Study of Light Pollution Risk in China based on EWM-TOPSIS and Correlation Analysis

Ke Lin Maximilian Wong *

College of liberal arts and sciences, University of Illinois Urbana Champaign, Champaign 61820 USA

* Corresponding Author Email: max_4980@163.com

Abstract. In order to study the influencing factors of light pollution in China, this study aims to explore the main influencing factors of light pollution in China. First, 30 indicators, including population density, electricity output, land and marine protected area, and GDP per capita, were screened from five dimensions: economic development, population, ecology and geography, and 197 regional samples were selected from the global scale for analysis. Through the factor analysis dimensionality reduction process, the 30 indicators were grouped into five dimensions: air pollution level, biodiversity, natural resource storage, population density and economic development index, with a cumulative explanation rate of 97.89%. Subsequently, the entropy weight method (EWM) and TOPSIS model were used to assess the light pollution risk of the sample. The entropy weighting method determined the weights of the factors as air pollution level (0.212), biodiversity (0.346), natural resource storage (0.144), population density (0.282) and economic development index (0.016). Based on these weights, the light pollution risk index was calculated for each region. Finally, Spearman correlation analysis showed that population density (0.8193) and air pollution level (0.5101) were significantly and positively correlated with the light pollution index, while the economic development index (0.07903), biodiversity (-0.095), and natural resource stock (-0.02292) were weakly correlated. The study suggests that high population density and air pollution are the main drivers of light pollution, and it is recommended that these factors be prioritized in urban planning and environmental management to effectively control light pollution.

Keywords: EWM-Topsis; Light Pollution Index; Spearman Correlation Analysis.

1. Introduction

In recent years, environmental pollution has been increasing. The occurrence of countless tragedies has made people more and more aware of the importance of the environment to the survival and health of human beings. People are concerned about water pollution, air pollution, noise pollution, etc., and take measures to vigorously rectify. Gorgeous neon lights to our city to bring a fashionable modern atmosphere, light pollution also comes with it. The so-called light pollution, simply put, is the negative impact of light on human beings and the natural environment, most of which comes from inefficient, non-essential artificial light sources. As early as the early twentieth century, astronomers have found that the negative impact of urban outdoor lighting on astronomical observation is becoming more and more serious, so gradually put forward the concept of “light pollution” or “light pollution”. Our night street lamps, building lighting, car lights, outdoor advertising and large public places of light, as well as daytime glass, marble, walls and other reflections of daylight may be the source of light pollution. Light pollution has already jeopardized people's daily lives and the safety of life and property. The development of a broad indicator for judging the level of light pollution and making decisions on it is of great significance to the healthy development of the society.

2. Research Ideas

- **Data collection and factor analysis for dimensionality reduction:** We collected data on 30 indicators related to light pollution from around the world, including population density, electricity output, area of terrestrial and marine protected areas, GDP per capita, etc., and factor analyzed the data for dimensionality reduction down to five main factors: air pollution level,

biodiversity, natural resource storage, population density, and economic development index. The significance of this is to simplify the data structure, making subsequent analysis and modeling more efficient, while ensuring that we have a comprehensive understanding of the main factors affecting light pollution.

- **Establishment of Light Pollution Risk Assessment System:** Based on the dimensionality-reduced factors, the entropy weighting method was used to determine their weights, then the TOPSIS model was applied to calculate the light pollution risk index of each sample, and finally the light pollution index was graded by the trichotomous method, and the light pollution risk assessment system was established. The significance of doing so is to provide a scientific and reasonable light pollution risk assessment method for different regions, so that we can objectively compare the light pollution level of each region and guide the corresponding management and policy making.
- **Correlation analysis of light pollution influencing factors:** The Spearman model is used to analyze the correlation between the dimensionality-reduced factors and the light pollution index, and ultimately to study the intrinsic connection between different factors and the light pollution index. The significance of doing so is to reveal the specific contribution of each influencing factor to light pollution, to help us understand which factors affect light pollution to what extent, so as to provide a scientific basis for the formulation of effective light pollution control strategies.

3. Data Processing and Analysis

This section describes the data sources and the data processing process.

