Quantitative Light Pollution Analysis Based On K-Means++ Cluster Analysis and Neural Networks

Jiaying You *

School of Economics and Management, China University of Geosciences, Beijing, China

* Corresponding Author Email: y1225jy@163.com

Abstract. The purpose of this study is to investigate the problem of light pollution, to quantitatively analyze a wide range of light pollution and to simulate the effect of regional control strategies. First, TANGY model was introduced to transform the light pollution evaluation from simple luminous flux calculation to comprehensive evaluation. TOPSIS evaluation method optimized by entropy weight method and BP neural network optimized by genetic algorithm were used to calculate the light pollution index. Secondly, three strategies to solve light pollution are proposed, and the implementation model is constructed by cluster analysis and regression modeling. The optimal strategy is determined by gradient enhanced decision tree algorithm, and the prediction model of light pollution intervention strategy is constructed. This study provides an effective reference for the formulation of light pollution prevention strategies in different regions, and is expected to have practical significance in the field of environmental protection.

Keywords: K-means++ cluster analysis; entropy weight method; TOPSIS evaluation method; neural network.

1. Introduction

Due to the impact of climate change, environmental pollution has attracted wide attention in recent years. Light pollution is another new type of environmental pollution after air pollution, water pollution, soil pollution and noise pollution. Light pollution can lead to wasted energy and harm humans and other living things. Although light pollution has not been included in the prevention and control of environmental pollution, its impact is absolutely obvious and serious.

In this paper, economic, population, wildlife and observatory data in recent years are obtained based on the "China Statistical Yearbook" issued by the National Bureau of Statistics of China, and night remote sensing data are obtained based on the Chinese Luojia-1 luminous remote sensing satellite. Based on the quantitative analysis of the relevant data, a quantitative comprehensive model for evaluating the impact of light pollution is developed. On this basis, a variety of feasible light pollution intervention strategies are proposed, and the implementation effects of the strategies are analyzed.

2. TANGY Quantitative Analysis Model of Light Pollution

The TANGY model consists of two parts, which are solved for the potential light pollution capacity index and the potential being light polluted capacity index. The model of the Entropy method [1-2] and the model of the TOPSIS method [3] are used for the solution of potential light pollution capacity index. The model of the neural network algorithm optimized by the genetic algorithm is used for the solution of the potential light pollution capacity and the potential being light polluted capacity. The difference between the two is interpreted as the light pollution index.

2.1. Model Construction

2.1.1 The model of the entropy method

Firstly, the different factors are normalized. The purpose of this step is to process all the different types of factors into very large factors. Our factors contain very large, very small and intermediate types of factors, where economic level and proportion of tertiary industries are very large factors,

light brightness area, mean of light brightness and maximum light brightness are very small factors, and population density is an intermediate type of factor.

For the very large factors, the data do not need to be normalized. For very small factors, since all our data are positive, the inverse is taken directly:

$$X_i = \frac{1}{X_i}, (i = 1,2,3 \dots) \tag{1}$$

For intermediate type factors, the best value x_{best} is the first thing to be determined, and x_{new} is the very large indicator after forwarding, the specific process is as follows:

$$M = max\{|x_i - x_{best}|\} \tag{2}$$

$$x_{new} = 1 - \frac{|x_i - x_{best}|}{M} \tag{3}$$

In order to remove the differences between factors or differences due to different scales, we standardize all indicators. The matrix X is developed as follows:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \tag{4}$$

The matrix after standardization is noted as Z. The normalization formula for each element in this matrix is as follows:

$$z_{ij} = \frac{x_{ij} - \min\{x_{1j}, x_{2j}, \dots, x_{ij}\}}{\max\{x_{1j}, x_{2j}, \dots, x_{ij}\} - \min\{x_{1j}, x_{2j}, \dots, x_{ij}\}} \tag{5}$$

After that, the weight of each element in the matrix is calculated and the sum of the weight of all elements is 1. The calculated matrix is called the probability matrix and is denoted as P:

$$p_{ij} = \frac{z_{ij}}{\sum_{i=1}^{n} z_{ij}} \tag{6}$$

Next, calculate the information entropy of each factor, noted as e_j. The formula is as follows:

$$e_j = -\frac{1}{lnn} \sum_{i=1}^{n} p_{ij}\, ln(p_{ij}), (j = 1,2, \dots, m) \tag{7}$$

