ISSN 2616-7433 Vol. 5, Issue 11: 117-121, DOI: 10.25236/FSST.2023.051118

Assessment of light pollution based on the k-mean clustering analysis

Minghao Guan[1,*,#], Yifeng Zhu[2,#], Xuan Liu[3,#]

[1]*School of Mathematics and Statistics, Central South University, Changsha, 410083, China*
[2]*School of Computing, Central South University, Changsha, 410083, China*
[3]*Business School, Central South University, Changsha, 410083, China*
[*]*Corresponding author: 8201210506@csu.edu.cn*
[#]*These authors contributed equally.*

Abstract: *With the development of science and technology and the increase of population, artificial light is used more and more. Abuse and misapplication of artificial light will cause light pollution. Light pollution may affect the nighttime environment, disrupt the living habits of organisms, affect human health and affect astronomical observations. The paper innovatively divided the light pollution risk model into two levels, which were equivalent to the correction of the light pollution degree index. After selecting appropriate evaluation indicators, the paper first collected data from 20 countries to satisfy the universality of the model. Next, this essay used a method which combined of subjectivity and objectivity to assign weights to each indicator and calculated the scores of 20 countries. Finally, it divided the light pollution risk score into four levels by K-means clustering, which were no risk, low risk, medium risk, and high risk.*

Keywords: *Light pollution, Subjective and objective empowerment methods, K-mean clustering*

1. Introduction

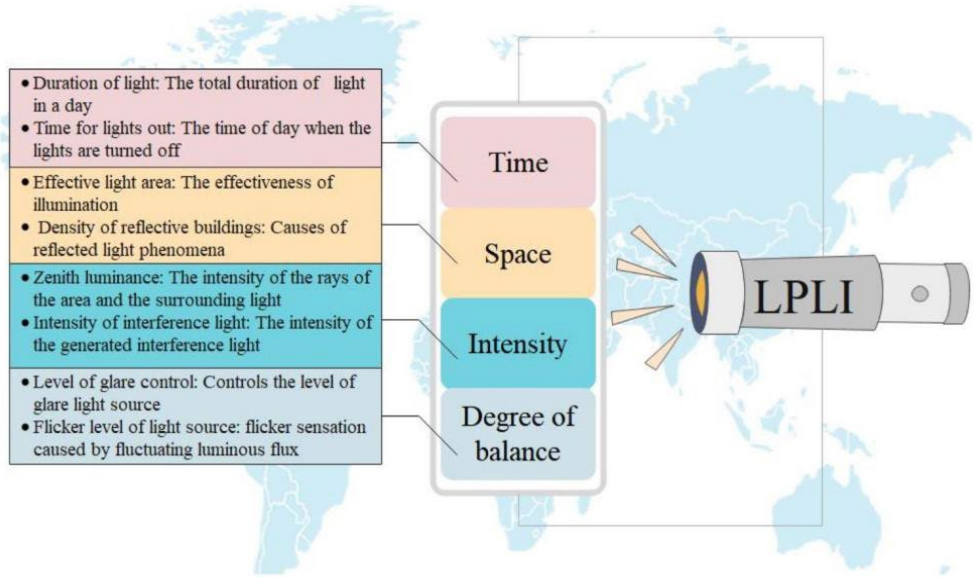
Light pollution, which refers to the pollution of the environment by dry light, is used to describe any excessive or undesirable use of artificial light. Light trespass, excessive illumination and light clutter are some of the phenomena it causes [1].These phenomena are most likely to be observed as glow in large cities, but they can also occur in some remote areas. Almost all countries in the world receive the effects of light pollution. Light pollution has a wide range of effects, for both plants and animals and humans. For wildlife, for example, artificial lighting may take them off their migration routes or spawning sites, causing a decline in reproduction and disrupting ecosystems. In addition, for crops, the blurring of night and day hinders growth and development, resulting in lower yields. In the case of humans, light pollution can cause insomnia, headaches, depression, emotional disturbance, stress, indigestion, breast and prostate cancers.[2]Due to the wide range and severity of the effects of light pollution, various governments attach great importance to the management and intervention of light pollution[3], but the intervention of light pollution needs to be combined with the characteristics of different regions, and the positive and negative effects of artificial light need to be considered at the same time. In order to develop a universal indicator to evaluate the level of light pollution risk in a location, the paper need to select representative indicators and assign weights to them to realize the construction of the indicator system. Subsequently, the paper need to calculate the scores of each country and derive the ranking classification, so that the paper can well evaluate the light pollution risk level of any region. After building the evaluation model, the paper need to bring the data of 4 different locations into the model, derive the scores, and analyze the causes according to the scores of different indicators.