3.1 Data Source

In order to study the causes of light pollution in China, we first need to establish a widely applicable quantitative indicator. To this end, we collected indicator data from 197 regions around the world, and screened and processed them in order to lay the foundation for modeling. Through this approach, we hope to develop a scientific and reasonable quantitative indicator of light pollution, so as to deeply analyze the causes of light pollution in the target cities.

The main data sources used in this paper are listed below:

Table 1. Data sources

Database Names	Database Websites
United States Indicators Database	http://genderstats.worldbank.org/dataonline/
African Indicators Database	http://genderstats.worldbank.org/dataonline/ http://genderstats.worldbank.org/dataonline/ http://genderstats.worldbank.org/dataonline/
China Indicators Database	http://genderstats.worldbank.org/dataonline/

After the initial screening, 30 indicators related to the level of regional development, population, biodiversity and geospatial factors were selected for further data processing.

3.2 Data Processing

Due to the large number of selected indicators, a single indicator is less descriptive of the light pollution risk system, and at the same time, in order to improve the broad applicability of the light pollution index, the 30 indicators of the constituency were factor analyzed for dimensionality reduction and fusion.

3.2.1 Factor Analysis of Evaluation Indicators

Since the original dataset contains 30 light pollution impact indicators, and there is correlation between these light pollution impact indicators. In order to facilitate the subsequent analysis, these 30 evaluation indicators were therefore factor analyzed, i.e., dimensionality reduction processing.

There are n samples and p indicators available, then a sample matrix of size $n \times p$ can be formed:

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = (x_1, x_2, \ldots, x_p) \tag{1}$$

Where $x_1, x_2, \ldots, x_p \Rightarrow f_1, f_2, \ldots, f_p (m \leq p)$, and they satisfy:

$$\begin{cases} x_1 = u_1 + a_{11}f_1 + a_{12}f_2 + \ldots + a_{1m}f_m + \varepsilon_1 \\ x_{21} = u_2 + a_{21}f_1 + a_{22}f_2 + \ldots + a_{2m}f_m + \varepsilon_2 \\ \vdots \\ x_p = u_p + a_{p1}f_1 + a_{p2}f_2 + \ldots + a_{pm}f_m + \varepsilon_p \end{cases} \tag{2}$$

Due to $x = u + Af + \varepsilon$, it is now assumed:

$$\begin{cases} E(f) = 0 \\ E(\varepsilon) = 0 \\ Var(f) = I \\ Var(\varepsilon) = D = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2) \\ cov(f, \varepsilon) = E(f\varepsilon') = 0 \end{cases} \tag{3}$$

Factor analysis was performed using SPSS25 software, and the gravel plots corresponding to the eigenvalues of each factor were first plotted as shown below.

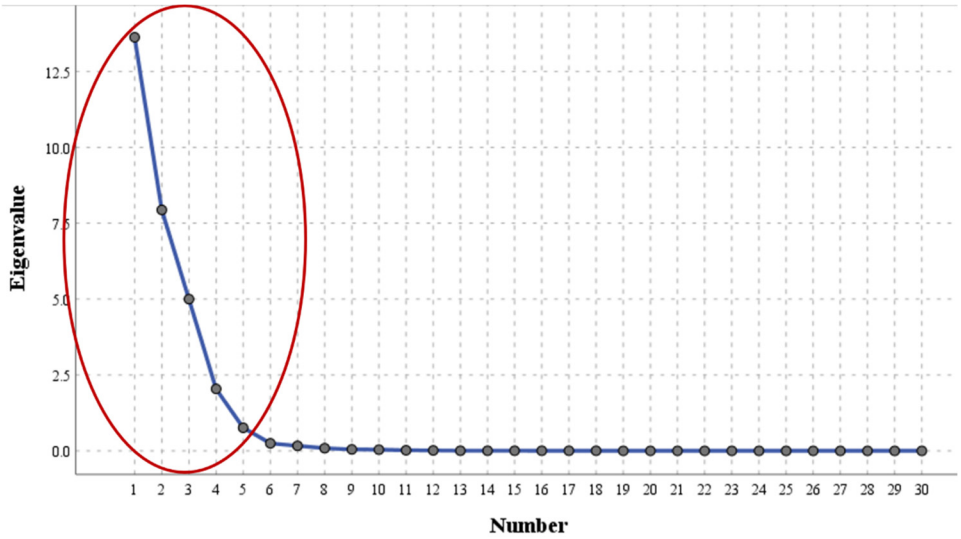


Figure 1. Factor analysis gravel plot

From Figure 1, it can be seen that the curves from component 1 to component 5 are steeper, while the curve obtained after component 5 is flatter, so component 1 and component 5 were retained for the analysis, and their explanation rate reached 97.89%.