Finally, the weights derived from the entropy weighting method are calculated, d_j is the information utility value of each factor and w_j is the weight of each factor:

$$d_j = 1 - e_j, (j = 1,2, \dots, m) \tag{8}$$

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j}, (j = 1,2, \dots, m) \tag{9}$$

2.1.2 The model of the TOPSIS method

Define the best and worst scenarios Z^+ and Z^- among all schemes after normalization first [4-6]:

$$Z^+ = (max\{z_{11}, z_{21}, \dots, z_{n1}\}, max\{z_{21}, z_{22}, \dots, z_{n2}\}, \dots, max\{z_{1m}, z_{2m}, \dots, z_{nm}\}) \tag{10}$$

$$Z^- = (min\{z_{11}, z_{21}, \dots, z_{n1}\}, min\{z_{21}, z_{22}, \dots, z_{n2}\}, \dots, min\{z_{1m}, z_{2m}, \dots, z_{nm}\}) \tag{11}$$

And then define the distance D^+ between the I evaluation scheme and the best scheme:

$$D_i^+ = \sqrt{\sum_{j=1}^{m} w_i \left(Z_j^+ - z_{ij}\right)^2} \tag{12}$$

The distance D^- between the $i-th$ evaluation scheme and the worst scheme:

$$D_i^- = \sqrt{\sum_{j=1}^{m} w_i \left(Z_j^- - z_{ij}\right)^2} \tag{13}$$

Finally, the score of the i-evaluation scheme is calculated:

$$S_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{14}$$

S_i is the potential light pollution capacity index mentioned above, which shows the synergistic relationship between regional economic development and light intensity.

2.1.3 The model of the neural network algorithm optimized by genetic algorithm

Different areas are different in how sensitive they are to light pollution. Thus, we need to classify these different areas according to the levels of risk according to the factors of different types of areas. In order to make the classification not affected by subjective considerations. However, many values of traditional neural networks [7] are difficult to find the best, so we use genetic algorithms to optimize the neural networks [8-9]. Finding the best values by multiple iterations. Then the neural network is trained based on these values. The specific operation flow is shown in the Fig.1.

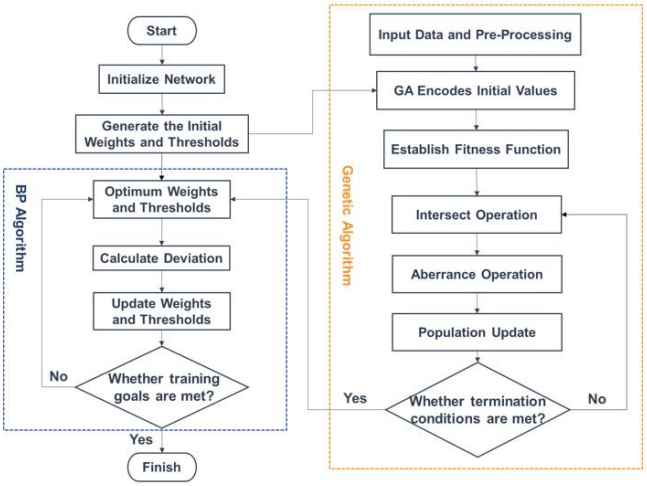

Fig. 1 The pattern of neural network algorithm optimized by genetic algorithm.

The encoding is performed first. Consider a neural network with input node i, implicit node j, and output node k, where the weight matrix from the input layer to the implicit layer is W, the threshold matrix of the implicit layer is γ, the weight matrix from the implicit layer to the output layer is V, and the threshold matrix of the output layer is h.

Next, set the initial population. The population size is set to 50 and let the individuals within the population be generated randomly. And define the fitness function.

$$F = \frac{1}{\sum_{K=1}^{NO}\sum_{i=1}^{Ne}|y_k^i - a_k^i|} \tag{15}$$

Where NO represents the number of training set samples, Ne represents the number of output nodes, y_k^i represents the i-th predicted output obtained by the $k-th$ sample through the neural network, and a_k^i represents the corresponding actual output.