2. Establishment of the multiplication model

2.1 Primary indicator system of LPLI

According to recent studies [4], the common evaluation index of the degree of light pollution includes three dimensions: time, space, and intensity, on the basis of which we add the degree of balance, defining the light pollution level index in four dimensions. The indicators of four dimensions are as follows in figure 1.

ISSN 2616-7433 Vol. 5, Issue 11: 117-121, DOI: 10.25236/FSST.2023.051118



(From the perspective of time, space, intensity and degree of balance, the model defines eight indicators that are incorporated into the light pollution level index.)

Figure 1: The process flow of establishing the criteria for evaluating the degree of light pollution

2.2 Primary indicator system of VI

Based on recent studies [5], as shown in Figure 2, we divided the evaluation indicators of light pollution vulnerability into four dimensions: level of development, population, biodiversity, geography, and climate, and further subdivided them into nine observations.

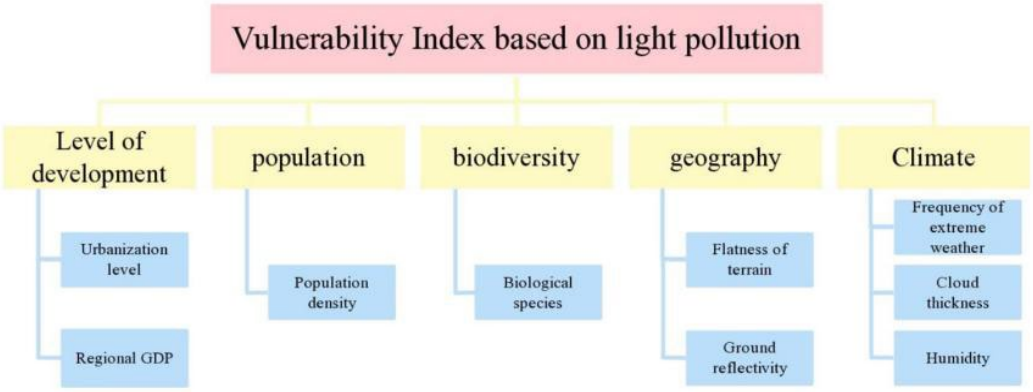


Figure 2: Indicator system of vulnerability index based on light pollution

(1) Level of urbanization: Urbanization level refers to the degree of urbanization reached by a region. With the accelerated urbanization of a region and the construction of urban brightening projects, the problem of light pollution becomes more and more serious. Therefore, we introduce urbanization level to reflect the development level.

(2) Population density: It is the average number of people on a certain unit area of land in a certain period of time.

(3) Biological species: The sum of ecological complexes formed by organisms (animals, plants, microorganisms) and the environment. The number of biological species can reflect the degree of light pollution in an area to a certain extent.

(4) Geography: An increase in ground reflectance leads to an increase in the intensity of reflected radiation from the ground, thus increasing the degree of light pollution.

(5) Climate: The frequency of extreme weather causes problems with the restorative nature of the ecosystem itself, which indirectly leads to light pollution not being better repaired by the ecosystem itself,

ISSN 2616-7433 Vol. 5, Issue 11: 117-121, DOI: 10.25236/FSST.2023.051118

thus increasing light pollution.

3. Introduction to the Index weight determining methods

This essay collected data of 17 different indicators from 20 countries. This section will briefly introduce the methods for assigning weights to the two indices, subjective weighting methods including hierarchical analysis and sequential relationship analysis, and objective weighting methods including entropy weighting method and random forest.

3.1 Subjective weight assignment method

The main idea of AHP is to use mathematical methods to determine the values that express the relative importance of all elements of each level, and through the analysis of each level to derive the weights of the importance of different solutions, to provide a basis for the selection of the best solution.

The following four steps are generally required:

Step 1: Establish hierarchical structure model.

Step 2: Construct A judgment matrix.

Step 3: Single hierarchical ranking weight.

