Research on population trafficking prediction based on K-means and BP neural network

Zhihong Liu[1,a], Zhongxian Zhu[1], Suxin Liu[1], Xiaoyu Han[2]

[1]Harbin Institute of Technology, Harbin, China

[2]Harbin University of Science and Technology, Weihai, China

2815806233@QQ.com

Keywords: TOPSIS method; K-MEANS; neural network; graph theory; human trafficking

Abstract: This paper mainly analyzes the issue of human trafficking. First, identify high-risk groups and prevent them in time. Using the data collected from countries around the world, the TOPSIS method was used to evaluate the risk of trafficking in countries around the world. The evaluation results were clustered by K-MEANS method and represented by different colors on the map. Second, locate the victim. Taking the United States as an example, we have further refined the location of the victims. By looking at the number of human trafficking cases in each state in 2012-2017, BP neural network was used to predict the number of human trafficking cases in each state in the next four years.

1. Introduction

There are varying degrees of human trafficking in every country, including the United States. Human trafficking is a profiteering industry that is estimated to generate more than $150 billion in illegal profits worldwide each year. In fact, human trafficking is one of the most profitable organized criminal activities in the world, second only to the drug trade.

In order to combat human trafficking, we need to use the information we have about victims, traffickers or buyers to break the supply chain. The specific tasks are:

(1) Identify high-risk groups

(2) Identify and locate the victims

(3) Destroy the trafficking network

2. Problem Analysis

In order to identify high-risk groups, it is necessary to identify countries with high risk of being trafficked. Therefore, we must first analyze the factors that cause national trafficking risks, such as economic development, unemployment rate, and war situation. According to the specific conditions of different countries, these factors are used to evaluate each country. According to the evaluation results, high-risk countries can be identified. Unemployed people in high-risk countries are the high-risk groups we are looking for. Positioning the victim is to determine the location of the victim. In other words, we need to find the location where the trafficking occurred. In order to determine the location of future trafficking cases, the number of existing cases can be collected, and a predictive model can be established to predict the number of cases in various regions in the future, and to predict where the most cases occur, that is, where the most likely victims are located. Police force and supervision should be strengthened.

In order to simplify the model, we only consider the impact of the four key factors of poverty rate, immigration rate, unemployment rate and number of wars on the risk of human trafficking in a country. Because of the limited data at the national level, we believe that the sales data of a region can represent General situation of all countries in the region; use local population cases to represent the severity of local trafficking.

DOI: 10.25236/mmmce.2019.124

3. Model establishment and solution

3.1 Screening for high-risk groups

To evaluate the risk of human trafficking in a country, it is first necessary to identify the factors that influence the human trafficking. We selected the four factors of poverty rate, immigration rate, unemployment rate [1] and the number of wars since 2008 [2] to evaluate the risk of human trafficking.

Poverty rate: the proportion of the population with income below $5.5 to the total population;

Immigration rate: the proportion of the population of immigrants to the total population;

Unemployment rate: the proportion of the unemployed population to the total population;

After the data of the office was sorted out, we got the value of the four influencing factors in the world. Later, we used the TOPSIS method to evaluate the risk of human trafficking in countries around the world.

Let the multi-attribute decision-making scheme set A= { a_1, a_2,... , a_m } , and measure the attribute vector of the scheme as X= { x_1, x_2,..., x_n }, then in the program set A The vector consisting of n attribute values of each scheme a_i(i=1,..., m) is X_i= { x_{i1}, x_{i2},..., x_{in} }, which is a point in the n-dimensional space and can be unique Ground representation scheme a_i

Algorithm steps:

Step one, using the vector planning method to obtain the normative decision matrix. A decision matrix with multiple attribute decision problems $X_i=(x_{ij})_{mxn}$, normalized decision matrix $Y_i=(y_{ij})_{mxn}$,

Step two, forming a weighted norm matrix $Z_i=(z_{ij})_{mxn}$.

Step three , Determine the positive ideal solution Z^* and the negative ideal solution Z^0 .

