ISSN 2616-5872 Vol.5, Issue 7: 21-27, DOI: 10.25236/AJEE.2023.050703

Research on light pollution problem based on K-means regression

Wenzhuo Chen[1], Ziqi Gao[2], Hao Wang[3], Zhongpeng Liu[4], Haojun Hu[5]

[1]School of Shipping and Naval Architecture, Chongqing Jiaotong University, Chongqing, 402247, China
[2]College of Civil Engineering, Chongqing Jiaotong University, Chongqing, 402247, China
[3]Institute of Finance and Economics, Qinghai University, Xining, Qinghai, 810016, China
[4]Clinical Medical Technology College, Sichuan Health Rehabilitation Vocational College, Zigong, Sichuan, 643000, China
[5]Sichuan Vocational College of Judicial Police, Deyang, Sichuan, 618000, China

Abstract: *In this paper, the light pollution problem is studied. Firstly, the principal component analysis method is used to delete and merge the existing data, so as to get the final judgment index. Secondly, the entropy weight method and TOPSIS evaluation method were combined to obtain the light pollution risk index (LDI) with comprehensive scores, and the K-Means clustering method was adopted to classify figure the light pollution risk levels according to LDI. Finally, indicators were collected and ranked for four regions: protected area, rural community, suburban community and urban community, so as to determine their risk levels. The study found that the levels of light pollution risk in protected areas, rural communities, suburban communities and urban communities gradually increased.*

Keywords: *Entropy weighting method; TOPSIS; K-means*

1. Introduction

With the rapid development of China's economy and the increase of social activities at night, light pollution has become a new kind of environmental pollution after the disintegration of waste gas, waste water, sludge and noise. [1] Too much artificial light may disrupt our circadian rhythms, leading to poor sleep quality and affecting people's physical and mental health.[2] At the same time, glare caused by artificial light may cause motor vehicle accidents and threaten life safety. Therefore, a light pollution risk assessment model and a light pollution impact regression prediction model are established in this paper to determine the light pollution risk levels in different regions.

2. Materials and methods

This paper adopts the 2023 American Mathematical Modeling Competition for College Students E Question (https://www.comap.com/contests/mcm-icm) to Analysis and Research.

3. Model establishment and solution

3.1 Light pollution risk level model

According to the characteristics of light pollution and the factors causing light pollution, considering the development level, population, biodiversity, geographical location, and climate of the region, a large number of primary indicators are selected using the McKinsey logic tree, and the indicators are censored according to whether they are quantified or not. Then, four primary indicators and 11 secondary indicators were obtained by using principal component analysis. These indicators are scored using TOPSIS based on the entropy weight method for weighting. Finally, the whole light pollution risk level model is completed according to the K-Means cluster analysis.

3.1.1 Selection of primary indicators

Based on the factors that lead to light pollution, we refer to the relevant literature below [3]. By combining spatial, temporal, lighting equipment, and the physical environment of that different area. In the model, we used McKinsey logic tree analysis to find a large number of primary indicators. We first

ISSN 2616-5872 Vol.5, Issue 7: 21-27, DOI: 10.25236/AJEE.2023.050703

selected 14 primary indicators and 47 secondary indicators.

Among these indicators, some of them are non-specific or fuzzy. Considering the applicability of the model comprehensively, we performed a small part of deletion and synthesis of these indicators according to whether they are quantified. The obtained indicators were then subjected to principal component analysis.

3.1.2 Principal component analysis

Principal component analysis (PCA) was introduced by Pearson to non-random variables in 1901, and Hotel extended this method to random vectors in 1933 [4]. The principal component analysis is a multivariate statistical analysis method that synthesizes several relevant variables into one or a few comprehensive indicators, which can reflect the information of the original variables to the greatest extent [5].

