Research on Opioid Propagation in the United States Based on Artificial Neural Network

Zexi Lin*, Bin Wang and Dongze Wu

School of Information Science & Technology, Dalian Maritime University, Dalian, Liaoning Province, 116026, China
*Corresponding author’s e-mail: flash@dlmu.edu.cn

Abstract. Nowadays the United State is going through a national crisis regarding the use of synthetic and non-synthetic opioids. In this paper, firstly, we will describe the spread and characteristic of opioid cases based on Multi-Layer Perceptron (MLP). Then find the possible place where opioid might have started with and predict the future situation. Secondly, we select some most important evaluation indexes by regression analysis. Thirdly, improve the model by using Principal Component Analysis (PCA) to have dimension reduction of the thirteen most important evaluation indexes and construct the comprehensive effect factor as model’s inputs to describe the influence of location and the U.S. Census socio-economic conditions.

1. Introduction
Nowadays, the opioids usually be used in health care field to reduce the pain for patients with critically ill. However, there are a lot of people selling the opioids to profiteer as well. The more important is that no matter using for the treatment or recreational purpose, the opioids may cause the negative health effects even some serious illness. Federal organizations such as the Centers for Disease Control are trying to prevent the negative health effects causing by the opioids. However, it is not easy, and the U.S. Drug Enforcement Administration just simply enforces existing laws. There are implications for important sectors of the economy of the United States too.

Now there are some data containing drug identification counts in years 2010-2017 for narcotic analgesics (synthetic opioids) and heroin in each of the counties from these five states (the data is provided by DEA/National Forensic Laboratory Information System (NFLIS)). Based on the assumption that the data are correct as provided, we try to use data science to describe the opioids crisis, find the reasons of opioids and give some suggestions to solve the crisis.

2. Description of opioid spreading

2.1. Data preprocessing
Before building the mathematical model to describe the spread of opioid, we have to preprocess data to lay the foundation for follow-up work.

2.1.1. Data reliability analysis. Now we do the reliability analysis on the data after data screening. Firstly, we analysis the 8 variables which include county total count of all substances identified (Total Drug Reports County) from 2010 to 2017. The Cronbach’s Alpha is 0.988 > 0.9 and the details show in Table 1. After that, do the reliability analysis on 10 variables which include latitude and longitude of each county and their total drug reports. The result show that Cronbach’s Alpha is 0.961 > 0.9 and the
details show in Table 2. It can prove that the modified scale has good reliability and can be used for modeling.

| Table 1. Reliability Statistics | Table 2. Reliability Statistics |
|---------------------------------|---------------------------------|
| Cronbach’s Alpha               | Cronbach’s Alpha               |
| 0.988                           | 0.999                           |
| Standardized Items              | Standardized Items              |
| N of items                      | N of items                      |
| 8                               | 10                              |

2.1.2. Data normalization. Due to the total drug report of county data (TDRC) range is too huge to fit directly, so we do the data normalization to turn the data range to [0, 1] for the further analysis. Here using the Max-Min method to normalize raw data [1].

Data set is

\[ y = \{y_1, y_2, ..., y_n\} \]

Normalization formula is

\[ y' = \frac{y - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}} \]

2.2. Data visualization:

We get the latitude and longitude of all the individual counties, and draw eight figures to show the relationship between locations (the location information comes from OpenStreetMap) and the total count of drug report of county (TDRC). The figures are as followed.

Figure 1. The relationship between locations and the total drug report of county (TDRC)

From the figure 1 we can find that the counties around the county with high drug abuse level tend to have high drug abuse level. Therefore, it is easy to get the conclusion that the count of drug reports has strong connection with county’s location, which just prove that our idea about location is right. Meanwhile the count of drug reports also has strong connection with county’s drug abuse situation last years.

According to the conclusion in preliminary analysis, we can know that the count of drug reports has strong connection with county’s location, but how is the location influence drug abuse situation? We all know that the method of opioid spread is transport it from one place to others therefore the closer two counties are, the easier opioid spread between them. Hence, geographical location information can turn into the distance of counties nearby combine with the drug abuse situation of counties nearby.