Step Four , calculate the distance from each scheme to the positive ideal solution and the negative ideal solution.

$$d_i^* = \sqrt{\sum_{j=1}^{n}(z_{ij} - z_j^*)^2}, \qquad i = 1,2,...,m$$

Step five : Calculate the queued index value of each scheme.

$$C_{\mathrm{i}}^* = d_i^0 /(d_i^0 + d_{\mathrm{j}}^*), \qquad i = 1,2,...,m$$

After the above steps, we obtained the weight vector *w*=(0.31487, 0.083779, 0.44584, 0.15551)^T, which consists of four factors: poverty rate, immigration rate, unemployment rate and number of wars since 2008. Using the TOPSIS method, we have obtained a comprehensive evaluation index of human trafficking risks in countries around the world. We then clustered the comprehensive evaluation index using K-MEANS cluster analysis and reflected each category on the map with different color depths. The darker the color, the greater the risk that the country's population will be trafficked. As shown in Figure 1.

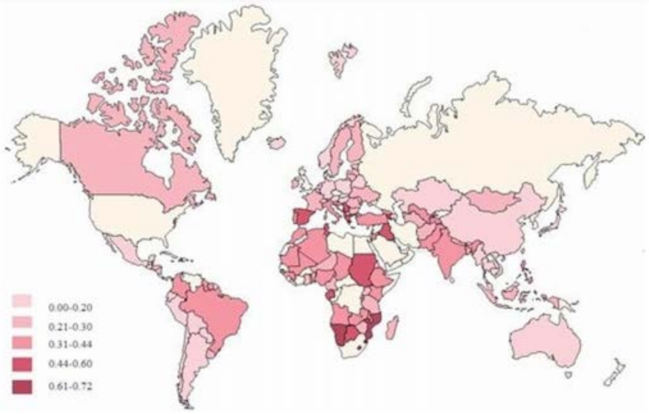


Figure 1 Human trafficking risk map

3.2 Model checking

To verify the correctness of the model, we found the number of victims in different regions at UNODC and used blue dots to mark areas with more victims on the map, as shown in Figure 2. The image shows that the results of the model are consistent with the actual situation, indicating the correctness of the model.

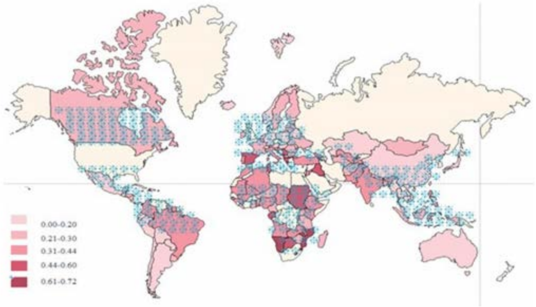


Figure 2 Trafficking risk assessment and victim concentration area comparison

3.3 Identify and locate victims

The location of the victimized population is the place where the human trafficking case takes place. Therefore, as long as the human trafficking case is predicted, the area where the human trafficking is most likely to occur in the future can be determined. A survey of identified areas reveals the affected population. We have collected the number of trafficking cases in various countries on the UNODC website [4]. According to the data, the number of victims found in the United States is the highest and is about twice that of the second-ranking Netherlands. Therefore, we will further refine the example of the United States with the most victims. The number of victims in various states in the United States was predicted using BP neural networks.

● BP neural network

The generation of BP neural network is attributed to the acquisition of BP algorithm. It has an input layer, an output layer, and one or more hidden layers. There is no correlation between the neurons in the same layer, and the neurons in the different layers are connected forward. According to the complexity of the object, the appropriate network structure can be selected to realize the mapping of any nonlinear function from input space to output space [5].

