyze_Matrix
    if(Analyze_Matrix(i,3)==1)
        Resistant_Matrix(j,1:2) = Analyze_Matrix(i,1:2);
        Resistant_Matrix(j,3) = Analyze_Matrix(i,4);
        j = j+1;
    end
end
%% Get the Attract Matrix according to threshold_low
j=1;
for i=1:row
    if(Analyze_Matrix(i,3)==2)
        Attract_Matrix(j,1:2) = Analyze_Matrix(i,1:2);
        Attract_Matrix(j,3) = Analyze_Matrix(i,4);
        j = j+1;
    end
end
%% Get the Unreachable Matrix
j=1;
for i=1:row
    if(Analyze_Matrix(i,3)==3)
        Unreachable_Matrix(j,1:2) = Analyze_Matrix(i,1:2);
        Unreachable_Matrix(j,3) = Analyze_Matrix(i,4);
        j = j+1;
    end
end
%% traverse all the discrete point to get each point's force
for i=1:point_num^2
    [F(i,1),F(i,2)] =
force(Attract_Matrix,Resistant_Matrix,Unreachable_Matrix,Point(i,1),Point(i,2)
); %The ith Point
    F_magnitude(i) = sqrt(power(F(i,1),2)+power(F(i,2),2));
    F_argument(i) = atan2(F(i,1),F(i,2));
end
%% plot
mapshow('cb_2017_us_state_500k.shp');
hold on
for i=1:point_num^2
    quiver(Point(i,1),Point(i,2),0.25*F(i,1)/sqrt(F(i,1)^2+F(i,2)^2),0.25*F(i,2)/s
qrt(F(i,1)^2+F(i,2)^2),'filled');
    hold on
end
% plot3(Analyze_Matrix(:,1),Analyze_Matrix(:,2),Analyze_Matrix(:,3));
title('2018-SO');

```

filename: "force.m"

```
%%
%input: attract matrix (3*m); reject matrix(3*n), edges:xmin,xmax,ymin,ymax
%return: F(Fx,Fy)
function [Fx,Fy]=force(A,R,U,x,y)
    da=50; %attract distance threshold
    dr=100; %reject distance threshold
    du=0.5; %unreach distance threshold
    Ka=10e2; %coefficient of attract force
    Kr=10e5; %coefficient of reject force
    Ku=10e7; %coefficient of unreach force
    [rowa,~]=size(A);
    [rowr,~]=size(R);
    [rowu,~]=size(U);
    Fx=0;
    Fy=0;
    %% Attract_Matrix and its force calculate
    for i=1:rowa
        dis=sqrt((A(i,1)-x)^2+(A(i,2)-y)^2);
        if(dis>da)
            ct=(A(i,1)-x)/dis;
            st=(A(i,2)-y)/dis;
            Fx=Fx+2*Ka*A(i,3)*da*ct;
            Fy=Fy+2*Ka*A(i,3)*da*st;

        else
            Fx=Fx+2*Ka*A(i,3)*(A(i,1)-x);
            Fy=Fy+2*Ka*A(i,3)*(A(i,2)-y);

        end
    end
    %% Resistance_Matrix and its force calculate
    for i=1:rowr
        dis=sqrt((R(i,1)-x)^2+(R(i,2)-y)^2);
        if(dis<=dr)
            Fx=Fx-Kr*R(i,3)*(1/dis-1/dr)/dis^3*(R(i,1)-x);
            Fy=Fy-Kr*R(i,3)*(1/dis-1/dr)/dis^3*(R(i,2)-y);
        end
    end
    %% Unreachable_Matrix and its force calculate
    for i=1:rowu
        dis=sqrt((U(i,1)-x)^2+(U(i,2)-y)^2);
        if(dis<=du)
```

```

%           Fx=Fx-Ku*U(i,3)*(1/dis-1/dr)/dis^3*(U(i,1)-x);
%           Fy=Fy-Ku*U(i,3)*(1/dis-1/dr)/dis^3*(U(i,2)-y);
%       end
%   end
end

```

filename: "FunK_mean3D.m"

