An Economic Model Processing Noise Method Based on Clustering in the Post-Epidemic Era

Tengyao Tu

School of Engineering Mathematics and Technology, University of Bristol, Bristol, UK

vs22203@bristol.ac.uk

Abstract. Since the 21st century, the world economic situation has undergone complex changes. Macroeconomic forecasting has become a hot topic of research for many scholars. An accurate macroeconomic forecast is of great significance to the country, enterprises, and individuals. However, in the post-pandemic era, some scholars have simply chosen all datasets for predicting macroeconomics. But, as the impact of the epidemic gradually decreases and the macroeconomic situation gradually returns to normal, the economic data on the impact of the epidemic is noisy.

The article uses TOPSIS scoring to discuss the extent to which the UK's macroeconomy has been affected after the outbreak of the epidemic, quantify the impact of the epidemic on the economic sector, and cluster the affected and unaffected intervals using clustering algorithms. We find that the period from Q1 2020 to Q1 2022 is the range of the impact of COVID-19 on the macro-economy. Moreover, the second quarter of 2020 is the period when COVID-19 has the greatest impact on the macro-economy. And the data after Q1 2022 is in the low-impact area, which is the post-pandemic period.

At the same time, we compared macroeconomic volume prediction methods using different datasets. When the impact of the epidemic on the economic volume was significant and intuitive, the model that excluded data from the epidemic period showed a significant performance improvement compared to the model with a complete dataset. When the impact of the epidemic on the economic quantity is not intuitive, removing data from the epidemic period will also improve the effectiveness of the model.

Keywords: Macroeconomics, Outlier, TOPSIS, K-means, Regression.

1 Introduction

Since the 21st century, prediction tasks have played an important role in the economic field, playing a crucial role in government policy formulation, enterprise strategic decision-making, and individual investment decision-making. However, COVID-19, a major global health event, has impacted many scholars' economic forecasts.

Xiao Dan used data from 1978 to 2022 for GDP analysis and forecasting in Sichuan Province[1], but in reality, the predicted values were far from the actual values published

K. Elbagory et al. (eds.), *Proceedings of the 9th International Conference on Financial Innovation and Economic Development (ICFIED 2024)*, Advances in Economics, Business and Management Research 281,
https://doi.org/10.2991/978-94-6463-408-2_24

in 2023. Similarly, Wang Zheng used data from 1995 to 2020 for GDP prediction and analysis in Guangdong Province[2], but the results were not ideal. Many articles have used data under the influence of the epidemic to predict future values, but the results are not ideal

We collected the GDP of the UK from the first quarter of 2010 to the third quarter of 2023[4], as shown in Fig 1:

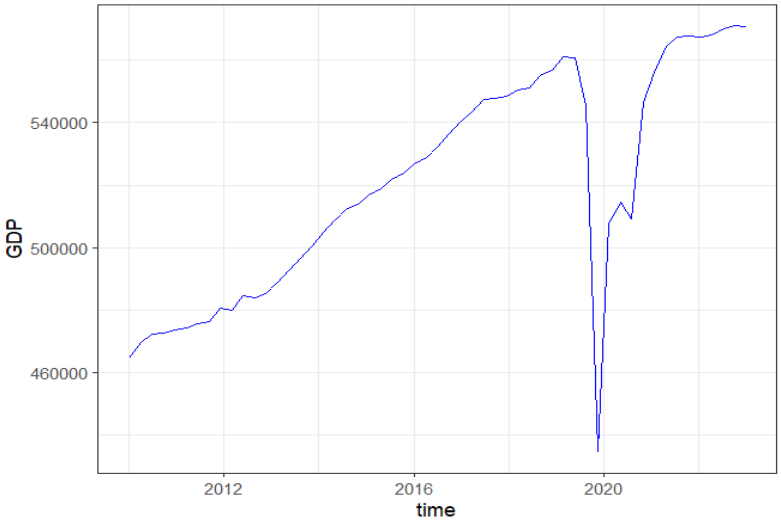

Fig. 1. GDP Time Series from Q1 2010 to Q3 2023[4]

It can be seen that although the COVID-19 epidemic has a great impact on Britain's GDP, with the end of the epidemic, the national economy is still in a normal state.