Determine that the original component matrix is multiplied with the rotation matrix k after retaining the 5 components.

$$k = \begin{bmatrix} 0.638 & 0.657 & 0.383 & -0.007 & 0.121 \\ -0.526 & 0.050 & 0.790 & 0.311 & 0.020 \\ 0.404 & -0.368 & -0.038 & 0.836 & 0.040 \\ -0.372 & 0.656 & -0.466 & 0.452 & -0.097 \\ -0.120 & -0.002 & -0.107 & 0.005 & 0.987 \end{bmatrix} \tag{4}$$

The rotated component matrices are obtained and the partial component matrices are shown in the table below:

Table 2. Component matrix after rotation

	1	2	3	4	5
Population growth rate	0.504	-0.598	0.217	-0.445	0.132
	0.209	0.377	0.822	0.363	0.005
Population density	0.364	-0.879	0.257	-0.058	0.015
	0.526	0.845	0.017	-0.076	-0.015
PM2.5 concentration	…	…	…	…	…
	0.511	0.050	0.131	-0.116	0.835
Electricity output	0.809	-0.379	0.189	-0.359	-0.068

From the results of the factor analysis, it can be seen that component 1 has higher PM2.5 and greenhouse gas emissions, so component 1 is interpreted as the level of air pollution; component 2 corresponds to a higher land and marine protected area and forest cover, so component 2 is interpreted as biodiversity; component 3 has higher freshwater resource storage and forest area, so component 3 is interpreted as natural resource storage; component 4 has a higher population density and water productivity are higher, thus interpreting Component 4 as population density; Component 5 has higher population growth rate and GDP per capita growth rate, thus interpreting Component 5 as economic development index; based on this, each of these five indicators is evaluated and studied below.

3.2.2 Factor Analysis Test

The factor analysis model in the previous paper was tested. Firstly, KMO test and Butterlit spherical test were conducted. KMO statistic is between 0 and 1. When the sum of squares of simple correlation coefficients among all variables is much larger than the sum of squares of partial correlation coefficients, the closer the KMO value is to 1, the stronger the correlation among variables is, and the more suitable the original variables are to be analyzed for factor analysis. When the sum of squares of simple correlation coefficients among all variables is close to 0, the KMO value is closer to 0, indicating that the correlation among variables is weaker and the original variables are not suitable for cooperative factor analysis [1]. The statistic of Bartlett's test of sphericity is obtained from the determinant of the matrix of correlation coefficients. The null hypothesis should be rejected if the value is large and the corresponding p-value is less than the level of significance in the mind of the user. The correlation coefficient cannot be a unit matrix, i.e., there is a correlation between the original variables, which is suitable for factor analysis [2].

The results of the factor analysis model given by SPSS were as follows:

Table 3. Factor analysis model test

KMO Sampling Appropriateness Measure		0.892
BST test	Approximate cardinality	164.275
		30
		0.000

From the above analysis, it can be seen that the p-value of KMO sampling tangent and Butterlit sphericity test are 0.892 and 0 respectively, indicating that the original data is suitable for factor analysis.

4. Establishment of Light Pollution Indicator Assessment System based on EWM-TOPSIS

4.1 Weighting by EWM

EWM is based on the size of the variability of the indicator to objectively determine the weight, if the smaller the information entropy of an indicator, it indicates that the indicator deserves a greater degree of variability, the more information it provides, the greater the role it can play in the comprehensive evaluation, and the greater its weight [3].

