Comprehensive Evaluation and Stepwise Regression Model of Light Pollution in Topsis Based on Entropy Weight Method

Qikai Sun [#, 1], Jiaming Yao [#, *, 2]

[1] Department of Artificial intelligence, Chongqing Jiaotong University, Chongqing, China, P.C: 402247

[2] Department of mathematics, Chongqing Jiaotong University, Chongqing, China, P.C: 402247

* Corresponding Author Email: Jiaming Yao, a18679310218@163.com

[#]These authors contributed equally

Abstract. Light pollution is a phenomenon that affects our environment. It is of great significance to make a quantitative analysis of light pollution and establish an evaluation system for energy saving, environment improvement and social and economic benefits. In order to establish a reasonable evaluation system, based on a comprehensive evaluation model of light pollution risk level based on the entropy weight method of Topsis analysis, combined a multiple linear regression model to test the significance of the indicator data and performed the White's test and the VIF test for multiple covariance on the model.Then this paper used the standard inverse stepwise regression model to solve the linear regression model, constructed a comprehensive evaluation model of light pollution.

Keywords: Light pollution, Topsis, Entropy method, Regression analysis, VIF test.

1. Introduction

Light pollution is a phenomenon caused by the excessive or poor use of artificial light, as a new type of urban construction pollution, will change our perception of the night sky, affecting our environment and our health, for the hazards of light pollution, the need to implement timely intervention to mitigate the negative impact of pollution[1]. As much as possible, the positive impact of human light, reduce to avoid the negative impact of excessive use of human light.

2. Materials and methods

2.1. Data acquisition and preprocessing

The data came from China Statistical Yearbook 2022 and were preprocessed:

- Step1: Harmonization of indicators to achieve positive indicators.

Since population density, GDP per capita, artificial light intensity, and very large indicators, no conversion is needed.

Seed plants, number of bird species, number of species, sleep quality, and SQM are very small indicators, so they need to be normalized to very large indicators.

Formula for converting very small indicators to very large indicators.

$$Max - x \tag{1}$$

- Step2: Normalize the forwarding matrix.

In order to eliminate the influence of 8 indicators with different scales, assume that there are (i =26) evaluated objects and (j =8) evaluated indicators, which constitute the forwarding matrix as follows:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} \end{bmatrix} \tag{2}$$

Then the matrix normalized to it is denoted Z, Each element of Z is:

$$z_{ij=} x_{ij} \Big/ \sqrt{\sum_{i=1}^{26} x_{ij}^2} \qquad i=1,2...26 \quad j=1,2...8 \tag{3}$$

2.2. Evaluation index construction method

This paper mainly establishes an evaluation system on light pollution, firstly, by using the Topsis method to synthesize the data of population density, GDP and other indicators, and by using the entropy weighting method to assign weights, [2]the comprehensive light pollution scores of Beijing, Shanghai and other cities are obtained and the light pollution is divided into four levels with the scores of 0.51, 0.45 and 0.37 as the dividing line, after that, the linear regression equation is constructed by stepwise regression model , using significance test to obtain the population density, artificial light intensity and other influential indicators, and through quantitative analysis, proposed that Shanghai and other places can effectively improve light pollution by reducing the range of strong light and other measures, this is shown in Fig1 below.