The selection method used in this paper is the roulette wheel selection method. Populations of the same size N are generated. The average fitness of the selected individuals is high. But there will be duplicate individuals, and the crossover of duplicate individuals is meaningless. Thus, duplicate individuals will be removed during the selection process.

The crossover method used in this paper is the real number crossover method. Suppose there are two chromosomes A and B of length l. Then:

$$a_k' = (1-r)a_k + rb_k \tag{16}$$

$$b_k' = rb_k + (1-r)a_k \tag{17}$$

Where r is a random number between [0,1] and $k = 1, 2, \ldots, l$.The advantage of this method is that the hybridized genes are a convex combination of the parent's gene values, thus, fully combining

the genes of parents. r is introduced to increase the diversity of individuals, which can enhance the local search ability of the algorithm and is beneficial for the algorithm to develop the optimal fusion ratio during reiteration.

The mutation method used in this paper is parenchymal variation. When the mutation operator mutates the $k-th$ gene of an individual A with a certain probability, the mutation operation is:

$$a_k' = \begin{cases} a_k + (a_{max} - a_k) * r_1 \left(1 - \frac{G}{G_{max}}\right)^2 & r_2 > 0.5 \\ a_k + (a_{min} - a_k) * r_1 \left(1 - \frac{G}{G_{max}}\right)^2 & r_2 \leq 0.5 \end{cases} \tag{18}$$

Where a_{max} and a_{min} are the upper and lower bounds of gene values, G represents the current iteration number, G_{max} represents the maximum iteration number, and r_1 and r_2 are the random numbers between [0,1]. The random number r_1 affects the level of variation. $r_2 > 0.5$ moves the gene value towards the maximum value, $r_2 \leq 0.5$ moves the gene towards the minimum value, ensuring that the gene value increases or decreases with equal probability, and the upper and lower bounds of the gene ensure that the gene value does not vary too much.

Finally, a termination condition is set for the iterations, and the algorithm terminates when the adaptation level of the optimal individual and the adaptation level of the population are no longer increasing.

Next, the neural network was constructed using the data obtained from the previous optimization. First, the structure of the neural network was determined. The input parameters included the number of observatories in the region, the total number of traffic accidents and the area of forested land. We hope to get the potential being light polluted capacity index through the neural network, so there is one output parameter of the neural network. According to the following experimental formula, we set the number of neurons in the hidden layer to four.

$$h = \sqrt{m+n} + a \tag{19}$$

2.2. Model Solving and Analysis

Four cities were selected as inputs to the model, which are Shanghai, Guangdong, Hunan and Tibet. These four cities represent the urban community, suburban community, rural community, and protected land. Shanghai is the most economically developed region in China, so it represents the urban community. parts of Guangdong are more developed, but overall, less developed than Shanghai, so it represents the suburban community. Hunan is one of the major cities in China, but has a large farming area, so it represents the rural community. Tibet is one of the least population-density areas in the world, and 90% of its area is a wildlife reserve, so it represents protected land.

Firstly, the weights of different factors are calculated by the model of entropy method. The factors we selected are economic level, proportion of tertiary industries, light brightness area, mean of light brightness, maximum light brightness and population density. The positive treatment of these factors has been mentioned in the previous section. After the calculation of entropy weighting method, the weights of these factors are shown in the Table 1.

Table 1. The weight of factors

Factor name	Weight
economic level	15.89012
proportion of tertiary industries	13.13101
light brightness area	24.01676
mean of light brightness	10.03436
maximum light brightness	10.02080
population density	26.90696

We then enter this weight and data into the TOPSIS model we build. The potential light pollution capacity index of the four cities was obtained.

Next, the model of the neural network algorithm optimized by the genetic algorithm was used to solve for the potential being light polluted capacity index. After the genetic algorithm was performed, the optimal parameters were input into the neural network. The results of the evaluation of the model of the neural network after the training and test sets are shown in the Table 2.