Calculate the ratio of each indicator under each scenario, i.e., the share of the jth indicator in the ith dimension for that indicator, and calculate the magnitude of the indicator variance.

$$p_{ij} = \frac{x_{ij}}{\sum_i^n x_{ij}} \tag{5}$$

Next, the information entropy of each indicator is calculated and the information utility value is calculated and normalized to get the entropy weight of each indicator:

$$E_j = -ln(n)^{-1} \sum_{i=1}^{n} p_{ij} \, ln \, p_{ij} \tag{6}$$

Where $E_j \geq 0, if\ p_{ij}$*=0, define* $E_j = 0$.

The EWM method weights are derived as follows:

Table 4. Weights of the five integration indicators

Indicator Name	EWM weight
Air Pollution Level	0.210955
Biodiversity	0.346124
Economic Development Index	0.144350
Population density	0.282462
Natural Resource Stock	0.016109

4.2 Evaluation System Establishment by Topsis

As a basic evaluation method, TOPSIS can make full use of the original data, and the results can reflect the gap between the programs [4], and its basic evaluation process is:

(1) Normalization of the original matrix

(2) Normalize the normalization index

(3) Calculate the score and normalization

Since the indicators have been normalized during data processing, the next steps are as follows:

Step1: Suppose we need to have an evaluation object, an evaluation indicator, and a forwarding matrix as follows.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix} \tag{7}$$

Then the matrix for which the maximum and minimum values are normalized is denoted Z, for each element in Z:

$$z_{ij} = \frac{x_{ij} - min}{max - min} \tag{8}$$

Step2: Calculate the score, let us have n objects to be evaluated and m metrics standardized matrices for evaluation:

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix} \tag{9}$$

z_{ij} is finally recorded as the data from the jth sampling point of the ith indicator after the normalization and standardization process.

Define the maximum value:

$$Z^+ = (Z_1^+, Z_2^+, \cdots, Z_m^+) = (\max\{z_{11}, z_{21}, \cdots, z_{n1}\}, \max\{z_{12}, z_{22}, \cdots, z_{n2}\}, \cdots, \max\{z_{1m}, z_{2m}, \cdots, z_{nm}\}) \quad (10)$$

Define the minimum value:

$$Z^- = (Z_1^-, Z_2^-, \cdots, Z_m^-) = (\min\{z_{11}, z_{21}, \cdots, z_{n1}\}, \min\{z_{12}, z_{22}, \cdots, z_{n2}\}, \cdots, \min\{z_{1m}, z_{2m}, \cdots, z_{nm}\}) \quad (11)$$

The $i = (1,2,\cdots,n)$ evaluation object is at a distance from the maximum value:

$$D_i^+ = \sqrt{\sum_{j=1}^{m} (Z_j^+ - z_{ij})^2} \quad (12)$$

Distance of the $i = (1,2,\cdots,n)$evaluation object from the minimum value:

$$D_i^- = \sqrt{\sum_{j=1}^{m} (Z_j^- - z_{ij})^2} \quad (13)$$

The final score can be calculated as $y_i = \frac{D_i^-}{D_i^+ + D_i^-}$, where $0 \le y_i \le 1$.

The five indicators of air pollution level, biodiversity, natural resource storage, population density and economic development index after dimensionality reduction and fusion were used as model inputs, and the light pollution risk index was derived by combining the Topsis model.

The final light pollution risk index is obtained in the following table:

Table 5. Light Pollution Risk Index (top 20 groups)

Air Pollution Levels	Biodiversity	Economic Development Index	Population density	Natural Resource Stock	Light Pollution Index
-1.26316	1.5393	0.00158	-0.57491	0.06802	0.111916284
-0.6159	0.90132	0.31406	0.5871	0.02488	0.214517746
0.21102	0.68416	0.30039	0.18792	-0.30051	0.357031159
1.64549	-0.58722	0.51601	1.1114	0.02311	0.72142202
2.16623	0.52013	0.32307	0.17277	0.87612	0.258689846
2.16335	0.49563	0.33102	0.1637	0.92926	0.260131827
2.01389	0.59387	0.30479	0.1494	0.75948	0.247492071
1.86717	0.56871	0.31221	0.09705	0.87866	0.245031552
1.86444	1.0767	0.44851	-0.00057	-0.30518	0.115816735
1.86398	0.71245	0.35105	0.11445	0.22056	0.206522425
1.85313	0.70242	0.35171	0.10736	0.24504	0.206786962
1.84901	0.73346	0.34059	0.12038	0.21594	0.207371131
1.84441	0.66857	0.29259	0.11874	0.61959	0.23325917
1.84089	1.08645	0.4457	0.0134	-0.34966	0.117649148
1.823	0.75763	0.32783	0.12507	0.21834	0.208455234
1.81765	0.83645	0.41235	0.12415	-0.12383	0.181816381
1.81088	0.70711	0.33799	0.09998	0.29944	0.209008897
1.80072	0.77714	0.32025	0.12843	0.21494	0.20888323
1.79887	0.85289	0.40446	0.12197	-0.10458	0.181883085
1.79824	0.79113	0.34235	0.13396	0.13246	0.204526326