As Zhang Zhenghua said, the monotonous growth of GDP data is a symbol of social and economic stability in a country[3]. He tested the two prediction models proposed before and after the epidemic and summarized the cyclical law of the economy. Therefore, in reality, data from the epidemic is noise to the overall model, and using data from the epidemic can lead to significant errors in the prediction model.

So how to find the influence range of COVID-19 in the macro-economy and remove the noise is particularly important in the economic prediction model in the post-epidemic era.

The most direct impact of COVID-19 on the macro-economy is the trade volume of each country, that is, the phenomenon of anti-globalization[5], and the most serious impact on industries, especially tourism. So we can choose a country's monthly trade volume data and tourist number data as input variables to determine the probability of a country's economic model being noisy at that time. Then, cluster analysis is performed using k-means to determine the impacted period in the economic model.

The objective of the study is to:

*Using TOPSIS to comprehensively score the quarterly trade data and tourist number data of the UK from 2016 to 2023, to judge the extent of the impact of COVID-19 on the economy in this period.

*Using K-means to score different levels of impact, to classify the impact period of the epidemic and the post-epidemic period.

*Using the OLS-based linear regression method to compare different linear regression models with and without added noise during the epidemic period.

2 Descriptive statistics and missing value handling of model variables

2.1 Selection of variables and descriptive statistics

To discuss the most intuitive and direct impact of COVID-19, We choose total imports[6], total exports[7], and number of tourists[8] from 2016 Q2 to 2023 Q3 quarterly as input variables for the model.

Table 1. Variable descriptive statistics

Variable	mean	median	std	N
Total Import	181644.1	174466.5	27566.74	30
Total Export	174993.1	170096	24401.86	30
Tourists	7563.034	9322	3885.154	29

From Table 1, It can be seen that the standard deviations of the three variables we selected are all large, indicating that trade and tourism in the UK have undergone significant changes from 2016 to 2023. Meanwhile, there are missing values in the number of tourists.

2.2 Moving average method for handling missing values

The data for the number of tourists in the UK in the third quarter of 2023 is missing, but we do not want to give up all the data for the third quarter of 2023, so we need to fill in the missing values for the number of tourists in the UK in the third quarter of 2023.

As they are time series, using data statistics such as mean, median, or even trimmed mean is unreasonable. We use a simple moving average method for forward filling. As it is quarterly data, we have chosen a time window size of 4.

For moving average method:

$$S_i = (\sum_{j=i-1-m}^{i-1} S_j)/m \tag{1}$$

*(1) S_i represents the missing value after filling, and m represents the size of time window. S_{na} represents data before filling.

According to the time series method, the number of tourists in the third quarter of 2023 is estimated to be 9279.25.

The image after missing value processing is shown in Figure 2:

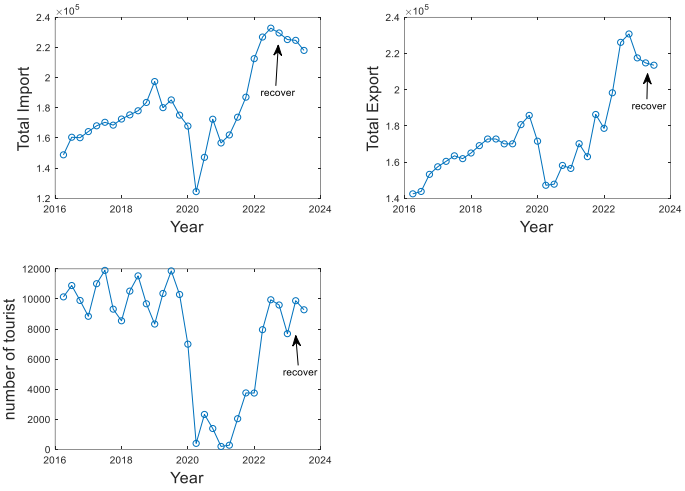

Fig. 2. Import and export trade data and tourist data

From Fig 2, It can be seen that these three indicators have significantly decreased during the epidemic, but in the post-epidemic era, these three indicators have gradually recovered, indicating that the UK's economic status has returned to normal.

So how do we quantify the data and periods affected by the epidemic?