```

function [ resX,resY, resZ,record,class] = FunK_mean3D( x,y,z,k )
    k1=0.7854;
    k2=0.1488;
    k3=0.0658;
    j = 1;
    seedX = zeros(1,k);
    seedY = zeros(1,k);
    seedZ = zeros(1,k);
    oldSeedX = zeros(1,k);
    oldSeedY = zeros(1,k);
    oldSeedZ = zeros(1,k);
    resX = zeros(k,length(x));
    resY = zeros(k,length(x));
    resZ = zeros(k,length(x));
    class = zeros(length(x),1);

    record = zeros(1,k);
    for i = 1:k
        seedX(i) = x(round(rand()*length(resX)));
        seedY(i) = y(round(rand()*length(resX)));
        seedZ(i) = z(round(rand()*length(resX)));

        if (i > 1 && seedX(i) == seedX(i-1) && seedY(i) == seedY(i-1) &&
seedZ(i) == seedZ(i-1))
            i = i -1;
        end
    end
    while (1)
        record(:) = 0;
        resX(:) = 0;
        resY(:) = 0;
        resZ(:) = 0;
        for i = 1:length(x)
            distanceMin = 1;
            for j = 2:k
                if (k1*power(x(i)-seedX(distanceMin),2)+k2*power(y(i)-
seedY(distanceMin),2)+k3*power(z(i)-seedZ(distanceMin),2))...

```

```

        > k1*(power(x(i)-seedX(j),2)+k2*power(y(i)-
seedY(j),2)+k3*power(z(i)-seedZ(j),2))
        distanceMin = j;
    end
end
class(i,1)=distanceMin;
resX(distanceMin,record(distanceMin)+1) = x(i);
resY(distanceMin,record(distanceMin)+1) = y(i);
resZ(distanceMin,record(distanceMin)+1) = z(i);
record(distanceMin) = record(distanceMin) + 1;
end
oldSeedX = seedX;
oldSeedY = seedY;
oldSeedZ = seedZ;

for i = 1:k
    if record(i) == 0
        continue;
    end
    seedX(i) = sum(resX(i,:))/record(i);
    seedY(i) = sum(resY(i,:))/record(i);
    seedZ(i) = sum(resZ(i,:))/record(i);
end

if mean([seedX == oldSeedX seedY == oldSeedY seedZ == oldSeedZ]) ==
1 %if seedX == oldSeedX && seedY == oldSeedY
    break;
end
end
maxPos = max(record);
resX = resX(:,1:maxPos);
resY = resY(:,1:maxPos);
resZ = resZ(:,1:maxPos);
end

```

filename: "gauss.m"

```

function weight = gauss(center,points)
    mu = center;
    sigma = [50,0;0,50];
    haha = mvnpdf(points,mu,sigma);
    wahaha = sum(haha);
    weight = haha/wahaha;
end

```

filename: "point.m"

```

clear
clc
%load data
% data=xlsread('kmeans_2010.xlsx','H2:K446');
% data=xlsread('kmeans_2011.xlsx','H2:K428');
% data=xlsread('kmeans_2012.xlsx','H2:K432');
% data=xlsread('kmeans_2013.xlsx','H2:K447');
%data=xlsread('kmeans_2014.xlsx','H2:K441');
% data=xlsread('kmeans_2015.xlsx','H2:K425');
% data=xlsread('kmeans_2016.xlsx','H2:K438');
data=xlsread('kmeans_2017.xlsx','H2:K429');
[ resX,resY,
resZ,record,class1]=FunK_mean3D(data(:,1),data(:,3),data(:,4),3);
[ resX,resY,
resZ,record,class2]=FunK_mean3D(data(:,2),data(:,3),data(:,4),3);

```

Appendix 2: ARMA

filename: "ARIMA_Predict.m"

```

function [predict,UB,LB] = ARIMA_Predict(y_ori,num)
%% ARIMA Model
p=3;d=0;q=1;
Mdl = arima(p,d,q);
EstMdl = estimate(Mdl,y_ori,'Display','off');
res = infer(EstMdl,y_ori);
[yF,yMSE] = forecast(EstMdl,num,'Y0',y_ori);
UB = yF + 1.96*sqrt(yMSE);
LB = yF - 1.96*sqrt(yMSE);
predict = yF;

```

filename: "predict_main.m"

```

close all;clear;clc
% threshold=10118.6; %threshold value
threshold=1462.3; %threshold value
year=3; %years need to be predicted
step=0.25;
TDRC=xlsread('TDRC_gauss.xlsx','A1:H101');
center_pos=xlsread('kmeans_center.xlsx','B2:C101');
x=TDRC(1,:);
for i=1:year
    x=[x,i+2017];
end
[row,~]=size(TDRC);

```