Table 2. Neural network model evaluation results

Assemblies	MSE	RMSE	MAE	MAPE	R^2
Training Assemblies	0.041	0.203	0.164	0.415	0.534
Test Assemblies	0.035	0.187	0.159	0.364	0.435

MSE is mean square error, RMSE is root mean square error, MAE is mean absolute error, MAPE is Mean absolute percentage error, the smaller these four values are, the more accurate the model is. R^2 represents the fit coefficient, the closer this number is to 1, the better the fit is. Our model fit is within the acceptable range.

Finally, the data from the four previously selected areas were input to the model to derive the potential being light polluted capacity index for these four cities. And the light pollution metric obtained by using the potential being light polluted capacity index minus the potential light pollution capacity index is shown in the Table 3.

Table 3. The light pollution metric of the four cities in China

City	light pollution metric
Shanghai, China	-0.4432
Guangdong, China	-0.3691
Hunan, China	0.1460
Tibet, China	0.6804

Light pollution metric has a value range of [-1,1]. In this range, the closer to -1, the more the city is polluted by light; the closer to 1, the less the city is polluted by light. It is clear from our results that light pollution is more serious in Shanghai and Guangdong, while Hunan and Tibet suffer from less light pollution, and Tibet seems not even suffer from light pollution.

2.3. Sensitivity Analysis

We analyze the factors that are more difficult to determine in the TANGY model. The more difficult factors to determine in the TANGY model are the proportion of tertiary industries and the area of forested land, and then the sensitivity of these two factors is tested. These data were scaled down and brought back to the original model for testing. The results of the graphs with this data are shown in Fig.2.

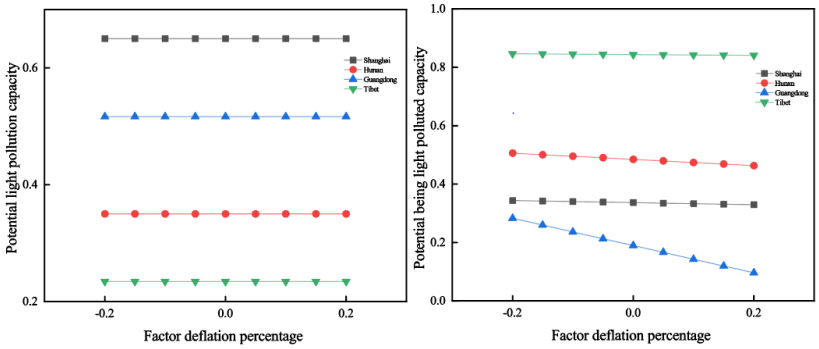

Fig. 2 Results of sensitivity test

As can be seen from the figure, the model for calculating the potential light pollution capacity is extremely stable. Because this model is a traditional evaluation type model, there is no uncertainty in the process of calculation. However, the model for calculating the potential being light polluted capacity is affected by changes in the values of the factors. In the Fig. 2, it can be seen that different regions have different sensitivities when using the model for prediction. The region with the highest sensitivity is Guangdong, and the region with the lowest sensitivity is Tibet. However, even for Guangdong, which has the highest sensitivity, the potential being light polluted capacity index changes by 12% when the factor is changed by 5%. This change is acceptable.

3. The Model of Light Pollution Intervention Strategy

3.1. Model Construction

First, we need to identify three strategies to reduce light pollution. The three strategies are: reduce the lighting time, reduce the light threshold, and reduce the light exposure range. These three strategies affect different factors in relative terms. And K-means++ cluster analysis was performed for the factors influenced by these three strategies, and the method of analysis is shown below.

Firstly, three centers of mass should be determined. A sample point c1 is randomly selected from the sample set X as the first cluster center of mass, and the distance $d(x)$ from the other sample points x to the nearest cluster center is calculated. The formula is as follows:

$$d(x) = \sqrt{\sum_{k=1}^{m}\left(x_{ik} - x_{jk}\right)^2} \tag{20}$$

A new sample point is subsequently selected to be added to the set of cluster centroids using the probability method. Where the larger the distance value $d(x)$, the higher the probability of being selected. The specific probability calculation formula is as follows:

$$P(x) = \frac{d(x)^2}{\sum_{\in x} d(x)^2} \tag{21}$$

Repeat the previous two steps to determine the center locations of the three clusters we need. Then perform the K-means calculation based on these three centers. The calculation is done by using (20) to calculate the distance of each remaining sample point with each center of mass and assigning it to the cluster where the center of mass with the smallest interphase distance is located. Then the distances of the sample points in each cluster with respect to the cluster center of mass are recalculated again to determine the new center of mass. This calculation process is repeated until the center of mass no longer changes or the maximum number of iterations is reached.