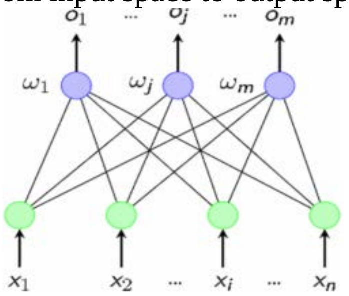


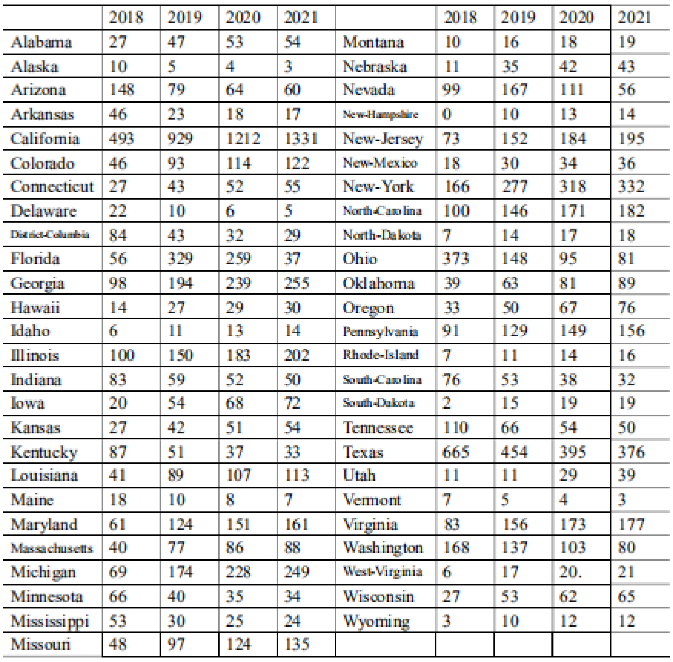
Figure 3 BP neural network

BP neural network is a multi-layer feedforward network trained by error back propagation (referred to as error back propagation). Its algorithm is called BP algorithm. Its basic idea is gradient descent method, which uses gradient search technology to make the network The error mean square error between the actual output value and the expected output value is minimal.

The basic BP algorithm includes two processes: forward propagation of the signal and back propagation of the error. That is, when calculating the error output, the direction from the input to the output is performed, and the adjustment weight and the threshold are performed from the output to the input direction. During forward propagation, the input signal acts on the output node through the hidden layer and undergoes nonlinear transformation to produce an output signal. If the actual output does not match the expected output, it is transferred to the backpropagation process of the error. The error back-transmission is to pass the output error back to the input layer through the hidden layer, and distribute the error to all the units in each layer, so that the error signal obtained from each layer is used as the basis for adjusting the weight of each unit. By adjusting the connection strength between the input node and the hidden layer node and the connection strength and threshold of the hidden layer node and the output node, the error is decreased along the gradient direction, and after repeated learning training, the network parameters (weights and thresholds corresponding to the minimum error) are determined.), the training will stop. At this time, the trained neural network can directly process the non-linear conversion information with the smallest output error for the input information of similar samples.

The outstanding advantage of BP neural network is that it has strong nonlinear mapping capability and flexible network structure. The number of intermediate layers in the network and the number of neurons in each layer can be arbitrarily set according to specific conditions, and their performance differs depending on the structure. However, BP neural network also has some defects. For example, it is easy to fall into local minimum values. There is no corresponding theoretical guidance for the selection of network layers and number of neurons.

- Neural network predictionk



	2018	2019	2020	2021		2018	2019	2020	2021
Alabama	27	47	53	54	Montana	10	16	18	19
Alaska	10	5	4	3	Nebraska	11	35	42	43
Arizona	148	79	64	60	Nevada	99	167	111	56
Arkansas	46	23	18	17	New-Hampshire	0	10	13	14
California	493	929	1212	1331	New-Jersey	73	152	184	195
Colorado	46	93	114	122	New-Mexico	18	30	34	36
Connecticut	27	43	52	55	New-York	166	277	318	332
Delaware	22	10	6	5	North-Carolina	100	146	171	182
District-Columbia	84	43	32	29	North-Dakota	7	14	17	18
Florida	56	329	259	37	Ohio	373	148	95	81
Georgia	98	194	239	255	Oklahoma	39	63	81	89
Hawaii	14	27	29	30	Oregon	33	50	67	76
Idaho	6	11	13	14	Pennsylvania	91	129	149	156
Illinois	100	150	183	202	Rhode-Island	7	11	14	16
Indiana	83	59	52	50	South-Carolina	76	53	38	32
Iowa	20	54	68	72	South-Dakota	2	15	19	19
Kansas	27	42	51	54	Tennessee	110	66	54	50
Kentucky	87	51	37	33	Texas	665	454	395	376
Louisiana	41	89	107	113	Utah	11	11	29	39
Maine	18	10	8	7	Vermont	7	5	4	3
Maryland	61	124	151	161	Virginia	83	156	173	177
Massachusetts	40	77	86	88	Washington	168	137	103	80
Michigan	69	174	228	249	West-Virginia	6	17	20.	21
Minnesota	66	40	35	34	Wisconsin	27	53	62	65
Mississippi	53	30	25	24	Wyoming	3	10	12	12
Missouri	48	97	124	135					