The specific steps of principal component analysis:

According to the original sample matrix X ,the indicator is positively processed: featureXi average value:

$$\overline{X_i} = \frac{1}{M}\sum_{i=1}^{M} X_1^i \tag{1}$$

featureXi average value:

$$\overline{X_2} = \frac{1}{M}\sum_{i=1}^{M} X_2^i \tag{2}$$

Standardize the data after forward processing:

$$X_{ij}^{*} = \frac{X_{ij} - X_i}{\partial_i} \tag{3}$$

In the formula $\bar{X}_t\partial_i$ respectively represent i sample mean and standard deviation of the indicators

Correlation judgment between indicators to get the correlation matrix R :

Obtained according to the equation R and P eigenvalue λ_i and eigenvectors

$U_j(i=1,2,3,...,p)$, cumulative contribution rate≥85% or eigenvalues greater than 1 the first m factors are determined as principal components:

$$\left|R - \lambda_i E\right| = 0 \tag{4}$$

$$Contribution\ rate = \frac{\lambda_i}{\sum_{k=1}^{n}\lambda_k} \tag{5}$$

$$Cumulative\ contribution\ rate = \frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{n}\lambda_k} \tag{6}$$

Calculate the principal component score F_i :

$$F_i = a_{1i}X_{x1} + a_{2i}X_{x2} + ... + a_{ni}X_n \tag{7}$$

Among a_{1i}, a_{2i}, …… a_{pi} for X covariance matrix the eigenvectors corresponding to the eigenvalue, X_{x1}……$X_p$is eigenvectors corresponding to the eigenvalues.

Use the formula to calculate the comprehensive score F :

ISSN 2616-5872 Vol.5, Issue 7: 21-27, DOI: 10.25236/AJEE.2023.050703

$$F = \frac{\lambda_1}{\lambda_1 + \lambda_2} F_1 + \frac{\lambda_1}{\lambda_1 + \lambda_2} F_2 \quad (i = 1, 2, ..., m) \tag{8}$$

The specific results are shown in Table 1.

Table 1: Variance Explanation Table

Element		Characteristic root	
	Characteristic root	Variance Rate (%)	Cumulative variance (%)
1	6.954	63.221	63.221
2	1.184	10.767	73.988
3	0.993	9.024	83.012
4	0.82	7.455	90.467
5	0.387	3.522	93.989
6	0.259	2.353	96.342
7	0.176	1.596	97.937
8	0.093	0.841	98.779
9	0.06	0.542	99.321
10	0.047	0.427	99.748
11	0.028	0.252	100

We plot gravel plots according to how well each principal component explains the variation in the data. Its function is to determine the number of principal components we need to select according to the slope of the eigenvalue decline, extracting the number of principal components through the unknown judgment of "slope tends to be flat." In addition, the variance interpretation table is used to confirm or adjust the number of principal components. The specific results are shown in Figure 1.

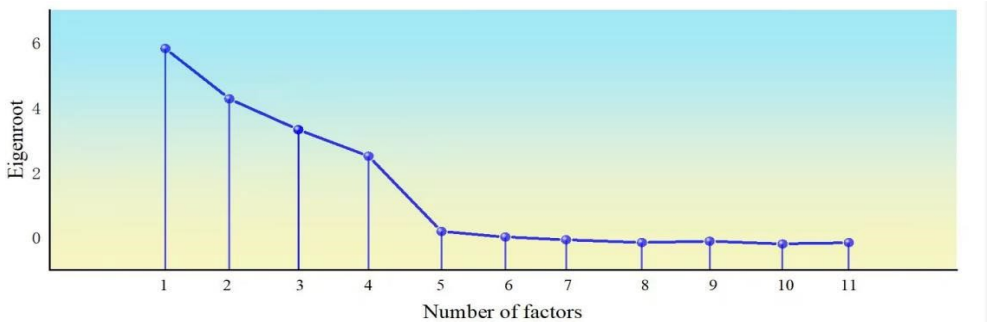

Figure 1: Gravel diagram

At the same time, we combined factor loading quadrant analysis to obtain the final division results. Light pollution, regional, biological, and policy indexes are four first-level and 11 second-level indicators.

(1) Light pollution index

Light pollution affects our urban environment and ecosystem to different degrees. A system of indicators to study the risk level of light pollution, the light pollution index is fundamental, which includes the value of upward light proportional illumination, artificial light intensity, and sky brightness.