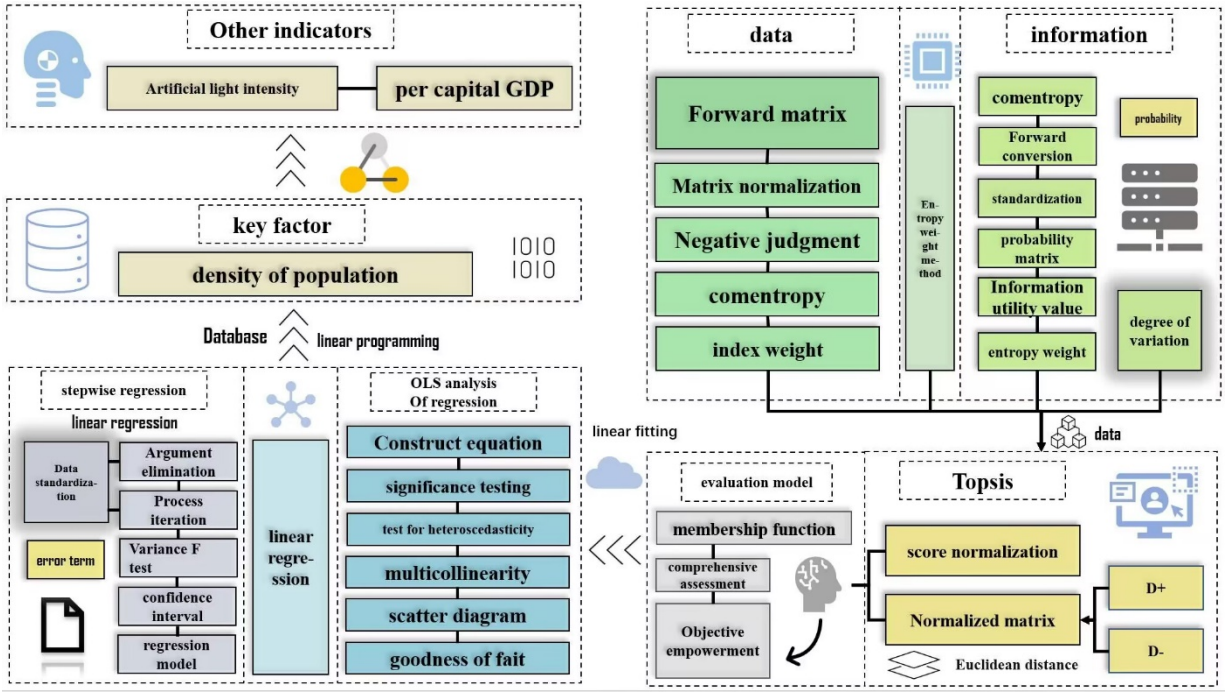


Fig 1. Our work

3. Modeling

3.1. Topsis comprehensive evaluation model of light pollution based on entropy weight method

The team will analyze the light pollution risk level based on the brightness of the influencing light by using the night light remote sensing image data of Luojia 1,this is shown in Fig2 below. 26 geographically representative areas in different provinces and cities in China were collected as sample data.

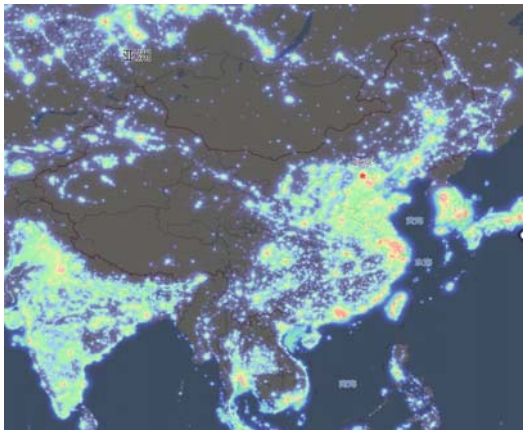


Fig 2. Luminous remote sensing image map

Based on our team's research and analysis, we reviewed the literature and selected eight indicators to evaluate the risk level of light pollution, namely, seed plants, number of bird species, population density, GDP per capita, number of species, sleep quality, SQM, and artificial light intensity, combined with 26 geographically representative areas in different provinces and cities across the country as samples to establish a comprehensive evaluation model based on entropy weighting method of topsis to score the risk level of light pollution, which divided into four risk levels to determine the risk level of light pollution in a location[3].

The data was pre-processed and visualized as shown in Fig3/4 below.