Figure 4 Forecast of the number of cases sold in 51 states in the next four years

We obtained the number of human trafficking cases in the US states in 2012-2017 [6] (see Appendix III for details). The year is the input, and the number of human trafficking cases per year is output, establishing a neural network with only one hidden layer. And set the number of neurons in the hidden layer to 10, and train the network with the existing data (see Appendix 4 for the MATLAB code). We used a well-trained network to predict the number of trafficking cases in 51 US states over the next four years.The results are rounded up and shown in Figure 4.

It is easy to know that in the following year, Texas, Florida, California, Ohio, Washington, New York, Arizona, Tennessee, North Carolina and Illinois are the top ten states, and the forecast is The number of cases has exceeded 100, and it should be supervised, and there is a high probability that there will be victims.

3.4 Destroy the trafficking network

Trafficking in human beings is a worldwide process of population movement, and different countries are linked by the inflow or outflow of human trafficking, forming a global network. Accurate identification of trafficking networks provides a guarantee for the effectiveness of reconnaissance operations. Destroying the important nodes of the criminal network can seriously attack criminals.

Social network analysis is a quantitative analysis method, so a formal description of social networks must be made. The social network G consists of a set of nodes N= { n_1, n_2,..., n_m } and a set of connections between a group of nodes L= { l_1, l_2,..., l_k } . From a mathematical point of view, there are two ways to describe social networks: graph theory and matrix method. Graph theory is a graphical method that uses points and lines as basic elements to represent actors and their relationships. It is more intuitive and can clearly reflect the actors and their relationships. , as shown in Figure 5. However, when the social network to be studied is large, the graphical representation will become very complicated, which is not conducive to the quantitative analysis of social relations, and the matrix method should be adopted.

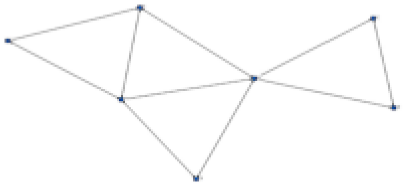


Figure 5 Schematic diagram

● Construction of human trafficking network model

There are many areas involved in human trafficking, and an area can be either the source of the victim or the destination of the transit station or victim. Therefore, there is more or less contact between the regions. From the perspective of data analysis, the human trafficking transaction between any two regions is equivalent to a simple graphic G_K consisting of a set of N_K with two nodes and a connection I_K between the two nodes. In this way, the human trafficking between the regions has the mathematical basis for social network analysis.

For the sake of research convenience, we have neglected the number of trading populations between regions and only studied the linkages between regions. We will link the various regions [6] as shown in Figure 6.

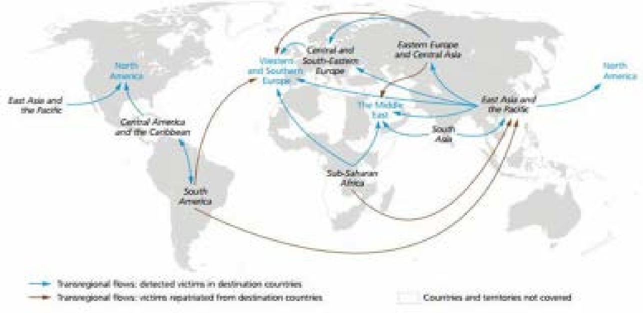

North America
East Asia and the Pacific
Central America and the Caribbean
South America
Western and Southern Europe
Central and South-Eastern Europe
Eastern Europe and Central Asia
The Middle East
South Asia
Sub-Saharan Africa
East Asia and the Pacific
North America
Transregional flows: detected victims in destination countries
Transregional flows: victims repatriated from destination countries
Countries and territories not covered

Figure 6 Inter-regional trafficking in humans

The human trafficking model network model reflects the relationship between regions and regions. It can be quantitatively analyzed from multiple research perspectives according to social network analysis methods, thus providing a theoretical basis for determining the importance of each region in the human trafficking network.