2.3. Description of Model

The basic variable definition is in Table 3 and the constant n is the number of counties.

| Table 3. variable definition |
|-----------------------------|
| Variable | Definition                           |
| t        | Year the evidence was received for analysis |
| i, j     | County’s index                       |
| d (i, j) | Distance between county i and county j |
After the analysis, we know that county’s total count of all substances identified in specific year depend on the nearby counties’ distance and these counties’ total count of all substances identified. Meanwhile it is also related with time. Therefore, we construct variable \( s \) to describe this relationship. And the definition of variable \( s \) are as following:

\[
s(i,j,t) = \frac{TDRC(j,t)}{d(i,j)}
\]

We can also know from the analysis that the total count of drug report of county \( i \) in year \( t \) is related with the total count of drug report of county \( i \) in year \( t-1 \). In other word TDRC\((i, t)\) is related with TDRC\((i, t-1)\).

The goal of our model is that try to find the relationship between TDRC\((i, t)\) and TDRC\((i, t-1), s(i, j, t-1) \) from the data set. Then we can predict next year situation according to last year situation. After that we predict the year after according to the next year situation we predicted. The further future situation can be predicted in the same manner.

In the similar way, if we train the model reversely, we will get the model which can estimate the last year situation according to this year. Then we use this model to estimate the situation in all beginning (the count of specific opioid close to 0) and the county which has most count of specific drug is the place this specific opioid may have started in most possible.

The model’s details are as following.

**Input variable:**

\( TDRC(i, t-1), s(i,j, t-1) \quad (j = 1, 2, ..., n and j \neq i) \)

**Output variable:**

\( TDRC(i, t) \)

In order to find relationship between input variable and output variable, we use Multi-Layer Perceptron (MLP, [2, 3]) to fit training data and use Adam optimizer[4]. Meanwhile, use logistic function as activation function.

The predict model’s neural network structure show in Figure 2. The input layer’s dimension is \( n \). The output layer’s dimension is 1. The hidden layer has two layers that has 3000 units and 30 units individually.

The estimate model’s neural network structure show in Figure 3. The input layer’s dimension is \( n \). The output layer’s dimension is 1. The hidden layer has two layers that has 3000 units and 30 units individually.

![Figure 2. Predict model’s network](image)

![Figure 3. Estimate model’s network](image)

**2.4. Result**

**2.4.1. Model evaluation.** Here dividing data into training data and testing data (70% training data, 30% testing data). Training data use for training model and testing data use for evaluate model.

And using the coefficient of determination \( R^2 \) of the prediction as the score of model evaluation[5].
The definition of $R^2$ is as following:

$$R^2 = 1 - \frac{\sum (y_{\text{real}} - y_{\text{predict}})^2}{\sum (y_{\text{real}} - \bar{y})^2}$$

The best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse) and a constant model that always predicts the expected value of $y$, disregarding the input features, would get a $R^2$ score of 0.0.

The score of predict model evaluation is 0.9485891794217512. And the detail of our evaluation result show in Figure 4.

![Figure 4. Predict model’s evaluation result](image)

![Figure 5. Estimate model’s evaluation result](image)

The score of estimate model evaluation is 0.9326503141248169. And the detail of our evaluation result show in Figure 5.

2.4.2. Model predict future. After evaluation we use all data to train the model. Then we use the trained model to predict the next 4 years (2018-2021) situation in all counties. After that we visualize the predict data for better demonstrate.

The result show in Figure 6.

![Figure 6. Predict result](image)

From figure 6 and the predict data with processed we can know following information.

- Most of counties’ total count of drug report will increase.
- The three counties (HAMILTON, OH; CUYAHOGA, OH; PHILADELPHIA, PA) with the most total count of drug report tend to be steady.
- The sum of 4 states’ total count of drug report in 2021 will be 3 times more than the count in 2017.

2.4.3. Model estimate past. After evaluation we use all data to train the model. Then we use the trained model to estimate the 10 years situation before 2010 (2000-2009) in all counties. After that we visualize the predict data for better demonstrate. And we show the 4 years result most representative in Figure 7.
3. Model modified
To put forward some suggestions to solve the opioids crisis, we have to study the relationship between trends-in-use and any of the U.S. Census socio-economic. We could modify the model in part 1 according to the relationship.

3.1. Analysis of main evaluation index
In this part, having a few of evaluation index, we will select some evaluation indexes that have a huge influence on the use or trends-in-use to describe the total count of drug report of county. In fact, this question is a regression problem. We can use regression analysis to predict the importance of evaluation indexes.

We set up a threshold to select the evaluation indexes that we have. If the importance of the evaluation bigger than the threshold, we think the evaluation index have a huge influence on the use or trend-in-use, else we think the evaluation index is not important enough to the use or trend-in-use. We set up the threshold as 0.03.