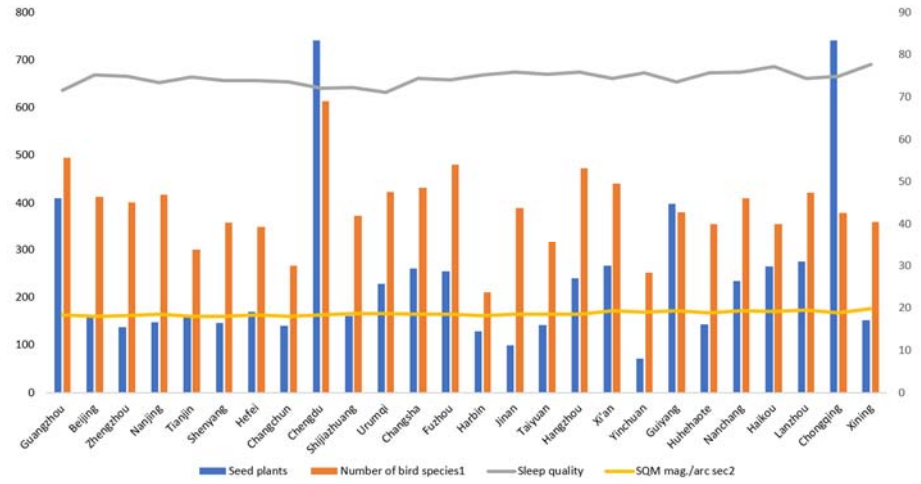


Fig 3. Data chart of each evaluation index

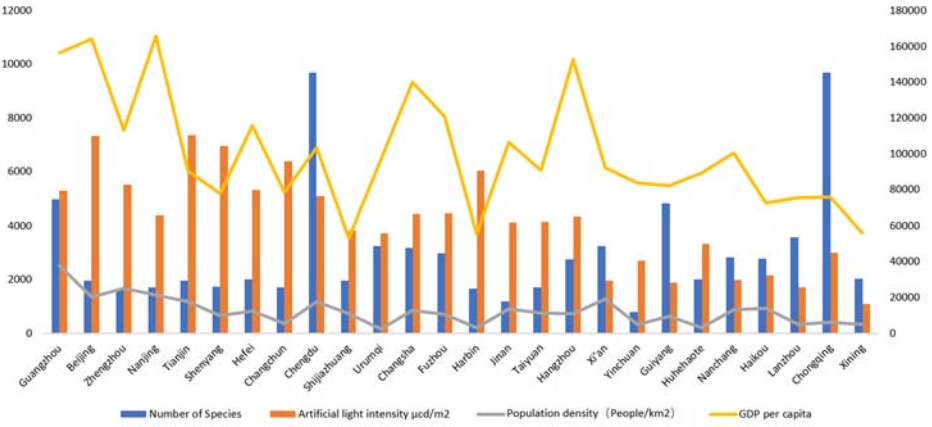


Fig 4. Data graph of each evaluation index

- **Step3:** Calculate the probability matrix based on the preprocessed data.

Denote the probability matrix as *P*

$$P_{ij} = \frac{z_{ij}}{\sum_{i=1}^{26} z_{ij}} \quad i = 1,2...26 \quad j = 1,2...8 \tag{4}$$

- **Step4:** The information entropy, information utility values of the eight indicators are calculated and normalized to obtain the entropy weight of each indicator.

The information entropy is calculated as in equation (5)

$$e_j = -\frac{1}{\ln 26}\sum_{i=1}^{26} P_{ij} \ln(P_{ij}) \tag{5}$$

The information utility value is as in equation (6)

$$d_j = 1 - e_j \tag{6}$$

The entropy weight of each indicator is as in equation (7)

$$W_j = d_j - \sum_{j=1}^{8} d_j \quad j = 1,2...8 \tag{7}$$

- **Step5:** The scores are calculated and normalized.

Define the maximum value as $Z^+ = \max(Z_1^+, Z_2^+, ...Z_8^+)$ (8)

Define the minimum value as $Z^- = \min(Z_1^-, Z_2^-, ...Z_8^-)$ (9)

Define the distance between i ($i = 1,2...26$) evaluation objects and the maximum value

$$D_i^+ = \sqrt{\sum_{j=1}^{8} W_j \ (Z_j^+ - z_{ij})^2} \tag{10}$$

Define the distance between i ($i = 1,2...26$) evaluation objects and the minimum value

$$D_i^- = \sqrt{\sum_{j=1}^{8} W_j (Z_j^- - z_{ij})^2} \tag{11}$$

Then we can calculate the resulting unnormalized score of the evaluation object:

$$S_i = \frac{D_i^-}{D_i^- + D_i^+} \tag{12}$$

We used MATLAB software to process the raw data. 5.1.2 Solving the model.
Results after normalization of normalization (see Appendix, Table 15 for specific data)
The entropy weighting method yields the weights in the following table1.