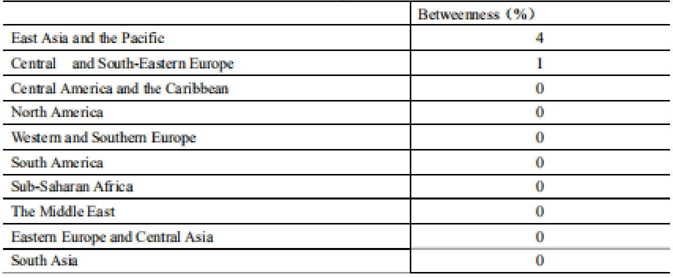
1) Identify core sales areas by point center analysis

In a social network, if an actor has a direct relationship with many other actors, the actor is at the core of the network and thus has a larger “right”. Actors at the core have more connections than other actors, which means they have more information and resources in the network and have an advantage in dealing with relationships. Under the guidance of this idea, the point centrality of a node in the network can be measured by the number of nodes in the network that are directly related to the point.

2) Identify important sales areas through intermediate central analysis

The intermediate centrality reflects the ability of an actor to act as an intermediary, that is, a node in the network is in the middle of many other two nodes, and thus the node is considered to be in an important position. The higher the mediation of a node, the more nodes that need to be contacted through it, the more information the node can obtain, and the greater its ability to manipulate resources. The basic calculation method is to calculate how many paths are connected to other nodes through which the nodes are connected. The value of the intermediate centrality is between 0 and 1. If the intermediate center of a node is 0, the node cannot control any other nodes and is at the edge of the network. If the centrality of a node is 1, it means that Points can completely control other nodes, at the most central location of the network.The central calculation results of the middleware of the human trafficking network model are shown in Figure 7.

As can be seen from figure 7, the central centrality of East Asia and the Pacific is the largest, and the central centrality of the central and southeastern regions is high, indicating that they play a vital role in transactional transmission. In summary, we can get East Asia and the Pacific region to play a pivotal role in both human trafficking destinations and human trafficking transit stations. Therefore, we believe that if the region's government can strengthen its regulatory measures, It will play a decisive role in the destruction of human trafficking networks.



	Betweenness (%)
East Asia and the Pacific	4
Central and South-Eastern Europe	1
Central America and the Caribbean	0
North America	0
Western and Southern Europe	0
South America	0
Sub-Saharan Africa	0
The Middle East	0
Eastern Europe and Central Asia	0
South Asia	0

Figure 7 Middleware centrality in each region

4. Sensitivity analysis

4.1 Sensitivity analysis of risk assessment model for trafficking

The use of the AHP model is highly subjective, especially in the process of pairwise comparisons, so sensitivity analysis becomes very important. There are three main types of general sensitivity analysis: numerical incremental analysis, probability simulation, and mathematical models. Here we use numerical incremental analysis, also known as the "one at a time" method, each time changing the value of a parameter to get a new solution and showing how the ranking changes. For the method of changing the parameter value, we use the method of selecting the largest weight and then slightly changing its weight. Use the following formula to adjust the weight.

$$W_j^{'} = \frac{1-W_p^{'}}{1-W_p} W_j^{'}$$

Using the above method, I will reduce the weight of the unemployment rate by 0.05 each time and increase other weights. The ranking change of the ranked area is shown in Figure 8.

