WEP
Warwick
Evans
Publishing

Transactions on Environment, Energy and Earth Sciences
ISSN: 2960-1924, eISSN: 2960-2300 | Volume 1
MSREE 2023

Evaluation model of light pollution level based on ID3 decision tree

Xin Wang, Xinshe Qi, Xinwei Gao, Cuicui Gao, Na Wang, Shasha Wu

Institude of information communication, National University of Defense Technology Wuhan, Hubei, China

Abstract. This article solved the problems like quantitative model of light pollution degree is established; we decision three model is used to predict the risk levels of light pollution in four different regions. In this paper, 16 evaluation indexes are selected, which are analyzed and processed by simple data reduction, and 5 indexes which have great influence factors and are widely concerned are selected. The classification model of light pollution level based on ID3 decision tree was established to evaluate the light pollution level of each region.

Keywords: Topsis, ID3 decision tree, PCA, SHAP, light pollution.

1. Introduction

Light pollution refers to any excessive or improper use of artificial light, it includes light intrusion, excessive lighting, light clutter and so on. Light pollution can occur during the day and at night. During the day, light pollution is mainly due to eye vertigo caused by mirror buildings. At night, light pollution can affect people's sleep quality and traffic safety. In recent years, with the development of science and Technology, industry and the increase of population density, the distribution and intensity of artificial light source is more and more obvious. Although these lamps and lanterns bring convenience to human social life, the contradiction of light pollution is more and more outstanding.

As a by-product of urban night scene lighting, light pollution is contrary to the concept of sustainable development. Therefore, it is necessary to establish a widely applicable light pollution assessment method and model based on the existing urban zoning and environmental development, and propose targeted intervention strategies for different areas.

2. Develop a broadly applicable metric to identify the light pollution risk level of a location.

In order to make a scientific analysis of urban lighting space, it is necessary to analyze and evaluate the light pollution from a city's big angle of view. As an academic problem, light pollution should have scientific basis and strict logic. After identifying and Quantitative analysis the adverse effects of light pollution, the severity of light pollution can be determined according to a set level. That's the foundation and the Operability.

2.1. The selection of evaluation index

To develop widely applicable metrics to predict and determine the level of light pollution risk in an area. After consulting data and analyzing, our group preliminarily selected the following evaluation indicators.cators and their inclusion relationships

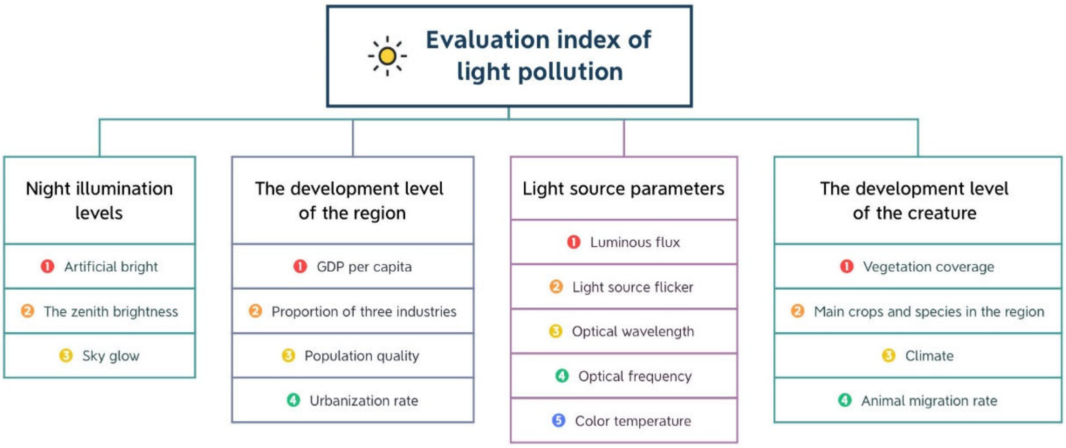

Figure 1: 16 evaluation indicators and their inclusion relationships

The evaluation indexes are divided into four categories: night lighting level, regional development level, light source parameters and biological development level, and these four categories are subdivided into several sub-categories. Using principal component analysis and cluster analysis, our team took artificial light intensity, population density, per capita GDP, vegetation coverage and urbanization rates into account. Here is the definition of these five indicators.