Table. 1 Indicator weighting table

weight	**Seed plants**	**Number of bird species**	**population density**	**GDP per capitaty**
W_j	0.10719	0.12241	0.13886	0.13543

weight	**Number of species**	**Sleep quality**	**SQM**	**Artificial light intensity**
W_j	0.10719	0.17088	0.10901	0.10904

The weights obtained from the combined entropy weighting method analysis were used to calculate the scores of light pollution risk levels for the above 26 regions, as shown in the following table2[4].

Table. 2 Scale of local light pollution risk level

City	Light pollution level score	City	Light pollution level score
Guangzhou	0.691173018	Jinan	0.417943217
Nanjing	0.538509977	Harbin	0.416660997
Zhengzhou	0.531791317	Taiyuan	0.414634775
Beijing	0.529319091	Xi'an	0.411976574
Tianjin	0.511723717	Hangzhou	0.402097924
Hefei	0.472387989	Yinchuan	0.373850146
Shenyang	0.470476633	Guiyang	0.368914738
Chengdu	0.466667463	Huhehaote	0.354520066
Shijiazhuang	0.461869489	Nanchang	0.352876545
Changchun	0.459854944	Haikou	0.335079057
Urumqi	0.446861774	Lanzhou	0.334831938
Changsha	0.441826179	Chongqing	0.303397231
Fuzhou	0.426123241	Xining	0.288024572

The higher the score, the greater the risk level of light pollution in the area, and the greater the degree of light pollution.

From the above figure, we can know the light pollution risk level score of each place, based on this data to define four light pollution risk level levels[5]. The light pollution risk level scale is shown in the following table3.

Table. 3 Light pollution risk level scale

Light pollution risk level	*Light pollution risk level score*
First level	Above 0.51
Second level	0.51-0.45
Third level	0.45-0.37
Fourth level	Below 0.37

3.2. Light pollution data fitting based on standardized backward stepwise regression model Model Preparation

3.2.1 Data Processing

First we start with descriptive statistics based on the data available to us, as shown in the following Table 4.

Table. 4 Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Seed Plants	30	275.8633	222.9267	70.7	1069.5
Bird species	30	410.8	115.6849	210	838
Population Density	30	912.7218	768.1986	161.058	3922.546
GDP per capita	30	102165.3	34070.84	52859	165681
Number of Species	30	3432.933	2907.755	785	13769
Sleep quality	30	74.46667	1.607382	71.1	77.6
SQM	30	18.564	0.5115669	17.83	19.83
Artificial light intensity	30	4292.667	1840.86	1090	7830

A. Assumptions of the linear model

- variable descriptions, as shown in Table 5.

Table. 5 Variable Description

Variable	Variable Description
a_1	Seed Plants
a_2	Bird species
a_3	Population Density
a_4	GDP per capita
a_5	Number of Species
a_6	Sleep quality
a_7	SQM
a_8	Artificial light intensity
b	Light pollution risk level score

- Constructing the equation

Since all nine variables are quantitative data, there is no need to set up dummy variables, assuming that light pollution indicators and ten factors obey the multiple regression model, and constructing the equation as

$$b = \beta_0 + \beta_1 a_1 + \beta_2 a_{_2} + \cdots + \beta_9 a_9 + \varepsilon \tag{13}$$

where ε is an unobservable and completely random perturbation term.

B. Solution of the model

i. Based on OLS regression analysis

First, the joint significance test of the regression coefficients before the ten independent variables was obtained $p = 0.0000 < 0.05$, Therefore, at the 95% confidence interval, the original hypothesis is rejected and we consider that β_i is not all 0.

The regression results show in the following Table 6.