As Figure 8, in the following discussion, we only consider the evaluation indexes above the dotted line. The meaning and the importance of the thirteen evaluation indexes are in appendix.

Using the data of the thirteen evaluation indexes, we could calculate the eigenvalues by Principal Component Analysis (PCA, [6]). The results are as following.

Figure 7. Estimate result

From figure 7 and the predict data with processed we can know that most of counties’ total count of drug report is increasing by time. And in the earliest time we have estimated (2000), most of counties’ count of drug report are in very low level. However, PHILADELPHIA, PA still has a very high drug abuse level. Therefore PHILADELPHIA, PA is the count these drugs have started most probably.
Figure 9. Total Variance Explained

As Figure 9, we select the two principal components whose eigenvalues is bigger than 1. The variance contribution rates of the two principal components are 72.732% and 18.458%. Then we could calculate the level of attraction to opioid in each of the individual counties.

3.2. The influence of model
The basic variable definition is in Table 4.

| Variable       | Definition                                |
|----------------|-------------------------------------------|
| t              | Year the evidence was received for analysis |
| i, j           | County’s index                            |
| s(i, j, t)     | Location effect factor                    |
| TDRC(i, t)     | Total count of drug report of county i in year t |
| p(i,j,t)       | Comprehensive effect factor               |
| u(i, q)        | The level of attraction to opioid of county i in qth year |

The constant definition is in Table 5.

| Constant | Definition |
|----------|------------|
| k        | The number of years |
| n        | The number of counties |

Now, we consider the two principal components’ influence of model in Part 1. The influence of the U.S. Census socio-economic is expressed in the two principal components. We could calculate a level of attraction to opioid in each of the individual counties every year.

As for every individual county, we could find the average of the level of attraction to opioid of each year in every individual county:

$$
\bar{u}(i) = \frac{\sum_{q=1}^{k} u(i, q)}{k}
$$

Then, we could build a new factor, the comprehensive effect factor to describe the reasons of total count of drug report of county:
As the same as Part1, we could build the Multi-Layer Perceptron:

**Input variable:**
\[ TDRC(i, t-1), p(i, j, t-1) \quad (j = 1, 2, ..., n \text{ and } j \neq i) \]

**Output variable:**
\[ TDRC(i, t) \]

### 3.3. Result and Discussion

We use the method same as Part1 to test the Multi-Layer Perceptron.

Score of model evaluation is 0.9291602799665424. And the detail of our evaluation result show in figure 10.

![Figure 10. Evaluation result](image)

In addition, because we do not use the data of 2017 to predict in this part. So we can use the model to predict the total count of drug report of county of 2017 and compare the result with the true condition in 2017. The detail of our result show in Figure 11 and Figure 12.

![Figure 11. Predict 2017](image) ![Figure 12. True 2017](image)

We could find that the prediction data is similar to the true data. Therefore, the model could describe how the comprehensive effect factor have an influence on the total count of drug report of county well.

### 4. Conclusion

In this paper, we presented a model to research opioid propagation in the United States. By using Multi-Layer Perceptron in a unique way to predict the future situation and estimate the past situation. Then, we can find the source of opioid propagation and know the future situation.

According to our estimate model, the source of opioid propagation is PHILADELPHIA, PA. Moreover, according to the predict model the count of drug report will increase fast in the next 4 years and the 3 counties (HAMILTON, OH; CUYAHOGA, OH; PHILADELPHIA, PA) with the most total count of drug report tend to be steady.

### Appendix

The meaning and the importance of the thirteen evaluation indexes:
Reference

[1] S GOPAL, P., Kishore Kumar, S. (2015). Normalization: A Preprocessing Stage. IARJSET. 10.17148/IARJSET.2015.2305.

[2] W Gardner, M., Dorling, S.R. (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric Environment, 32: 2627-2636.

[3] Nazzal, J.M., El-Emary, I.M., Najim, S.A. (2008). Multilayer Perceptron Neural Network (MLPs) For Analyzing the Properties of Jordan Oil Shale. World Applied Sciences Journal, 5.

[4] Kingma, D., Ba, J. (2015) Adam: A method for stochastic optimization. In: International Conference of Learning Representation. San Diego, CA.

[5] Cameron, A., Windmeijer, F. (1996) R-Squared Measures for Count Data Regression Models With Applications to HealthCare Utilization. Journal of Business & Economic Statistics - J BUS ECON STAT. 14.

[6] Lever, J., Krzywinski, M., Altman, N. (2017) Principal component analysis. Nature Methods, 14: 641-642.