Step 4: Consistency check.

$$CI = \frac{\lambda \max - n}{n-1} \tag{1}$$

3.2 Objective weight assignment method

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

3.3 Random Forest

A random forest is a combinatorial classification model that consists of multiple decision trees. The parameter sets are independently and identically distributed random vectors, and given the independent variables, the optimal classification result is voted on by each decision tree model. In the random forest model, the training samples of each tree and the splitting attributes of the nodes are chosen randomly, and the combined effect of the two randomnesses avoids overfitting of the model to a certain extent and makes the model more robust. In addition, a large number of theoretical and applied studies have demonstrated the accuracy of the random forest model from different perspectives. The model is well tolerant of outliers and noise in the dataset and is currently recognized as one of the best machine learning models.

The process of using the random forest model is shown in Figure 3:

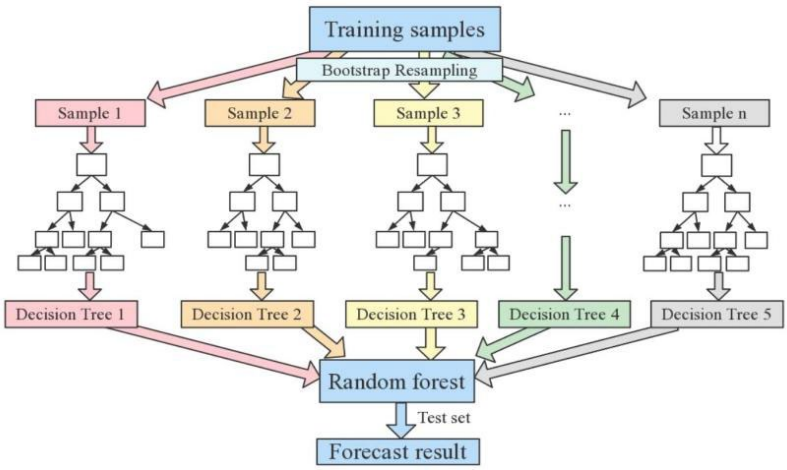


Figure 3: Stochastic forest prediction flow chart

ISSN 2616-7433 Vol. 5, Issue 11: 117-121, DOI: 10.25236/FSST.2023.051118

4. Determination of the weights and the K-means clustering algorithm

This essay applied the four assignment methods in the model introduction to the processed data, including two subjective and two objective methods. Then, we arithmetically averaged the weight vectors derived from the four methods and finally arrived at Optimum weighting.

This paper also performed a consistency test for the AHP method and performed an error analysis for the random forest, and the results are presented in table 1.

Table 1: The CI value of AHP and the error of random forest

	CI	Standard Deviation	Accuracy
LPLI	0.0680	0.1023	0.973
VI	0.0330	0.1012	0.978

It can be seen that the CI values of both indices are less than 1, and both of them pass the consistency test of AHP. The random forest accuracy for both indices was within the interval [0.9-1] and the mean square error was within the interval [0.0998-0.15], indicating that the accuracy of the assignment using random forest was high.

Finally, we use K-means clustering algorithm to cluster these 20 countries. K-means algorithm measures the similarity of different data objects by selecting appropriate distance formula. The distance between data is inversely proportional to the similarity, that is, the smaller the similarity, the greater the distance.

K-means algorithm first needs to randomly specify the initial clustering number k and corresponding initial clustering center C from a given data object, and calculate the distance between the initial clustering center and other data objects. In this paper, Euclidean distance is selected, and the Euclidean distance formula between the clustering center and other data objects in the space is as follows:

$$d(x,C_i)=\sqrt{\sum_{j=1}^{m}(x_j-C_{ij})^2} \tag{2}$$

The x above is the data object, C_i is the ith clustering center, m is the dimension of the data object, X_j and C_{ij} are the attribute values of the jth dimension of the data object C_i and clustering.

X center C_i. According to Euclidean distance, the similarity is measured, and the target data with the highest similarity to the clustering center is allocated to the C_i cluster. After the allocation, the data objects in the k clusters are averaged to form a new round of clustering center, to reduce the sum of squares of errors of the data set.

The value of SSE is used as the basis to measure the quality of clustering results. When it no longer changes or converges, the iteration is stopped, and the final result is obtained. The flow chart is shown in figure 4.