Table. 6 Results of OLS-based regression analysis

Light pollution risk level score	Coef.	Std.Err.	t	P>t	[95%Conf.	Interval]
Seed plant	-0.0000176	0.000224	-0.08	0.938	-0.0004835	0.0004482
Bird species	-0.0000989	0.0000502	-1.97	0.062	-0.0002033	5.63E-06
density(people/km^2)	0.0000769	5.76E-06	13.34	0	0.0000649	0.0000889
GDP per capita	3.54E-07	1.35E-07	2.62	0.016	7.27E-08	6.35E-07
Number of species	-6.04E-06	0.0000171	-0.35	0.727	-0.0000415	0.0000295
Sleep quality	-0.0254155	0.0022515	-11.29	0	-0.0300978	-0.0207332
SQM $mag./arcsec^2$	0.0091254	0.0289651	0.32	0.756	-0.0511108	0.0693616
Artificial light intensity	0.0000188	8.00E-06	2.35	0.029	2.12E-06	0.0000354
_cons	2.037707	0.5244049	3.89	0.001	0.9471475	3.128267

For all regression coefficients β_i a significance t-test was performed, and as shown by the results in the table above, population density, GDP per capita, sleep quality, and artificial light intensity were significant, and next we performed the White's test for heteroskedasticity and the VIF test for multicollinearity on the model data.

ii. Heteroscedasticity test

First we plot the scatter plot of residuals versus fitted values by stata, this is shown in Fig5 below, and intuitively, we cannot directly determine whether heteroskedasticity exists.

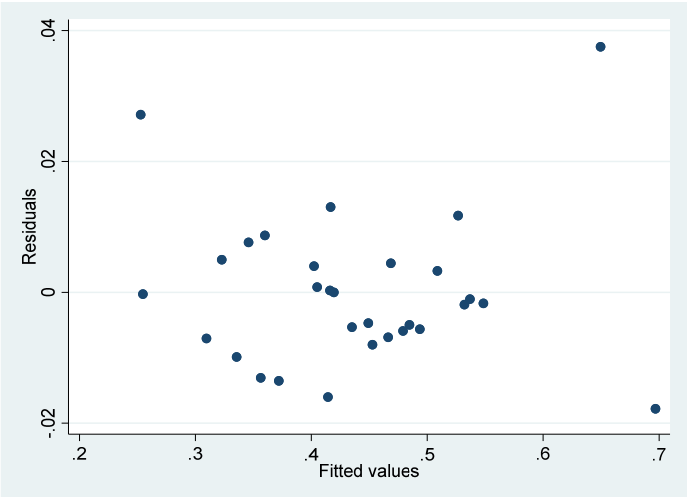


Fig 5. Scatter plot of residuals vs. fitted values

In order to confirm more accurately whether there is a variance problem, we performed White test on the OLS model by stata software, and obtained $\chi^2(29) = 30$, p=0.414>0.05, Therefore, we cannot reject the original hypothesis and can assume that there is no heteroskedasticity.

iii. VIF test for multicollinearity

To test for the presence of multicollinearity problems, we used the variance inflation factor VIF.

$$VIF_m = \frac{1}{1-R^2_{1-k/m}} \tag{14}$$

$R^2_{1-k/m}$ is the goodness-of-fit obtained by regressing the first m independent variable as the dependent variable on the remaining $k-1$ independent variables[6].

Define VIF of the regression model as $VIF = \max\{VIF_1, VIF_2 \ldots VIF_8\}$, The regression equation is considered to have multicollinearity if $VIF > 10$.

We tested the results by stata software as follows Table 7.

Table. 7 VIF test results

Variable	VIF	1/VIF
Seed plants	313.25	0.003192
Number of Species	309.54	0.003231
SQM $_{\text{mag./arcsec}^2}$	27.58	0.036255
Artificial light intensity μcdm^2	27.25	0.036701
Number of bird species	4.24	0.235621
GDP per capita	2.66	0.375576
Density(people/km^2)	2.46	0.406045
Sleep quality	1.65	0.60776
Mean VIF	86.08	

From the table, it can be seen that VIF=86.08>10, It can be considered to have multicollinearity, and since the OLS model cannot pass the multicollinearity test, we propose to take a backward stepwise regression for it.

iv. Standardized backward stepwise regression model solving

To explore the important factors affecting the light pollution risk level scores, we took a standardized backward stepwise regression approach[7].