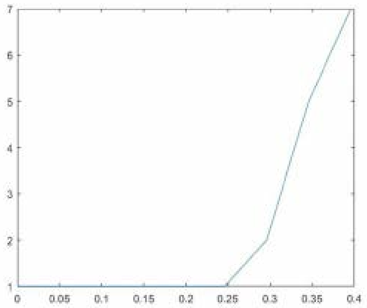


Figure 8 Changes in rankings of regions when weights are changed by 0.05

4.2 Sensitivity analysis of trafficking case prediction model

Considering that the data on trafficking cases may be biased in statistics, we have added a random disturbance to the number of crimes in various states in the United States, so that the number of cases is randomly changed from 0.8 to 0.9 times. It can be concluded that the top three states are basically always in the top three, and only one Florida has withdrawn from the top three, considering that this is due to our large random disturbance, which is between 10% and 20%. Further consideration can be given to reducing the disturbance and observing the sensitivity of the model. Overall, our predictive model is well-stabilized. The performance of the top three states in the first perturbation program in the second and third times is shown in Table 1.

Table 1. State ranking changes in the top three after joining random disturbances

Undisturbed	**Ranking**	**First disturbance**	**Ranking**	**Second disturbance**	**Ranking**	**Third disturbance**	**Ranking**
Texas	1	Texas	2	Texas	3	Texas	3
Florida	2	Florida	3	Florida	1	Florida	2
California	3	California	1	California	2	California	1

4.3 Sensitivity analysis of trafficking network model

We randomly changed the relationship matrix between regions and added a relationship. The centrality of the points in each region is shown in Table 2.

Table 2. State ranking changes in the top three after joining random disturbances

	First change	**Second change**	**Third change**
East Asia and the Pacific	14%	0%	0%
Central and South-Eastern Europe	0%	0%	0%
Western and Southern Europe	0%	0%	0%
North America	0%	0%	33%
Central America and the Caribbean	0%	0%	33%
South America	0%	0%	0%
Eastern Europe and Central Asia	0%	0%	0%
The Middle East	0%	0%	0%
South Asia	50%	0%	0%
Sub-Saharan Africa	0%	0%	0%

It is easy to know from the table that the centrality of the point is very sensitive to the relationship between the regions. Accurate regional relationships are critical to the importance of regional importance

5. Advantages and disadvantages analysis

5.1 Advantage

The TOPSIS method can sort the scenarios by their match to the target. If the judgment of the relative importance of the indicator and the judgment of the candidate's ability to satisfy the optimization goal are both accurate and effective, then the calculation result of the TOPSIS method is that there are reasonable results derived from those judgments.

The application of social network analysis methods can make the complex relationships between countries clearly expressed in numbers and easy to understand.

We compared the results of the country's evaluation with the actual situation and proved the accuracy of the evaluation model.

Sensitivity analysis was carried out and the results proved that our model has better robustness.

5.2 Disadvantage

The use of AHP's method to construct a comparison matrix has a large human subjectivity and may have a greater impact on the country's ranking.

Due to the shortcomings of the data, we are unable to accurately locate the victim population to the provincial and municipal levels, and only to the regional level. In addition, we have not been able to accurately measure the human trafficking network to the provincial and municipal level through network analysis.

References

[1] United Nations Office on Drugs and Crime. http://www.unodc.org/unodc/human-trafficking/, 2018-2-1

[2] https://wenku.baidu.com/view/5486b551d0d233d4b14e69ee.html

[3] Si Sukui. Mathematical Modeling Algorithms and Procedures. Yantai: Naval Aeronautical Engineering Institute, 2007

[4] United Nations Office on Drugs and Crime. http://www.unodc.org/unodc/human-trafficking/, TEAM# Page17 of 2018-2-1.

[5] Zhuo Jinwu. Application of MATLAB in Mathematical Modeling. Beijing: Beijing University of Aeronautics and Astronautics Press, 2011.1 18.

[6] United Nations Office on Drugs and Crime. http://www.unodc.org/unodc/human-trafficking/, 2018-2-1

[7] Zhang Chenghu, Lü Yi. Research on Customer Identification of Commercial Banks Based on Social Network Analysis[J]. Financial Forum, 2012, 17(0 8): 68-72

[8] Zhao Zeyun,Wang Xuefeng.Research on Urban Traffic Flow Prediction Based on Neural Network[J].China New Telecommunication , 2019 , 21(04) : 59.