Uplighting ratio: Uplighting is a significant type of pollution that causes urban skies to shine. Among them, the illumination of lamps and the reflection characteristics of urban surfaces are the main factors affecting the proportion of uplight, so we can quantitatively evaluate the uplight from these two aspects.

Illuminance value: It is a parameter describing the amount of luminous flux irradiated to a particular surface[6]. In some areas with close road layouts, dense traffic tracks, and a high level of economic development, as well as a large number of urban shopping malls, surrounding industrial night plants and landscape lighting will cause higher illumination at night. The abbreviation symbol is E, and the unit of measurement is lux(LX). The calculation formula is as follows:

$$E = \frac{LM}{M^2} \tag{9}$$

Artificial light intensity: With the development of urbanization, artificial light is an essential factor influencing the degree of light pollution. It can also be used as a potential factor to measure the degree

ISSN 2616-5872 Vol.5, Issue 7: 21-27, DOI: 10.25236/AJEE.2023.050703

of light pollution[7].

Sky brightness: The brightness of the sky gradually extends from the horizon to the zenith. In general, the higher the brightness of the sky, the more serious the light pollution, so the team selected sky brightness as one of the essential indicators for evaluating light pollution. Then, by grading the night sky's brightness, people can intuitively know the pollution level of the area.

(2) Regional Index

The development level, terrain, and climate of an area can all be used as indicators of light pollution in the area. The development level of a region refers to the amount of development in this region, which reflects the scale or level of social and economic phenomena in different periods. In general, light pollution is relatively severe in regions with high levels of economic development.

Different terrains have varying degrees of impact on light pollution, and complex terrain has a diffuse impact on light pollution. Some terrains have very few people living in closed forests, and the level of light pollution is low. Some areas are flat, populated by many people, and have relatively light severe pollution. There are also some mountainous areas with high terrain, so the pollutants emitted by the pollution sources are blocked by the mountains and form reflections, resulting in light severe pollution in some areas.

Although light pollution does not change the climate more than greenhouse gas emissions, to a certain extent, light pollution will also affect the climate.

(3) Biological index

Both humans and non-humans contribute to light pollution. A city with a higher population density will have a relatively higher level of light pollution because the increase in population will drive the development of shopping malls, factories, transportation, and tourism, and the development of these factors will generate more light pollution, thereby increasing the risk level of light pollution.

Except for very few animals active at night, most are quiet and do not like intense light. However, the interfering light and reflected light produced by outdoor lights and decorative lights at night seriously affect the life and rest of animals and disrupt the circadian rhythm of animal life. In the long run, the balance of the ecological system in this area will be seriously damaged. The richness will decrease, so biodiversity[8]can also be used as an essential indicator of light pollution levels.

$$R = \frac{N_1 + N_2}{2} \tag{10}$$

Where, R ——Richness of species. N_1The——number of wild animal species in the assessed area. N_2——Number of wild vascular plant species in the assessed area.

3.1.3 Establishment of comprehensive evaluation model

In order to further evaluate the light pollution risk level by weighting the four first-level and 11 second-level indicators, we have determined. We use the entropy weight method to determine each index's weight, then use TOPSIS to score the weight, and finally define the score as the light pollution risk index.

(1) Entropy weight method

The basic idea of the entropy weight method is to determine the objective weight according to the size of the index variability.

Generally speaking, the smaller the information entropy in the index, the greater the variation degree of the index value, the more information it provides, the more significant the role in the comprehensive evaluation, and the greater the weight [9]. The greater the information entropy of an indicator, the smaller the change in the indicator value, the less information it provides, the smaller the role in the overall evaluation, and the smaller the weight.

Weighting steps of entropy weight method:

Data standardization: standardize the data of each indicator. For positive indicators:

$$X_i^{'} = \frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)} \tag{11}$$

ISSN 2616-5872 Vol.5, Issue 7: 21-27, DOI: 10.25236/AJEE.2023.050703

Direction indicator:

$$X_i^{'} = \frac{\max(X_i) - X_i}{\max(X_i) - \min(X_i)} \tag{12}$$

Among X_i'represents the value that normalizes the data for each metric.