2.1.1. Artificial light intensity

Illumination intensity refers to the amount of visible light that is received on a unit of area, referred to as illuminance, or Lux, and is used to indicate the illuminance and illuminance of an object's surface. It is defined as:

$$Average\ illumination = \frac{Total\ luminous\ flux\ of\ light\ source(N{\cdot}\emptyset)\times utilization\ factor(CU)\times maintenance\ factor(MF)}{area\ of\ region(m^2)} \tag{1}$$

Artificial light source is to make up for the lack of natural light or completely replace the natural light of artificial light source. However, in the case of light pollution, both day and night artificial light can promote its formation. However, night artificial lighting is the main factor causing light pollution, so this paper mainly discusses night artificial lighting.

2.1.2. Population density

Population density refers to the number of people per unit of land area. It is an important indicator to measure the population distribution in a country or region.Define D_P as the population density，N_P is the population of a country or region, S is the land area of the country or region, then

$$D_p = \frac{N_p}{S} \tag{2}$$

Population density plays an important role in light pollution risk level. In general, the higher the population density, the more residents per unit area, resulting in more light use at night. If everyone lights a lamp, it will undoubtedly cause excessive use and waste of light resources, and also lead to light pollution.

2.1.3. Per capita GDP

Per capita GDP is calculated by comparing the gross domestic product achieved in a country's accounting period (usually one year) with the country's resident population (or registered population) to get the gross domestic product per capita. It is an important standard to measure the living standard of people in all countries.

Set G_a as the per capita GDP,as M_P a country (or region) accounting period (usually one year) of gross domestic product (GDP), N_u as the country (or region) of the population of permanent residents (or census register population).Then the per capita GDP is defined as

$$G_a = \frac{M_p}{N_u} \tag{3}$$

Per capita GDP is an important standard to measure the living standard of a country, and people's living standard will affect the consumption of various resources. In general, the higher the per capita GDP, the higher the living standard of the residents in the region, the greater the corresponding consumption of various resources, among which the demand for light will be more. Inadvertently, many people use more light than they need. This often goes unnoticed, perhaps by forgetting to turn off a light in the room the person has left.

2.1.4. Vegetation cover rate

Vegetation coverage rate refers to the ratio between the vertical projection area of plants and the area of a certain region, which is an important index reflecting the forest resources and afforestation level. Set the vegetation coverage rate as C_f, the vertical projection area of vegetation stems and leaves at a measured point of one square meter as S_{fi}, the total vegetation area as S_{tf}, and n as the number of selected measured points of one square meter, then the vegetation coverage rate is defined as

$$C_f = S_{tf} \bullet \frac{1}{n} \sum_{i=1}^{n} S_{fi} \tag{4}$$

Trees can effectively block out some of the incoming light, and lawns can also reduce the intensity of light in the area by scattering light. At the same time, the vegetation coverage rate is often closely related to the development level of the region. Generally, the higher the vegetation coverage rate of an area, the fewer residents in the area, or the farther away from the city center and business district with high nighttime light intensity, so the lower the light pollution risk level.

2.1.5. Urbanization rate

Urbanization rate is a measure of urbanization, generally using demographic indicators, that is, the proportion of urban population in the total population (including agricultural and non-agricultural). Set the urbanization rate as P_C, the number of urban population as N_C , and the number of agricultural population as Na, then the urbanization rate is defined as

$$P_c = \frac{P_c}{P_c + N_a} \tag{5}$$

The urbanization rate reflects the proportion of urban population. The higher the urbanization rate of a region, the larger the area of urban area within the region. Compared with non-urban areas such as rural areas, urban areas have much higher demand for and use of light at night, which is more likely to cause light pollution.

2.2. Establishment of the evaluation model

Before analyzing the degree of light pollution, we have subjectively listed some influential indicators, which have some correlation with the dependent variable, whether positive or negative. The change of index value will directly affect the change of factors. The greater the change, the more obvious the effect of index on the change of factors should be. As an objective assignment method, entropy weight method calculates the entropy weight of each index by using information entropy based on sample data, and then corrects the weight of each index by entropy weight, so as to obtain relatively objective index weight.