Once we have this result, we can calculate the relationship between all these factors on the light pollution level by least squares regression. Firstly, all the factors are normalized to the matrix Z. The matrix takes the following form:

$$Z = \begin{bmatrix} z_{11} & \cdots & z_{i1} \\ \vdots & \ddots & \vdots \\ z_{ij} & \cdots & z_{ij} \end{bmatrix} \tag{22}$$

Prediction models were then developed:

$$y^{(j)} = w_0 + w_1 x_1^{(j)} + w_2 x_2^{(j)} + \cdots + w_i x_i^{(j)} \tag{23}$$

The accuracy of the prediction is measured by calculating the mean square error between the predicted and true values of the light pollution metric for each sample. The specific equation is as follows:

$$J_{(j)} = \frac{1}{2i}\sum_{i=1}^{i}\left(y^{(t)} - h_j\left(x^{(j)}\right)\right)^2 \tag{24}$$

Where $J_{(j)}$ represents the error between the predicted and true values of the sample, $h_j(x^{(j)})$ represents the true value of the corresponding sample, and $y^{(t)}$ represents the predicted value of the corresponding sample. When $J_{(j)}$ takes the minimum value, the error is the lowest, which means the more accurate the model is. When taken to the extreme value, the partial derivative of $J_{(j)}$ is 0, and the parameters corresponding to the factors are solved:

$$\frac{\delta J}{\delta j} = \frac{1}{i} Z^T (Z_j - y) \tag{25}$$

Usually, policies are not developed with one factor in mind, but rather multiple perspectives. Thus, we need to determine how strong the three policies should be for a particular region. However, using traditional traversal algorithms can make this process extremely long. Therefore, we use the gradient regression tree algorithm [10-11] to optimize this process. At this point, the model of light pollution intervention strategy is completed.

3.2. Model Solving and Analysis

12 statistical factors are selected as the input of the model, and the 12 influencing factors are shown in Table 4. Firstly, through K-Means ++ cluster analysis, three main influencing factors of light pollution intervention strategies were obtained. Next, the least squares regression was carried out to determine the coefficients of each factor, and the results were shown in Table 4.

Table 4. The coefficient of the 12 factors

Factors	Coefficient
Economic level	-3.6959886853880115e-8
Proportion of tertiary industries	-3.1505417541035e-10
Light brightness area	-8.607144820710941e-10
Mean of light brightness	-1.3991181625694836e-9
Maximum light brightness	-7.446408678447474e-10
Population density	-1.261829483514949e-9
Number of observatories in the region	0
The total number of traffic accidents	7.964841890604137e-8
Area of forested land	1.3218751010896492e-10
Budling construction area	0.000001197511096911006
Urban street lighting	-4.104605048928162e-7
Total electricity consumption	2.52316473135583e-10

Notice that the coefficient of Number of observatories in the region is 0. This is because there are no observatories in the two regions we have chosen. Thus, we also take this into account when training the model. The data from the cities with observatories were purposely excluded to ensure accuracy of the mode.

Finally, the same dataset is used to train the gradient boosting tree model. Finally, the strongest learner is obtained. At this point, we have successfully built this model. Next, this model is used to analyze our two selected regions.

The two cities we selected are Zhejiang and Chongqing in China. The true values of light pollution metric for these two regions are -0.4596 and -0.3444, respectively. Their other data will be changed by the policy change. And the changed values are put into the model for prediction. The final data obtained for the best policies in these two regions are shown in the Table 5. After the policy regulation, there is a significant decrease in the light pollution level in the two regions.