5. Judgment of Light Pollution Influencing Factors based on Spearman Correlation Analysis

5.1 Introduction to Spearman's Correlation Analysis

The commonly used correlation coefficient matrix method mainly contains three typical algorithms of Pearson, Spearman, and Kendall, due to the type of variables in our study, the existence of non-continuous variables and does not conform to the characteristics of the normal distribution, therefore, Spearman coefficients are selected for the mathematical modeling of redundant feature filtering.

Spearman's (Spearman) correlation coefficient: also known as Spearman's rank correlation coefficient, is a type of rank correlation coefficient. "Rank," or order, can be understood as an order or ordering [5], and is calculated based on the position of the variable within the data, with the formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{14}$$

In contrast to the Pearson correlation coefficient, the Spearman correlation coefficient has no linear constraints and does not need to be concerned with how the data varies and what distribution it conforms to, but only with the position of the corresponding values of each variable. Two variables are significantly correlated if their corresponding values, within each group, are arranged in the same or similar order.

5.2 Correlation Analysis of Regional Light Pollution Factors in China

For the five fusion indicators and the light pollution comprehensive assessment index after factor analysis dimensionality reduction, the correlation of variables within the class was solved, and the Spearman coefficients between the two were calculated to obtain the correlation coefficient matrix of variables [6]:

$$Matrix_\rho = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1j} \\ \vdots & \ddots & \vdots \\ \rho_{i1} & \cdots & \rho_{ij} \end{bmatrix} \tag{15}$$

where ρ_{ij} denotes the value of the Spearma correlation coefficient between the molecular descriptor variable i and variable j, which is plotted as a matrix chunked plot in terms of the magnitude of the correlation coefficient, as shown in the following figure:

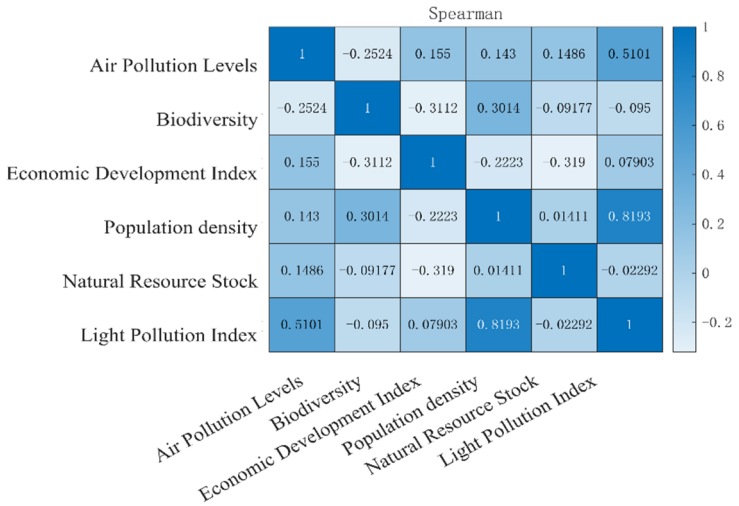


Figure 2. Color block diagram of correlation coefficients for each indicator

By analyzing the Spearman correlation analysis charts, it can be seen that the light pollution index has varying degrees of correlation with each factor. Among them, population density and air pollution level show a strong positive correlation with the light pollution index, indicating the importance of these factors in light pollution. In contrast, the economic development index, biodiversity and natural resource stock show a weaker correlation with the light pollution index, indicating that their influence is relatively small or more complex.

