A study of light pollution risk level assessment based on multiple linear regression

Xiao Chen*

Chongqing University-University of Cincinnati Joint Co-op Institute, Chongqing University, Chongqing, China, 400044

* Corresponding Author Email: chen3x6@mail.uc.edu

Abstract. Light pollution is affecting modern people's lives, and it may damage people's health and safety. To develop an applicable metric to identify the light pollution risk level of a location, this paper first analyzes the correlation of factors affecting light pollution, and obtains that the correlation coefficient with light pollution mainly includes population density, per capita income, GDP, green coverage, the proportion of the tertiary industry and employment rate. Select these explanatory variables which have high linear correlation with the degree of light pollution to establish a multiple linear regression model. Then, carry out heteroscedasticity and multicollinearity tests on the model, and next, after removing the three variables of population density, per capita income and the proportion of tertiary industry, the optimal linear regression model is determined through OSL. Finally, different levels of light pollution are classified by DN value, which represents the night light of a location, to achieve the assessment of light pollution risk level.

Keywords: Light Pollution; Multiple Linear Regression Model; Risk level.

1. Introduction

In modern society, artificial light at night widely exists in people's production and life [1-2], but the light pollution caused by the bad use of artificial light is also seriously affecting the behavior and health of human and non-human organisms [3]. COMAP's Illumination Control Mission (ICM) wants to measure the risk level of light pollution so as to improve people's understanding of light pollution and develop effective intervention strategies to reduce the phenomenon of light pollution. Light pollution in different locations is closely related to population, economic level, area and other factors [4-5], and the differences of these factors in different locations lead to the differences in the severity of light pollution. China light pollution distribution is shown in figure 1.



Figure 1. Map of China light pollution distribution.
(Data from the web site http://wwwnoaa.gov/).

To develop an applicable metric to identify the light pollution risk level of a location [6], this paper performs linear regression on the degree of light pollution to obtain a model.

2. Light Pollution Assessment Model

2.1. The Data

2.1.1. Data Collection

The data source website is table 1:

Table 1. Data Source.

Database Names	Database Websites	Data Type
NOAA	http://wwwnoaa.gov/	Geography
GoogleScholar	http://scholar.google.com/	Academic Paper
ColoradoSchoolofMines	https://payneinstitute.mines.edu/	Geography
NationalBureauofStatistics	http://wwwstats.gov.cn/	Statistics

First, we downloaded DMSP/OSL Night Light Data and VIIRS/DNB Night Lighting Data at the National Oceanic and Atmospheric Administration (NOAA). It should be noted that the image data of DSMP/OSL and VIIRS/DNB is not comparable due to the difference in resolution and sensor parameters. The VIIRS/DNB image data adopted in our model was reduced the spatial resolution to 1000m during data processing, and used DMSP/OSL image data for reference for reference and return to correction[7-8]. Therefore, the VIIRS/DNB image data used in the model can be compatible with DMSP/OSL. Then based on the DMSP/OSL night lighting data and vector chart (China prefecture level, 2015) and ARCGIS software spatial analysis module extraction function, we superimposed night light data and grid data[9], and finally obtained China Night Lighting Imaging Data (1992-2013) .Figure 2 is the F182012 Chinese area night lighting image after the mask treatment.



Figure 2. China Night Lighting Image.
(Data from the web site http://wwwnoaa.gov/).

We use ArcGIS software to cut 34 provinces in China to obtain the corresponding regional boundaries. Then use the mask processing tool to superimpose the Chinese lighting data and the corresponding provinces to obtain the light data of various provinces. The following is a brief display in Beijing as an example [10].

Sensor imaging environmental differences, decreased detection performance (over time), and sensor's own performance differences will affect abnormal fluctuations of DN values (multi-sensor image data). Based on the comparability and continuity of night light data (long-term sequence), we follow the technical ideas of "Night Lighting Image → Data Correction of the F101992-F182013" in accordance with the technical idea of "acquiring the Chinese area F101992-F182013". Correction. Under the ARCGIS platform, we used the grid calculator to construct the map algebraic expression to complete the DN value correction of the multi-sensor's image data of the same year.