Model principle: We first put all variables into the model, after which we try to raise one of the independent variables from the model to see if there is a significant change in the variance of light pollution risk level scores across the model. After that, the independent variable with the least explanatory power is removed, and this process is iterated until no independent variable meets the removal condition.

- F-test of variance for the model was obtained F(6.23)=332.78,p=0.0000<0.05, Therefore, at the 95% confidence interval[8-10], the original hypothesis is rejected and β_i is considered not all 0.
- Regression results by stata software as follows Table8.

Table. 8 Results of standardized backward stepwise regression analysis

Light pollution risk level score	Coef.	Robust Std	t	P>t	Beta
Artificial light intensity	0.0000164	2.26E-06	7.24	0	0.305236
Bird specie	-0.000098	0.0000438	-2.24	0.035	0.114966
Density(people/km^2)	0.0000774	7.74E-06	10	0	0.602746
GDP per capita	3.38E-07	1.05E-07	3.23	0.004	0.116826
Number of species	-7.42E-06	1.45E-06	-5.12	0	0.218912
Sleep quality	-0.0251418	0.0031817	-7.9	0	0.409798

- Obtain linear regression model

$$b = 2.197761+(-9.8a_2+7.74a_3+0.0338a_4-0.742a_5-2514a_6+1.64a_8)\times 10^{-5} \quad (15)$$

4. Conclusion

This paper establishes a comprehensive evaluation model of light pollution risk level based on the entropy weight method of Topsis analysis to determine the light pollution risk level of a location. Combined with the luminous image data of representative cities, four representative cities were selected to measure the light pollution risk level according to the evaluation model established in Model I. The idea of regression was put forward, and multiple linear regression models were constructed to test the significance of indicator data, so as to obtain indicators with greater influence. White heteroscedasticity test and VIF test of multiple covariance were conducted on the models. Since the data failed to pass the multivariate covariance test, we used the standard inverse stepstep-by-step regression model to solve the linear regression model, and developed three intervention strategies for population density, light pollution intensity and sleep quality respectively.

References（10 articles, and cited in the text）

[1] Pan Jialing.Research on the analysis method of curtain wall light pollution based on parametric model [J].Green building,2014,(3):10-12.

[2] Liu Chenchen.Study on Light Pollution Control of Landscape Lighting in Urban Leisure Square [J].Light source and illumination,2021,(8):62-63.

[3] Huang Ling.Light pollution, solid waste, carbon peak: focus on Shenzhen environmental protection keywords [J].Procuratorial situation,2022,No.667(23):13-15.

[4] Xu Cuiru.The "invisible" killer of human health - light pollution[J].Mathematics, Physics and Chemistry for Secondary School Students (Grade 8 Physics),2022,No.1313(10):19+31.

[5] Feng Kai, Hao Luoxi, Zeng Kun.Evolution characteristics of urban lighting pollution: A case study of VIIRS image [J].Journal of lighting engineering,2022,33(06):1-10.

[6] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.

[7] He Jia,Gao Ersheng,Lou Chaohua. Research on weighting coefficients and standardization methods in comprehensive evaluation [J]. China Public Health,2001,17(11):1048-1050.

[8] Li Jing. Discussion on the inclusion of light pollution in environmental impact assessment [C]// China Society of Lighting. China Lighting Forum 2020 -- Proceedings of Semiconductor Lighting Innovation Application and Intelligent Lighting Development Forum.[Publisher unknown.,2020:6.DOI:10.26914/c.cnkihy.2020.030160.

[9] Ying Shen, Tang Mo, Zhang Xinyue et al. Analysis of Light pollution of Glass Curtain Wall in 3D City [J].Wuhan University (Information Science Edition),2021,46(05):610-619.DOI:10.13203/j.whugis20200492.

[10] [Chen C N. Research on light pollution problems and countermeasures in urban nightview lighting [J]. Low-carbon world.,2021,11(05):18-19.DOI:10.16844/j.cnki.cn10-1007/tk.2021.05.009.