```

notice=[]; %when UB meets threshold
warning=[]; %when center meets threshold
for i=21:30
    disp(i)
    data=TDRC(i,:);
    [predict,UB,LB] = ARIMA_Predict(data,year);
    data_center=[data;predict]';
    data_UB=[data;UB]';
    xx=2010:step:(2017+year);
    center_predict=spline(x,data_center,xx);
    UB_predict=spline(x,data_UB,xx);
    center_after17=center_predict(1,(7/step+1):length(center_predict));
    UB_after17=UB_predict(1,(7/step+1):length(UB_predict));
    xx_after17=xx(1,(7/step+1):length(xx));

    figure(i)
    plot(x,data_center,'bx',xx,center_predict);
    hold on;
    plot(x,data_UB,'ro',xx_after17,UB_after17);
    hold on
    plot(x,threshold+zeros(1,length(x)),'-');

    flag_w=0;
    flag_n=0;
    for j=1:(length(xx)-7/step)

if(center_after17(1)<threshold&&center_after17(j)>=threshold&&flag_w==0)
        warning=[warning;[i-1,center_pos(i-1,1),center_pos(i-
1,2),xx_after17(1,j)]];
        flag_w=1; %once record, no more further information to be recorded
in warning matrix
    end
        if(UB_after17(1)<threshold&&UB_after17(j)>threshold&&flag_n==0)
            notice=[notice;[i-1,center_pos(i-1,1),center_pos(i-
1,2),xx_after17(1,j)]];
            flag_n=1; %once record, no more further information to be recorded
in notice matrix
        end
    end
end
end

```

Appendix 3: PSO && LeastSquare

filename: "LeastSquare.m"


```

clc;clear;close all
%% data read
X1 = xlsread('aandb.xlsx',1); %drugi
X2 = xlsread('aandb.xlsx',2); %social-economic
Y = xlsread('aandb.xlsx',3); %drugi+1
%% Least Square Method
X = [X1,X2];
p = X\Y; %coefficient
Y_predict = X*p;
%% to value the good or bad
SSE = sum((Y_predict-Y).^2);
MSE = SSE/length(X);
RMSE = sqrt(MSE);
y_av = sum(Y)/length(Y);
SST = sum((Y-y_av).^2);
R_square = 1-SSE/SST;

```

filename: "euclidean_dis.m"

```

%fx = euclidean_dis(x,W1,W2,W3,W4,drugreports)
function fx = euclidean_dis(x,W1,W2,W3,W4,drugreports)
    [N,~] = size(x);
    for i=1:N
        weight = x(i,:); %1*4
        W5 = weight(1)*W1+weight(2)*W2+weight(3)*W3+weight(4)*W4;

        C = (W5-drugreports).^2;
        fx(i,:)=sqrt(sum(C(:)));
    end
end

```

filename: "PSO.m"

```

clc;clear;close all;
N = 1000;
d = 4;
ger = 10000*d;
limit = [-10, 10];
vlimit = [-(limit(2)-limit(1))/10, (limit(2)-limit(1))/10 ];
w = 1.4;
c1 = 2;
c2 = 2;
x = limit(1) + (limit(2) - limit(1)) * rand(N, d);
v = vlimit(2)*((rand(N, d)-0.5)/0.5);

```

```

xm = x;
ym = zeros(1, d);

fxm = 1000000000+zeros(N, 1);
fym = 1000000000;
drugreports = xlsread('drug.xlsx',1,'B2:H378');
W1 = xlsread('1.xlsx',1,'B2:H378');
W2 = xlsread('2.xlsx',1,'B2:H378');
W3 = xlsread('3.xlsx',1,'B2:H378');
W4 = xlsread('4.xlsx',1,'B2:H378');
iter = 1;
record = zeros(ger, 1);
while iter <= ger
    fx = euclidean_dis(x,W1,W2,W3,W4,drugreports) ;
    for i = 1:N
        if fxm(i) > fx(i)
            fxm(i) = fx(i);
            xm(i,:) = x(i,:);
        end
    end
    if fym > min(fxm)
        [fym, nmin] = min(fxm);
        ym = xm(nmin, :);
    end
    v = v * w + c1 * rand * (xm - x) + c2 * rand * (repmat(ym, N, 1) - x);

    v(v > vlimit(2)) = vlimit(2);
    v(v < vlimit(1)) = vlimit(1);
    x = x + v;

    x(x > limit(2)) = limit(2);
    x(x < limit(1)) = limit(1);
    record(iter) = fym;
    iter = iter+1;
end

weight = ym;
SE = weight(1)*W1+weight(2)*W2+weight(3)*W3+weight(4)*W4;
for i=1:7
    SE(:,i) = SE(:,i)/norm(SE(:,i));
end
for i=1:length(SE)
    s=corrcoef(drugreports(i,:)',SE(i,:));
    Pearson(i,:) = s(1,2);
end

```

Appendix 4: PART3

filename: "PSO.m"