For the first small question, we use DN value as an indicator of the level of light pollution risk level in the region. The level of light pollution risks with high DN values is high. Conversely, the level of light pollution risk is low. The night light images we use ARCGIS software to process ash images are used. DN values (like metad gray value) represent the average lighting strength.

The DN value is essentially the accidental noise of clouds, fire light, and oil and gas burning in different time. Except for conditions, the number of all-like cloud-free observations (derived from visible light, near-infrared channel data) and the result corresponding to them in the whole year. The sum of the brightness of all grid lights in a country or region is its light brightness. Each grid represents 30 seconds (30 seconds of latitude grid network), 0-63, when a grid's DN value reaches 63, it will no longer continue to increase.

The "ceiling" problem of lighting data is derived from such phenomena, especially in developed countries or regions. Such situations are more likely to occur. At present, the results of literature collection, combing, and analysis have confirmed that this phenomenon has not yet appeared in China. Therefore, we use DN value as a widely used indicator to evaluate the level of light pollution risk in a region.

2.1.2. Data Analysis

The Pearson correlation coefficient between two variables is defined as the quotient of the covariance and standard deviation between the two variables:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

The above formula defines the overall correlation coefficient, which is usually represented byρ. The Pearson correlation coefficient can be obtained by estimating the covariance and standard deviation of the sample, which is usually represented by$r$:

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{2}$$

r Can also be estimated from the mean value of the standard scores of the sample points (X_i, Y_i), and the expression equivalent to the above formula is obtained:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma_X}\right)\left(\frac{Y_i-\bar{Y}}{\sigma_Y}\right) \tag{3}$$

Where$\frac{X_i-\bar{X}}{\sigma_X}$, $\bar{X}$ and σ_X are the standard score, sample mean value and sample standard deviation of X_i . The correlation analysis of these data is shown in the figure 3.

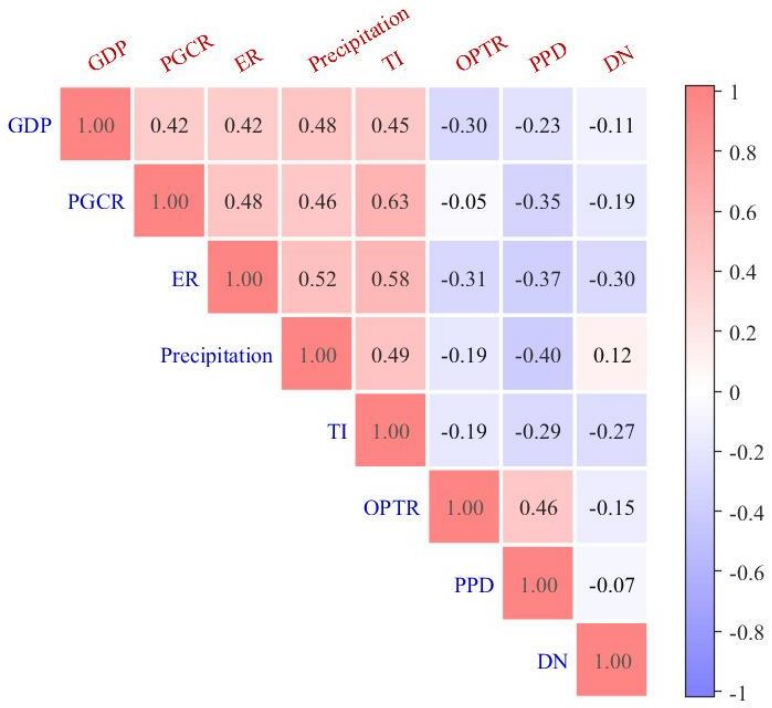


Figure 3. Correlation Analysis.

The correlation coefficients of GDP, green coverage, employment rate and the DN value of night light data are 0.787, 0.94 and 0.965, respectively, and the p value is less than 1%, which indicates that GDP, green coverage, employment rate and the DN value of night light data are significantly positively correlated. We will use the DN value of GDP, green coverage, employment rate and night light data to establish the in this part, the multiple regression model is used to analyze, taking Beijing as an example, to get the evaluation indexes of light pollution severity under the influence of economic development level, employment rate and green coverage rate. The levels of light pollution are graded by the radiance in different digital ranges, and an index is developed to assess the level of light pollution risk in different locations. Finally, we applied the model to four different locations: a protected land location, a rural community, suburban community, and an urban community to achieve the assessment of the light pollution risk grade.