2.2.1. Using the entropy weight method to establish model

Indicator forward

Different indicators represent different meanings, some indicators the bigger the better, known as very large indicators. Some indicators, the smaller the better, are called very small indicators, while some indicators are the best at a certain point, called intermediate indicators. In order to facilitate evaluation, all indexes should be converted into extremely large indexes. There are m objects to be evaluated and n evaluation indicators, which can form a data matrix

$$\boldsymbol{X} = (\boldsymbol{x_{ij}})_{\boldsymbol{m\times n}} \tag{6}$$

Let very small indicators be expressed as $(x_{ij})'$, then

$$(\boldsymbol{x_{ij}})' = \mathbf{max}(\boldsymbol{x_{ij}}) - \boldsymbol{x_{ij}} \tag{7}$$

Normalized processing   is every element of the normalized forward matrix,then

$$P_{ij} = \frac{Z_{ij}}{\sum_{i=1}^{n} Z_{ij}} \tag{8}$$

get the probability matrix

$$P = (P_{ij})_{m\times n} \tag{9}$$

Calculate each index of information entropy and information utility value

Let S_j be the information entropy of index j, then

$$S_j = -\frac{1}{\ln n}\sum_{i=1}^{n} P_{ij} \ln(P_{ij}) \tag{10}$$

Guarantees that S_j is on the interval [0,1].Let A_j be the utility value of the j index, namely the redundancy degree of information entropy, then

$$A_j = 1 - S_j \tag{11}$$

The greater the utility of information, the greater the amount of information. Calculate the weight coefficient of the index

By normalizing the information utility value, the weight of each index can be obtained, let $\mathrm{W_j}$ be the weight of the j-th index, then

$$\omega_j = \frac{A_j}{\sum_{j=1}^{m} A_j} \tag{12}$$

Compiled by MATLAB software, the weight coefficient results of each index are returned after data import, as shown in the following table:

Table1. Each index weight coefficient comparison table

Symbol	Weight Coefficient	
$\boldsymbol{L_n}$	ω_1	0.1235
$\boldsymbol{D_p}$	ω_2	0.2111
$\boldsymbol{G_a}$	ω_3	0.1403
$\boldsymbol{C_f}$	ω_4	0.0106
$\boldsymbol{P_c}$	ω_5	0.5145

2.2.2. Establishing an evaluation system

The quantitative fraction of light pollution in different regions is calculated by the following formula

$$Y = \omega_1 L_n + \omega_2 D_p + \omega_3 G_a + \omega_4 P_c + \omega_5 C_f \tag{13}$$

Thus, the evaluation model of regional light pollution degree is established.

2.3. Result

2.3.1. Based on Topsis method to determine the various areas of light pollution level of risk

Topsis, also known as the good and bad solution distance method, is a sort method approximating to the ideal solution, ranking according to the degree of proximity between a finite number of evaluation objects and an idealized target. In 5.2, we obtained the weight coefficient of each evaluation index, calculated the quantitative fraction of sample light pollution degree based on Topsis method, and ranked the risk level of each region. For the sample space with sample size m and evaluation index n, Z is the matrix after standardization of this data set. $\mathrm{Z_{max}}$ is defined as the set of the maximum values of each indicator, and $\mathrm{Z_{min}}$ is defined as the set of the minimum values of each indicator

Define the distance between the ith evaluation object and the maximum value

$$D_i^{max} = \sqrt{\sum_{j=1}^{m} \omega_j (Z_j^{max} - z_{ij})^2} \tag{14}$$

Define the distance between the ith evaluation object and the minimum value

$$D_i^{min} = \sqrt{\sum_{j=1}^{m} \omega_j (Z_j^{min} - z_{ij})^2} \tag{15}$$

Thus it can be concluded that the unnormalized score of the ith evaluation object is

$$S_i = \frac{D_i^{\min}}{D_i^{\max} + D_i^{\min}} \tag{16}$$

it is obviously to see that $0 \leq S_i \leq 1$， and the bigger $\mathrm{S_i}$ is the bigger $\mathrm{D_i^{max}}$ is，

That is, the closer the score is to the maximum.The normalized score is

$$S_{0i} = \frac{S_i}{\sum_{i=1}^{n} S_i} \tag{17}$$

Based on the above formula, we can obtain the corresponding light pollution degree of 20 samples. Its physical meaning is the quantitative fraction of light pollution degree of these 20 cities under the constraints of five evaluation indicators in 5.1: artificial light intensity, population density, per capita GDP, vegetation coverage rate and urbanization rate. According to the degree of light pollution, light pollution is divided into four levels: mild, moderate, severe and very severe.