5.3 Interpretation of Correlation Analysis Results

The significant positive correlation (0.5101) between air pollution level and light pollution index may be due to the significant interaction between air and light pollution. High concentrations of suspended particulate matter can scatter and reflect light and increase nighttime brightness, thus exacerbating light pollution. In addition, industrial activities and dense traffic in urban areas are not only important sources of air pollution, but also important contributors to light pollution. Industrial lighting and traffic lights at night can appear more visible and widespread in the presence of air pollution.

Biodiversity was negatively correlated (-0.095) with the light pollution index, but the correlation was weak. This may be due to the fact that light pollution is mainly concentrated in more urbanized and industrialized areas, which tend to have lower biodiversity. Biodiversity reflects the health of the natural environment to a certain extent, but its direct relationship with light pollution is not significant, probably because the destruction of biological habitats and the increase in nighttime lighting during urbanization are not directly related.

The economic development index was positively correlated with the light pollution index (0.07903), but the correlation was weak. This may be due to the fact that economic development is usually accompanied by urbanization, which leads to an increase in nighttime lighting, thus increasing light pollution. However, the economic development index covers a wider range of economic activity data, making its direct correlation with light pollution insignificant. The increase in light pollution due to economic development may be diluted by data on other socio-economic activities.

Population density has a significant positive correlation (0.8193) with the light pollution index. This may be due to the fact that high population density implies more demand for nighttime activities and more nighttime lighting facilities, which directly leads to an increase in light pollution. Especially in large cities, areas with high population density tend to have a concentration of commercial and residential areas, which are more densely populated with nighttime lighting facilities, leading to serious light pollution problems. Dense populations also imply a higher frequency of living and commercial activities, which require significant nighttime lighting support.

Natural resource stock is negatively correlated with the light pollution index (-0.02292), but the correlation is very weak. This may be due to the fact that natural resource stocks are usually correlated with the health of the natural environment. Areas rich in natural resources tend to be less developed and have less need for nighttime lighting, thus less light pollution. However, the correlation between natural resource stocks and light pollution is weaker due to their complex relationship with urban development and human activities. In resource-rich areas, although there is less development activity, there is not necessarily a complete absence of night-time lighting demand.

5.4 Causality Determination

Akaike Causality is a method to show the strength of causality between multiple variables based on dividing the power spectral density of an optimal autoregressive model [7]. The denser the power spectral density, then the stronger the causality between two indicators [8]. The power spectrogram of each data sample indicator and light pollution index brought into the model is shown below:

It can be seen from the graph that the power spectral density of population density, economic development index and biodiversity and light pollution index are all at a high level, indicating that there is a high causality between light pollution index and these three indicators. That is to say, the

results of correlation analysis have practical significance, and there is causality between the influence indicators and light pollution.