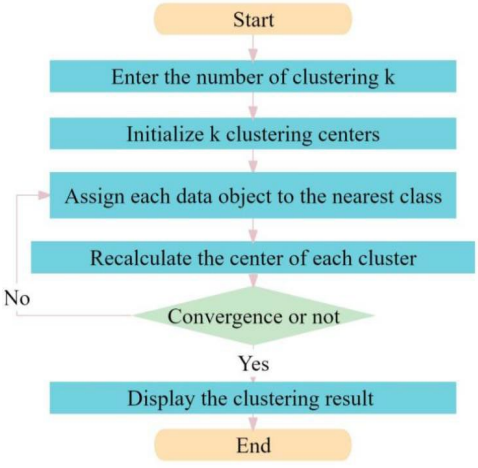


Figure 4: The flow chart for K-means clustering

ISSN 2616-7433 Vol. 5, Issue 11: 117-121, DOI: 10.25236/FSST.2023.051118

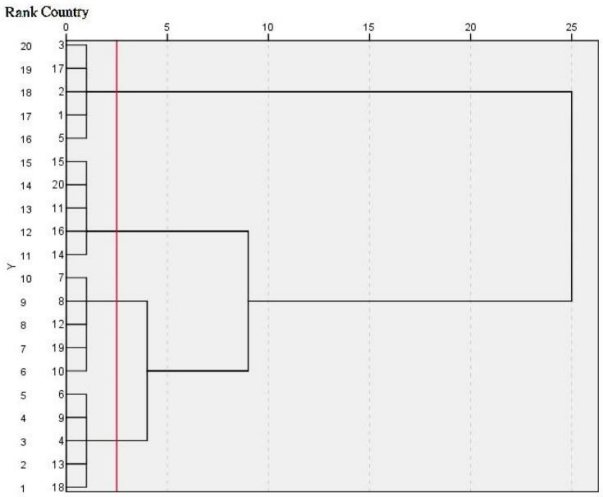


Figure 5: Clustering results

Serial numbers in figure 5 along the axis of the label represent countries:(1)Kuwait (2)Singapore(3)Qatar (4)Saudi Arabia (5)South Korea (6)United Arab Emirates (7)France (8)Italy (9)Germany(10)Canada (11)Central African Republic (12)Australia (13)United States (14)South Africa(15)Britain (16)Madagascar (17)Zimbabwe (18)China (19)Holland (20)Chad. After traversing the process, we can classify the different indicators of the selected countries into three levels: instability, warning, and stability. The higher the number, the better the indicator. The clustering results are shown above in figure 5. Light pollution risk level: A value between 0 and 0.1 indicates No risk, between 0.1 and 0.2 indicates low risk, between 0.2 and 0.6 indicates medium risk and above 0.6 indicates high risk.

5. Conclusions

This paper use both subjective and objective weighting methods to integrate new weights based on moment estimation, which considers the subjective non-negligibility of events and the intrinsic linkages of the indicators to be evaluated. But our model is based on data that can be collected on the web and published by governments. Due to the large number of countries involved, the reliability of our data sources needs to be further verified, which may lead to some bias in the resulting weights.

From the perspective of LPLI alone, protected land is the least serious. Rural community is the least serious. Suburban community is the most serious. And urban community is the most serious. In fact, due to the dense population and rapid economic development in the urban community, after nightfall, the advertising lights and neon lights on shopping malls and hotels are dazzling and dazzling, causing serious light pollution. Protected land, on the other hand, is closer to nature and far from Artificial light, so light pollution is less. This is consistent with our conclusions. In terms of vulnerability, too. On the whole, protected land is at a risk-free level. Rural community is at a low risk level. Suburban community is at a moderate risk level, and urban community is at a high-risk level.

Protected land is risk-free and all indicators are in good shape, especially related indicators such as biodiversity and climate. So, we need to keep light pollution away from this last pure land.

References

[1] Bará Salvador, Bao-Varela Carmen, Falchi Fabio. Light pollution and the concentration of anthropogenic photons in the terrestrial atmosphere [J]. Atmospheric Pollution Research, 2022, 13(9).
[2] Ayudyanti Amalia Gita, Hidayati Iswari Nur. Impact of Optical Aerosol Depth (AOD) on Light Pollution Level: a spatio-temporal analysis [J]. IOP Conference Series: Earth and Environmental Science, 2021, 884 (1).
[3] Hee-Kyung Yun. Study on Standardization of Environmental Impact Assessment Guidelines in the field of light pollution, Environmental impact assessment. 2019, 28(1): 63-70.
[4] Mykyta Peregrym, Erika Pénzesné Kónya, Fabio Falchi. Very important dark sky areas in Europe and the Caucasus region [J]. Journal of Environmental Management, 2020, 274.
[5] Hao Ying. China Population, Resources and Environment. 2014, 24(S1): 273-275.