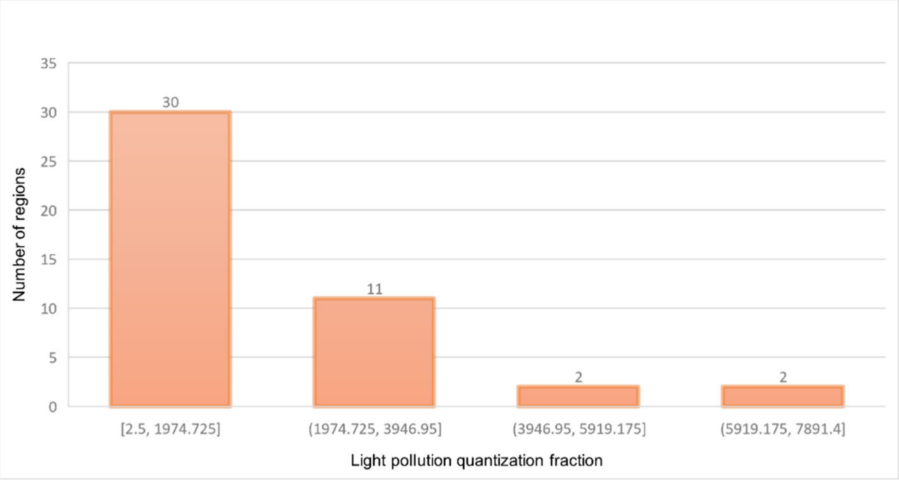

Figure 2. Result graph

3. The evaluation model is used to explain the prediction results of light pollution rate in different types of areas

In order to more intuitively and quickly judge the light pollution rate level of a region, we use the four light pollution levels defined in question 1 to assess the light pollution risk level of the region in question. Firstly, ID3 decision tree model is established, that is, the best features are selected to segment the data according to the "maximum information entropy gain", and the data is recursively divided from root to leaf. Until you have a complete tree. Finally, test the light pollution risk levels of the four regions according to the model questions.

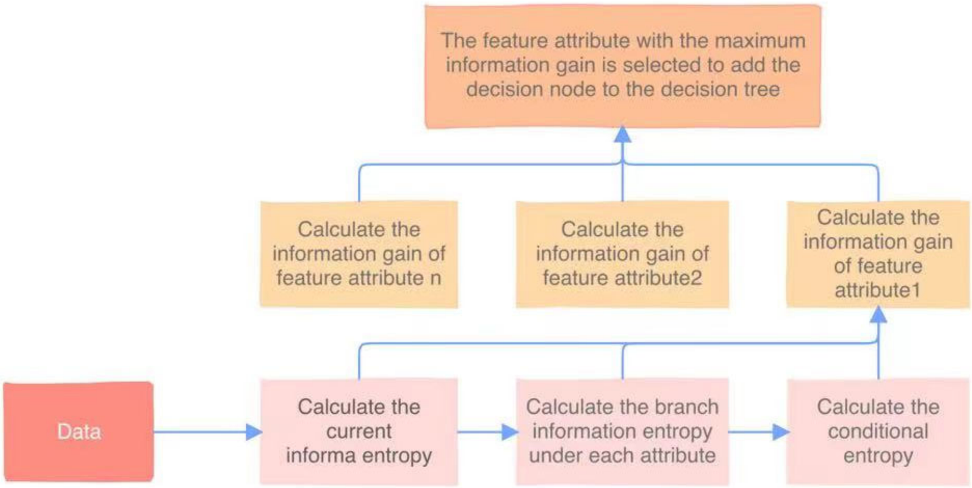
The feature attribute with the maximum information gain is selected to add the decision node to the decision tree
Calculate the information gain of feature attribute n
Calculate the information gain of feature attribute2
Calculate the information gain of feature attribute1
Data
Calculate the current informa entropy
Calculate the branch information entropy under each attribute
Calculate the conditional entropy

Figure 3. ID3 Decision tree flow chart

3.1. Grading of evaluation index

Considering that ID3 decision tree model is unable to deal with continuous eigenvalues, the evaluation indexes should be graded first. The sample data of each evaluation index are divided by the division of the tripartite number and the dichotomy method, and the division results are as follows

Table 3. Evaluation index classification

Symbol	Grade		
	High	Middle	Low
L_n（ucd/m^2）	（13500,∞]	（6740,13500]	（0,6740]
D_p（person/km^2）	（1000,∞)	（100,1000]	（0,100]
G_a（Ten thousand dollars）	（10,∞]	（5,10]	（0,5]
C_f（%）	（54.0,∞]	（28.7,54.0]	（0,28.7]
P_c（%）	(69.5,100]	(40.4,69.5]	（0,40.4]