2.2. Multiple Linear Regression Model

Suppose our model is as shown in equation (4):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu \tag{4}$$

μ Is a perturbation term that cannot be observed and meets certain conditions, and the perturbation term needs to be uncorrelated with all independent variables, making the model exogenous. $\hat{y}$ Is Obtained from regression analysis, and its formula is shown in equation (5):

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \widehat{\beta_k} x_{ki} \tag{5}$$

$\widehat{\beta_m}$ (m = 1, 2… k) indicates the influence of each additional unit of i on i when other independent variables. This is shown in equation (6):

$$\widehat{\beta_m} = \frac{\partial y_i}{\partial x_{mi}} \tag{6}$$

First, the multiple linear regression model requires us to process part of the data so that the data conforms to the linear relationship with the DN value. Next, we need to estimate all the regression coefficients β_1 = β_2 =, ……, = β_{30} from the known data and test the regression coefficients by the F statistic whether it is joint significant, if it passes the test, it indicates that the explanatory variables have an effect on the DN value.

After adding various explanatory variables linearly related to DN, and estimating all $\hat{\beta}$ by OSL, the goodness-of-fit R^2 is calculated by equation (7):

$$R^2 = 1 - \frac{SSE}{SST} SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \quad (7)$$

However, when more independent variables are introduced, the goodness-of-fit becomes larger. We tend to favor adjusted goodness-of-fit, which will actually decrease if the newly introduced independent variables reduce SSE particularly little. The adjusted goodness-of-fit is shown in equation (8):

$$R^2_{\text{adjusted}} = 1 - \frac{SSE/n-k-1}{SST/n-1} \quad (8)$$

By changing the type and number of explanatory variables and the selection of different regression coefficients in regression analysis, different fitting effects are obtained, and the optimal regression model can be obtained by comparison, so as to achieve a reliable assessment of the risk level of light pollution.

2.3. The solution of model

Firstly, the green coverage data is processed to obtain the $PGCR$ data, which has a linear relationship with the DN value. The transformation relationship is shown in equation (9):

$$PGCR = ln\, GCR \quad (9)$$

The factors with high correlation, including GDP, population density, proportion of tertiary industry, green coverage and employment rate, were added to the model as explanatory variables, and regression analysis was carried out, and the regression coefficient, corresponding standard error and p-value test results were obtained, as shown in table 2.

Table 2. First multiple linear regression.

Dnvalue	Coef.	Std. Err.	P>t
GDP	-0.0000502	0.0000259	0.064
PGCR	14.00188	3.436491	0
ER	0.095561	0.0212966	0
PPD	-0.0034786	0.0019639	0.089
TI	5.771397	2.52176	0.031
cons	-62.32316	3.944285	0

As can be seen from the table, the regression coefficients corresponding to population density and tertiary industry proportion receive the null hypothesis, which can be approximately equal to 0, so these two variables can be removed from the regression model. Modify the variables in the model and make linear regression again, and the results are shown in table 3.

Table 3. Changed Model.

Dnvalue	Coef.	Std. Err.	P>t
GDP	-0.0000733	0.0000188	0.001
PGCR	19.56385	0.9629892	0.000
ER	0.054977	0.0174223	0.004
cons	-58.77028	3.137015	0.000

Conduct joint significance test on this model:

$$F(3,27) = 546.07 \quad (10)$$

$$Prob > F = 0.000 \quad (11)$$

The results show that the model passes the joint significance test, and the regression coefficients are significantly different from 0. The goodness-of-fit R^2 of the model is calculated by OSL and the result is 0.9838. The adjusted goodness-of-fit R^2 result was 0.9820, which showed that the model had good fit performance.

White's test was performed on the remaining three explanatory variables to rule out the effect of heteroscedasticity, and the result was Prob>chi2=0.3783, the original hypothesis of White test is accepted, and there is no heteroscedasticity.