```

clc;clear;close all
co_drug = 0.8;co_rela = 11/12;co_edu = 1.1;co_soc = 1;co_famhis = 14/15;
a=1.0154;b=-0.0106;
w1=-0.97585;w2=0.934164;w3=0.474319;w4=-0.56624;
% COEFF = [co_drug;co_rela;co_edu;co_soc;co_famhis];
drug_now = xlsread('drug.xlsx',1,'H2:H378');%2016
rela_now = xlsread('1.xlsx',1,'H2:H378');
edu_now = xlsread('2.xlsx',1,'H2:H378');
soc_now = xlsread('3.xlsx',1,'H2:H378');
famhis_now = xlsread('4.xlsx',1,'H2:H378');
ES_now =w1*rela_now+w2*edu_now+w3*soc_now+w4*famhis_now;

drug_n1 = a*(co_drug+0.05*randn(1,1))*drug_now+b*ES_now;%2017
rela_n1 = rela_now*(co_rela+0.03*randn(1,1));
edu_n1 = edu_now*(co_edu+0.03*randn(1,1));
soc_n1 = soc_now*(co_soc+0.03*randn(1,1));
famhis_n1 = famhis_now*(co_famhis+0.03*randn(1,1));
ES_n1 = w1*rela_n1+w2*edu_n1+w3*soc_n1+w4*famhis_n1;

drug_n2 = a*(co_drug+0.05*randn(1,1))*drug_n1+b*ES_n1;%2018
rela_n2 = rela_n1*(co_rela+0.03*randn(1,1));
edu_n2 = edu_n1*(co_edu+0.03*randn(1,1));
soc_n2 = soc_n1*(co_soc+0.03*randn(1,1));
famhis_n2 = famhis_n1*(co_famhis+0.03*randn(1,1));
ES_n2 = w1*rela_n2+w2*edu_n2+w3*soc_n2+w4*famhis_n2;

drug_n3 = a*(co_drug+0.05*randn(1,1))*drug_n2+b*ES_n2;%2019
rela_n3 = rela_n2*(co_rela+0.03*randn(1,1));
edu_n3 = edu_n2*(co_edu+0.03*randn(1,1));
soc_n3 = soc_n2*(co_soc+0.03*randn(1,1));
famhis_n3 = famhis_n2*co_famhis;
ES_n3 = w1*rela_n3+w2*edu_n3+w3*soc_n3+w4*famhis_n3;

drug16to19 = [drug_now,drug_n1,drug_n2,drug_n3];
ES16to19 = [ES_now,ES_n1,ES_n2,ES_n3];

for i=1:4
    druggg(:,i) = sum(drug16to19(:,i))/377;
    ESSSS(:,i) = sum(ES16to19(:,i))/377;
end

T = 2010:2019;
druggg = [0.024968,0.025931,0.028959,0.035892,0.041583,0.03181,druggg];

```

```

figure(1);
plot(T(1:8),druggg(1:8),'xr-');
hold on
plot(T(8:10),druggg(8:10),'ob-');
xlabel('Year');
ylabel('Drug Index');
title('Average county's Drug Index - Year');
rela = xlsread('1.xlsx',1,'B2:G378');
edu = xlsread('2.xlsx',1,'B2:G378');
soc = xlsread('3.xlsx',1,'B2:G378');
famhis = xlsread('4.xlsx',1,'B2:G378');
ES2 = w1*rela+w2*edu+w3*soc+w4*famhis;
for i=1:6
    ES3(i) = sum(ES2(:,i))/377;
end
ES = [ES3,ESSSS];
figure(2);
plot(T(1:8),ES(1:8),'xr-');
hold on
plot(T(8:10),ES(8:10),'ob-');
xlabel('Year');
ylabel('Social Index');
title('Average county's Social Index - Year');

```

Appendix 5: PCA_result

file index"Relation"

file index "Education"

file index "Society"