Calculate the probability matrix:

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y}_{im})(y_i' - \overline{y}'_{im})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y}_{im})^2 \sum_{i=1}^{n}(y_i' - \overline{y}'_{im})^2}} \tag{13}$$

Find the information entropy of each indicator:

$$E_j = k\sum_{i=1}^{n} P \ln P_{ij} \tag{14}$$

Calculation of information utility value:

$$D_i = 1 - E_j \tag{15}$$

Calculate the weight coefficient of the index:

$$W_j = \frac{D_i}{\sum_{i=1}^{m} D_i} \tag{16}$$

The specific indicators are shown in Figure 2.

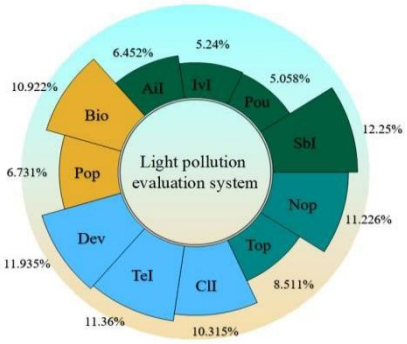

Figure 2: Weight calculation chart

(2) TOPSIS Evaluation method

The TOPSIS method is a commonly used comprehensive evaluation method [10]. The comprehensive distance between any program in the evaluation index system and the optimal and worst solutions is calculated through specific calculations.

TOPSIS algorithm specific steps:

Forwarding the original matrix: intermediate metrics:

$$M = \max\{|x_i - x_{best}|\} \tag{17}$$

$$\overline{x_i} = 1 - \frac{|x_i - x_{best}|}{M} \tag{18}$$

Normalize the normalization matrix:

$$Z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=0}^{n} x_{ij}^2}} \tag{19}$$

ISSN 2616-5872 Vol.5, Issue 7: 21-27, DOI: 10.25236/AJEE.2023.050703

Calculation score normalization: the final score results are as Figure 3.

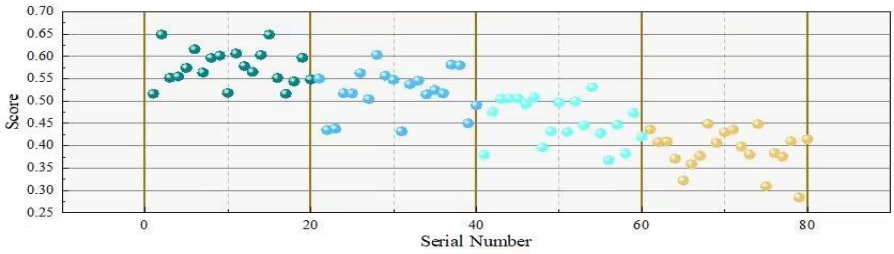

Figure 3: Light pollution level final score chart

Finally, we collated data from 80 different regions and clustered their final scores into four categories based on the evaluation index system through K-Means, divided into four types: tiny, low, medium, and high. Each type corresponds to a corresponding segment of values, which is used to evaluate the risk level of light pollution. The classification results are as follows:

Tiny: Score results between 0.562531 and 0.648535.

Low: Score results between 0.475682 and 0.556426.

Medium: Score results between 0.396107 and 0.473378.

High: Score results between 0.283883 and 0.383046.

3.2 Indicators are applied to four different types of positions

We now have a model for measuring the level of light pollution risk that incorporates many factors and a system of indicators with broad applicability as required by topic one. There is no doubt that light pollution will affect some of these factors to different degrees in different locations and thus to different degrees for humans and non-humans. In this section, we will apply our indicator system at four different types of locations: a protected land location, a rural communities, a suburban communities, and an urban communities, and we will interpret the results.