To test the multicollinearity of the model, in this section, we use the variance inflation factor (VIF):

$$VIF_m = \frac{1}{1-R^2_{1-\frac{k}{m}}} \tag{12}$$

Assuming that there are k independent variables, then the $R^2_{1-\frac{k}{m}}$ of the mth independent variable is the goodness-of-fit obtained by regression of the remaining k-1 independent variables using the mth independent variable as the dependent variable. The larger theVIF_m, the greater the correlation between the mth variable and the other variables. The VIF of the regression model is defined as shown in Equation (11):

$$VIF_m = max\{VIF_1, VIF_2, VIF_3, \ldots VIF_k\} \tag{13}$$

The results of this model are shown in table 4:

Table 4. VIF Of the model.

Variable	VIF	1/VIF
ER	7.71	0.129766
GDP	6.94	0.144029
PGCR	4.15	0.240790
MeanVIF	6.27	

As can be seen from the table, the VIF of this model is less than 10. Therefore, we do not consider this model to be multicollinear.

Finally, we get a linear regression model of DN values for per capita annual income, green coverage and employment rate, and the model formula is shown in Equation (12):

$$DN = -0.0000733GDP + 19.56PGCR + 0.055ER - 58.77 + \mu \tag{14}$$

Set the light pollution index according to different DN values, we divide the light pollution into 5 levels, as shown in table 5 below:

Table 5. Classification standards for light pollution of different grades.

Class	Degree of light pollution	DN
0	none	<5
1	Light	[5,9]
2	medium	[9, 15]
3	heavy	[15,20]
4	severe	>20

3. Conclusions

This paper first analyzes the correlation of factors affecting light pollution, and obtains that the correlation coefficient with light pollution mainly includes population density, per capita income, GDP, green coverage, the proportion of the tertiary industry and employment rate. Select these explanatory variables which have high linear correlation with the degree of light pollution to establish a multiple linear regression model. Then, carry out heteroscedasticity and multicollinearity tests on the model, and next, after removing the three variables of population density, per capita income and

the proportion of tertiary industry, the optimal linear regression model is determined through OSL. This model can get the risk level of light pollution from GDP, green coverage and employment rate as explanatory variables. The model formula is DN = -0.0000733GDP + 19.56PGCR + 0.055ER - 58.77 + μ, PGCR is the greening coverage after pretreatment, and ER is the employment coverage. Finally, different levels of light pollution are classified by DN value, which represents the night light of a location, to achieve the assessment of light pollution risk level.

References

[1] Liu Xiaoshuang. Research on analysis and prediction of urban nighttime light environment evolution based on time series [D]. Dalian University of Technology, 2022.

[2] Hou Xiaoli, Cheng Wanru. Research on factors influencing customer satisfaction of commercial.

[3] Chen Dongjie, Wang Yingxu, Guo Yanru. Research on the development of forest tourism experience products based on tourists' satisfaction--a case study of 49 national forest parks in Shandong Province [J]. Forestry Economics, 2021, 43(08): 62-79.

[4] Shi S, Huang X, Zhang M. Is immersion enough? -- A study of the effects of immersion and meaningful experiences on tourist satisfaction in tourism performing arts [J]. Journal of Tourism, 2021, 36(09):46-59.

[5] Yang Leyi, Chen Jieying, Zhang Luziyu, Zheng Yingying, Yu Yangfan, Fu Yingchun. Study on the sleep effect of nighttime light pollution based on NPP-VIIRS nighttime light data--a case study of Guangzhou City [J]. Tropical Geomorphology, 2021, 42(01):47-53.

[6] Liu Yuchuan. Construction and application of inversion method of urban nighttime light environment by correlation of remote sensing and actual measurement [D]. Dalian University of Technology, 2021.

[7] Yu Siyi. Research on the measurement and evaluation of the impact of urban three-dimensional spatial form on the habitat environment [D]. East China Normal University, 2020.

[8] Li Kunling. Single wood characteristics of dominant tree species and response to urban lighting in green areas of Fuzhou City [D]. Fujian Agriculture and Forestry University, 2020.

[9] Tian Jingjing. A brief exploration of the humanistic return of urban night landscape design--conceptual design of Jinji Lake lighting enhancement and transformation [J]. Urban construction theory research (electronic version), 2019(03):24-25.

[10] Yu Juan. Research on the evaluation of urban lighting color light interference to residents' indoor activities [D]. Tianjin University, 2017.