3.2. Selection of features

Calculate the information entropy of the root node

Information entropy is the most commonly used index to measure the "purity" of a sample set.Let it be the information entropy of the root node, and P_k（k=1,2,3,4） be the probability of mild, moderate, severe and very severe.

$$\mathrm{Ent(D)} = -\sum_{k=1}^{4} P_k \log_2 P_k \tag{18}$$

Calculate the information gain of each evaluation index

After calculating the entropy of the decision attributes, we need to order the features according to the information gain. Information gain is relative to features, and represents the degree to which the uncertainty of information of class Y is reduced when the information of feature A is known. The greater the information gain, the greater the influence of the feature on the final classification result, so this feature is selected as our classification feature. When the first classification criterion is obtained, it continues to extend downward as the root of the decision tree.

Define the information gain of feature A

$$\begin{aligned} Gain(D,A) &= Entropy\ of\ decision\ attributes \\ &\quad - The\ average\ information\ expectation\ of\ feature\ A \\ &= Ent(D) - \textstyle\sum_{v=1}^{k} \frac{|D^v|}{D} Ent(D^v) \end{aligned} \tag{19}$$

Repeat the previous two steps to calculate $\mathrm{Gain(D,A)}$for all indices

Information gain of each index is compared, and the evaluation index with the maximum information gain is placed at the front of the decision tree.

Recurse the first classification feature is used as the new root node, and the recursive idea is adopted until the leaf node, which is the decision attribute.

3.3. Verify

Through literature review and data collection and analysis, we obtained the confidence interval of the mean value of the light pollution rate evaluation index for the following four different types of regions at the significance level of $\partial = 0.05$ .The median value of each interval was taken as the reference data of this region under the index, as shown in the following table.

Table 4. Reference data for each region

	L_n	D_p	G_a	C_f	P_c
Protected land	97.2	69	0.3	52.1	66.98
Rural community	552.7	146	1.77	71.3	53.40
Suburban community	2638	350	4.12	77.8	45.55
Urban community	8764	5691	7.07	82.4	29.58

The reference data of each region is taken as the test set and input into the established ID3 decision tree model, and the light pollution degree risk level of the four regions is

Table 5. Result

	Level	Value
Protected land	Mild	61.63
Rural community	moderate	932.24
Suburban community	Severe	1798.5
Urban community	Very severe	4150.6

The results were compared with the evaluation results of the four regions obtained through the formula of quantization fraction of light pollution degree established in Question 1 and the method of risk classification of light pollution degree. The results were consistent, which further proved the credibility of the two models.

4. Test the Model

- The evaluation model we developed is widely applicable. Because the selection of evaluation indicators is relatively comprehensive. At the same time, the selection of evaluation samples is relatively comprehensive, which comprehensively considers developed and developing countries, suburbs, cities, rural areas and protected areas, so as to make the model as universal as possible.

- Our model is intuitive and simple. In the process of model development, through analysis and screening, we cluster or discard the evaluation indicators that affect the degree of light pollution, which greatly improves the interpretability and accuracy of the model prediction results.

- Our model is sufficiently reliable. We used various algorithms and tools, such as entropy weight method, Topsis, principal component analysis, ID3 decision tree, MATLAB Curve Fitting Tool and SHAP interpretation model, to analyze the abstract and complex influencing factors of light pollution qualitatively and quantitatively. The model is based on strict mathematical derivation, the solution process is rigorous and rigorous, and the results are reliable and persuasive.

- Our model has both theoretical value and practical significance. Our model is based on big data analysis and uses algorithms such as SHAP, which has full theoretical connotation and practical guiding significance for light pollution standard evaluation and prevention in real life.

References

[1] Liu M. (2012). Study on Evaluation Indexes and Methods of Light Pollution in Urban Lighting Planning. (Dalian University of Technology).

[2] Liu T Y. Light pollution treatment: a two-way study of domestic practice and foreign experience [J]. Journal of Northwest University for Nationalities (Philosophy and Social Sciences Edition), 2022,(1): 109-116

[3] ZHANG C H. (2020). Research on the application of quantitative trading strategy model. (Master dissertation, Yunnan University of Finance and Economics).

[4] SUN,W.W. ,&Hu,Y. M.,&Liu,X.C. (2005).Soil quality grade evaluation based on decision tree . (South China Agricultural University).