Table 5. Changes and strategies of light pollution in Zhejiang and Chongqing

Area	Light pollution metric		Strategies
	Before	After	
Zhejiang, China	-0.4596	-0.0129	RLH 62%
			RLT 83%
			RRLE 29%
Chongqing, China	-0.3444	0.0435	RLH 51%
			RLT 84%
			RRLE 67%

The RLH, RLT and RRLE in the table correspond to the three management strategies mentioned in the previous section, respectively. That is, reducing light hours, Reducing the light threshold and reducing the range of light exposure.

Based on the results of the two regions above, different strategies should be developed for different regions. For Zhejiang, the strongest policy should be Reduce the range of light exposure, because Zhejiang is one of the most economically developed regions in the world, where commercial areas are large, and businesses use a lot of light to decorate their stores. For Chongqing, the strongest policy should be Reducing light hours. Most communities in Chongqing are not designed intelligently enough, and the lights in many communities are not turned on as they are used.

4. Summary

Firstly, the potential light pollution capacity index is solved based on entropy weight method and TOPSIS method, and the potential light pollution capacity is solved by neural network algorithm optimized by genetic algorithm, so that TANGY model is established to evaluate the light pollution level in a region. Four regions, Shanghai, Guangdong, Hubei and Tibet, are selected to represent the four places in the question for analysis. The results showed that the degree of light pollution was higher in Shanghai and Guangdong, and lower in Hubei and Xizang.

Second, three policies to control light pollution are identified, and then K-Means ++ and least squares regression are used to obtain a model that can predict light pollution levels using policy influencing factors. The GBDT algorithm was used to predict the light pollution level after the policy adjustment. The model is applied to Chongqing and Zhejiang in China, and the results show that Zhejiang should increase the light reduction range, and Chongqing should increase the light reduction hours.

References

[1] Weiqiang Liao;Shixian Lin.Prediction of Photovoltaic Output Power Based on Match Degree and Entropy Weight Method[J].International Journal of Pattern Recognition and Artificial Intelligence,2023,Vol.37(7): 2350018.

[2] Jie Ma;Yilei Zeng;Dawei Chen.Ramp Spacing Evaluation of Expressway Based on Entropy-Weighted TOPSIS Estimation Method[J].Systems,2023,Vol.11(139): 139.

[3] Zeyun Chen.Port Logistics Function Evaluation Model Based on Entropy Weight TOPSIS Method[J]. Discrete Dynamics in Nature and Society,2022, Vol.2022.

[4] Lu, Jin-Cheng;Li, Mei-Juan.An Improved TOPSIS Within the DEA Framework.[J].Asia-Pacific Journal of Operational Research,2023,Vol.40(6): 1.

[5] Vicente Liern;Blanca Pérez-Gladish.Building Composite Indicators With Unweighted-TOPSIS[J].IEEE Transactions on Engineering Management,2023,Vol.70(5): 1871-1880.

[6] Mohamed Marzouk;Maryam El-Maraghy;Mahmoud Metawie.Assessing retrofit strategies for mosque buildings using TOPSIS[J].Energy Reports,2023,Vol.9: 1397-1414.

[7] Yuanbin Wang; Yu Duan; Yuanyuan Li; Huaying Wu. Smoke Recognition based on Dictionary and BP Neural Network. [J]. Engineering Letters,2023, Vol.31(2): 554-561.

[8] Libiao Bai; Yuqin An; Yichen Sun.Measurement of Project Portfolio Benefits With a GA-BP Neural Network Group[J]. IEEE Transactions on Engineering Management,2024, Vol.71: 4737-4749.

[9] Qinwei Li;Lunxiao Wang;Xiaoguang Lu;Dequan Ding;Yang Zhao;Jianwei Wang;Xinze Li;Hang Wu;Guang Zhang;Ming Yu;Ping Han.Classification and Location of Cerebral Hemorrhage Points Based on SEM and SSA-GA-BP Neural Network[J].IEEE Transactions on Instrumentation and Measurement,2024,Vol.73: 1-14.

[10] Luca Di Persio; Nicola Fraccarolo.Energy Consumption Forecasts by Gradient Boosting Regression Trees[J]. Mathematics,2023, Vol.11(1068): 1068.

[11] Authors: Alberg, Dima | Tessler, Nina.Predicting aircraft maintenance timetofailure using the interval gradient regression tree algorithm[J]. Intelligent Decision Technologies, 2022, 1-8.