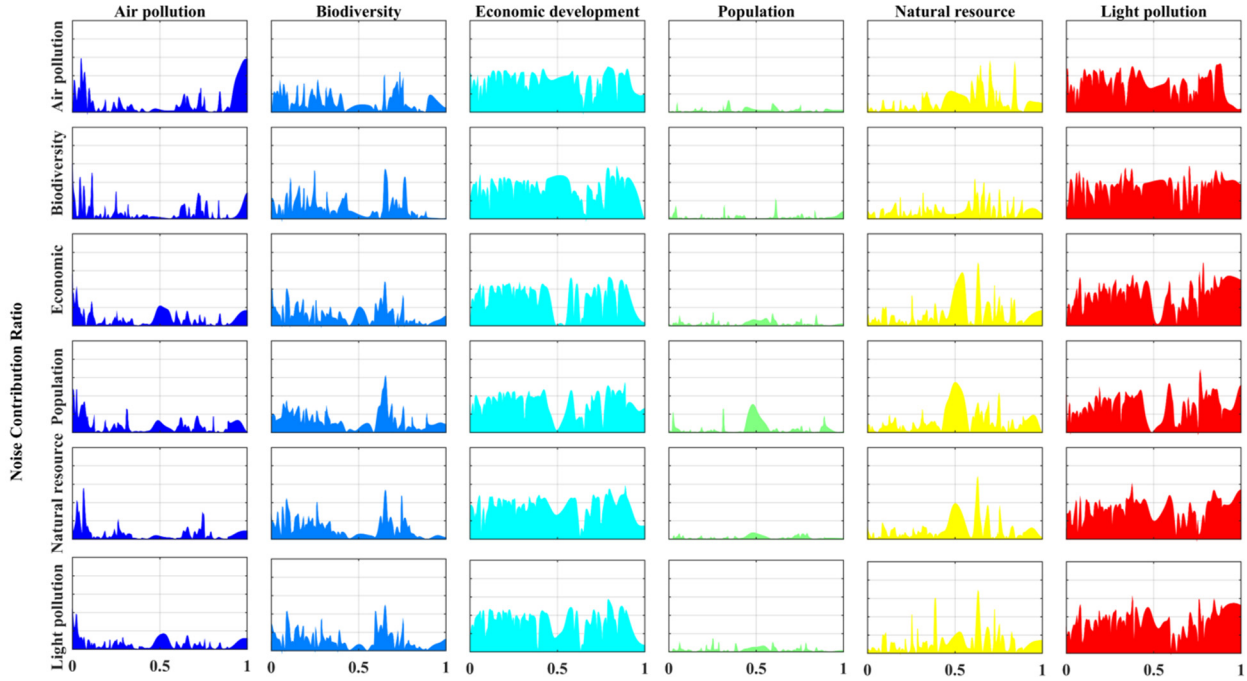


Figure 3. Causality spectrogram

6. Conclusion

In summary, the degree of light pollution is related to several factors. In this paper, a total of 30 indicators from 197 regions around the world are selected to establish the light pollution assessment index using the EWM-Topsis assessment algorithm, and the light pollution risk system is established. Spearman's correlation analysis is used to quantitatively analyze the influence of each index on the comprehensive light pollution assessment index. Then use Causality model to do causal analysis to prove that the model has practical significance, finally, this paper draws the following conclusions:

1. The use of EWM-Topsis assessment algorithm to establish the light pollution assessment index can reduce the weight fluctuation caused by data changes. It makes the system more interpretable.

2. Population density and air pollution level are the main influencing factors of light pollution, while the correlation between economic development index, biodiversity and natural resources stock and light pollution is relatively weak. Understanding the influence of these factors on light pollution can provide a scientific basis for urban planning and environmental protection, and help develop more effective light pollution control measures.

References

[1] CHEN Dongsu, LI Qiaoyuan, YAN Li, et al. Analysis of major meteorological factors affecting AQI in Hunan Province[J/OL]. Journal of Natural Sciences of Hunan Normal University,1-10[2024-07-25].

[2] YUAN LIGE, GENG MINGMING, LI FENG, et al. Water quality change and driving factor analysis of Dongting Lake from 2015 to 2020[J]. Yangtze River Basin Resources and Environment,2024, 33(07): 1576-1588.

[3] CHENG Qiyue. Structural entropy weighting method for determining the weights of evaluation indexes[J]. Systems Engineering Theory and Practice,2010,30(07):1225-1228.

[4] Lei Xunping, Qiu Guanghua. Empirical study on regional resource and environmental carrying capacity evaluation based on entropy weight TOPSIS model[J]. Journal of Environmental Science,2016, 36 (01): 314-323.

[5] HU Jun, ZHANG Chao, CHEN Pingyan. Comparison of non-parametric bivariate correlation analysis methods Spearman and Kendall's Monte Carlo simulation[J]. China Health Statistics,2008,25(06):590-591.

[6] Xu Weichao. A review of correlation coefficient studies[J]. Journal of Guangdong University of Technology, 2012,29(03):12-17.

[7] Wong Kin Foon Kevin, Ozaki Tohru. Akaike causality in state space. instantaneous causality between visual cortex in fMRI time series.[J]. Biological cybernetics,2007,97(2).

[8] Kin Foon Kevin Wong, Tohru Ozaki. Akaike causality in state space.[J]. Biological Cybernetics, 2007, 97 (2).