3.2.1 Data selection

According to the four different indicator locations provided by the topic, we select some environmental characteristics of these locations and make some sub-regions in these environmental characteristics to make our data source more adaptable for that region. We also consider the differences between the four types of area tables because, generally, artificial light brightness values are high in urban communities. In contrast, suburban communities have somewhat lower brightness values, and rural ones are even lower, and finally, protected land. We find the most significant differences among the four regions, i.e., the light pollution thresholds accepted by different regions differ.

3.2.2 Indicator Application

We apply our indicators to these four different types of locations. The data for each region is different. Based on the model in the first question, we can calculate the light pollution intensity of each region separately.

$$\beta_{actual} = \beta_{total} - \alpha_i \tag{20}$$

β_{total} represents the actual calculated light pollution, β_{actual} indicates the actual risk of light pollution in the area, $\alpha_i$indicates the light pollution that should be generated. The specific results are shown in Table 2.

Table 2: Comprehensive scoring and ranking of the four regions

Index	Score	Rank	Index	Score index	Rank
E1	0.623140192	8	E3	0.529063611	61
E2	0.531117524	34	E4	0.376349601	79

ISSN 2616-5872 Vol.5, Issue 7: 21-27, DOI: 10.25236/AJEE.2023.050703

4. Conclusions

This study found that the levels of light pollution risk in protected areas, rural communities, suburban communities and urban communities increased gradually. The results were closely related to the geographical location of the four regions.

For protected areas, the government introduced light pollution prevention and control policies earlier and more forcefully. For rural communities, there is a lack of diverse entertainment and shopping centers as in cities, a low overall level of development, low population density, and little light pollution. As for the suburbs, there are more people and fewer businesses, mainly distributed in residential areas. For cities, there is more housing and more traffic, which results in high population density and high levels of artificial light intensity. Although some big cities have introduced light pollution prevention and control policies in the early stage, compared with other factors causing light pollution, the improvement effect of the policies is less reflected to some extent. Therefore, the government should constantly improve and introduce new prevention and control measures.

References

[1] Hao Ying, Li Wenjun, Zhang Peng, Zhang Jinyan, Xu Yang, Sun Hongbo. China Population, Resources and Environment, 2014, 24(S1): 273-275.
[2] Chen S L. Effects of light pollution on environment and health [J]. Chinese Journal of Tropical Medicine, 2007(06):1005-1009.
[3] Liu M, Zhang BG, Pan XH, Yuan J. Research on light pollution evaluation indexes and methods in urban lighting planning [J]. Journal of Lighting Engineering, 2012, 23(04): 22-27+55.
[4] Du Min. Research on comprehensive index of environmental quality based on principal component analysis [D]. Sichuan University, 2006.
[5] Bai Yi. Principal component analysis model and principle of multi-indicator comprehensive evaluation [J]. Journal of Shaanxi Normal University (Natural Science Edition), 1998(02): 109-110.
[6] Li Jiayi, Xu Yongming, Cui Weiping, et al. Nighttime light pollution monitoring in Nanjing based on Liaojia-1 nighttime light remote sensing data [J]. Remote Sensing of Natural Resources, 2022, 34(2): 289-295.
[7] Bedi T.K., Puntambekar K., Singh S., 2021. Light pollution in India: appraisal of artificial night sky brightness of cities. Environ. Dev. Sustain. 1-16
[8] Guo Y, Feng N, Christopher S A, et al. Satellite remote sensing of fine particulate matter (PM2.5) air quality over Beijing using MODIS[J]. International Journal of Remote Sensing, 2014, 35(17-18):6522-6544.DOI:10.1080/01431161.2014.958245.
[9] Zhang Z, Zhang Yichen, Zhang Jikuan et al. Urban resilience assessment based on entropy weighting method and TOPSIS model - an example of Changchun city [J/OL]. Disaster Science: 1-12 [2023-02-20].
[10] Yuan Ying, Wu Weijie, Zhang Yining et al. Preferential evaluation of pumped storage sites in Guangdong Province based on AHP-entropy weight method-TOPSIS [J]. Guangdong Water Conservancy and Hydropower, 2023, No. 323(01): 37-42.