3 Classification Outlier Interval Based on IQR Method

we first thought of using IQR to judge the outlier when judging the impact of the epidemic. The influence range of the COVID-19 epidemic situation can be judged by the outliers.

$$IQR(Interquantile\ range) = quantile_{0.75} - quantile_{0.25} \tag{2}$$

$$x_i > quantile_{0.75} + 1.5IQR\ or\ x_i < quantile_{0.25} - 1.5IQR \tag{3}$$

*(2)(3) x_i represents outliers, $quantile_{0.75}$ is a value with a quantile of 0.75, $quantile_{0.25}$ is value with a quantile of 0.25.

The result calculated through R is as follows:

Table 2. IQR Method result

Variable	Quantile25	Quantile75	IQR	count
Import	165153.5	194806	29652.5	0
Export	158744.5	184534.5	25790	2
Tourists	3750	10313.75	6563.75	0

Also, We made a box diagram based on the results, as shown in Figure 3:

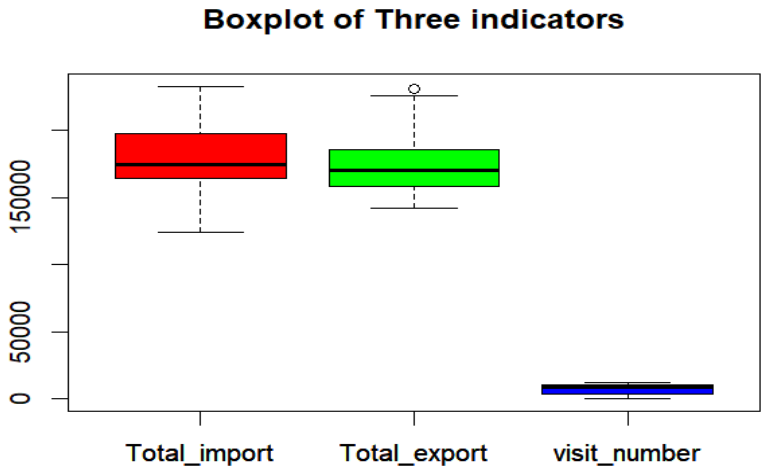

Fig. 3. Boxplot of three indicators

From Table 2, It can be seen that the IQR method is not feasible in finding the impact of the COVID-19 epidemic, because the post-epidemic era has not lasted for a long time, and the amount of data from the post-epidemic era is relatively small.

The IQR method has a high tolerance, so we need a more sensitive method to classify time intervals between epidemic and non-epidemic periods.

4 Score of influence degree of COVID-19 based on TOPSIS

For the TOPSIS method, we only need to select data from after the epidemic (Q1 2020). as there is no need to discuss the scores of normal economic growth before the epidemic.

In terms of the impact of COVID-19, the lower the trade volume, the fewer the number of tourists, and the greater the impact of COVID-19. So the total import value, total export value, and number of tourists are the minimum indicators.

For the TOPSIS algorithm[9], it is necessary to convert all minimum indicators into maximum indicators.

$$S_i = S_{max} - S_i \tag{4}$$

*(4) S_{max} represents the maximum number of the indicators.

when discussing multiple indicators comprehensively, the dimensions of different indicators are different. Simple addition is extremely unreasonable. Our preferred approach is to standardize different indicators. standardize using TOPSIS's standardization method[9]:

$$S_{sta} = S_i / \sqrt[2]{\left(\sum_{i=1}^{m} S_i^2\right)} \tag{5}$$

*(5) S_{sta} represents the data after standardizing, S_i represents the original data. After transformation and standardization, all data is shown in Figure 4:

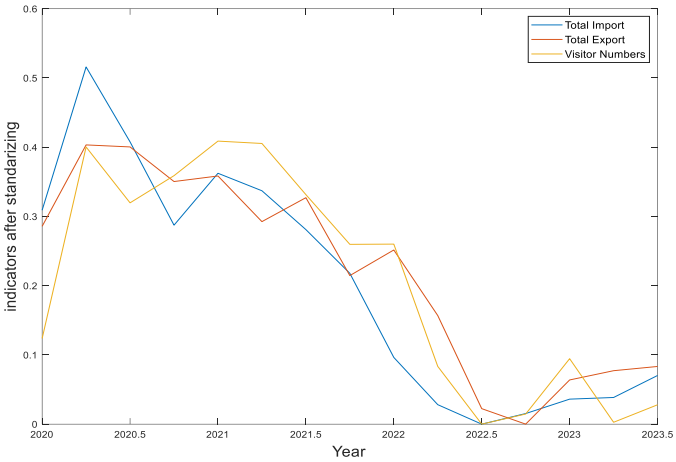

Fig. 4. Indicators after transformation and standardization

From Fig 4, It can be seen that the three indicators are rising rapidly in 2020 Q2, which means that they will be greatly affected by COVID-19 in 2020 Q2. Then calculate the distance between the maximum and minimum solutions(S_1, S_2) at different times:

The distance calculation method adopts Euclidean distance:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{6}$$

*(6) X=(x_1, x_2, x_3,…, x_n), Y=(y_1, y_2, y_3,…, y_n)

Calculate TOPSIS score(7):

$$Score_i = \frac{D_2}{D_1 + D_2} \tag{7}$$

*(7) D_i represents the distance from different indicators to S_i , i represents time step.

We calculated the TOPSIS score as shown in Figure 5:

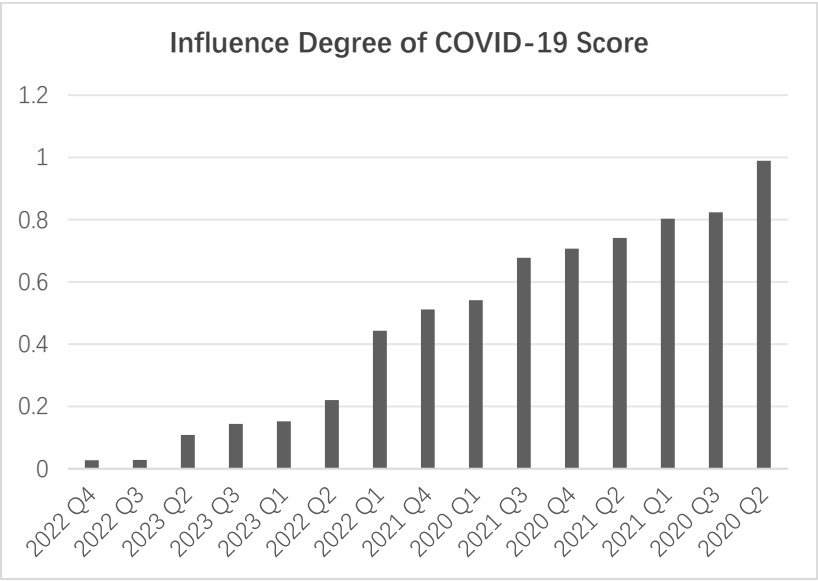

Fig. 5. Influence degree of COVID-19 Score

From Fig 5, It can be seen from the figure that after 2022 Q1, the impact of the COVID-19 epidemic on the economy will be significantly reduced. Then we use the clustering algorithm to classify the scores that are greatly affected and those that are less affected.

5 K-means clustering to confirm the influence range of COVID-19

After we quantified the impact of COVID-19 through TOPSIS if we use IQR for classification, we will also face the same problem, that is, half of the data set is affected data, so the tolerance of IQR is too high. So we use k-means to classify the degree of impact of COVID-19, to determine the range of significant impact of COVID-19.

The steps of the k-means algorithm are as follows:

Step 1:Joint each column X_i into a matrix X as a dataset.

Step 2: Initialize k category centers C_1, C_2 (k selected in the article is 2).

Step 3: Label each sample as the category closest to a certain cluster center. The distance calculation method adopts Euclidean distance(6)

Step 4: Update the center C_1, C_2 to the mean of all samples belonging to that category.

$$C_i = (\mu_1, \mu_2, \mu_3, \ldots, \mu_n) \tag{8}$$

*(10) μ_i Represents the mean of all sample points belonging to this category.

Step 5: Repeat steps 2 and 3 until the termination condition is reached(The termination condition of the article is that the cluster center is no longer changing, or the iteration number is equal to 10).

We obtained clustering results using K-means and visualized them as shown in Figure 6:

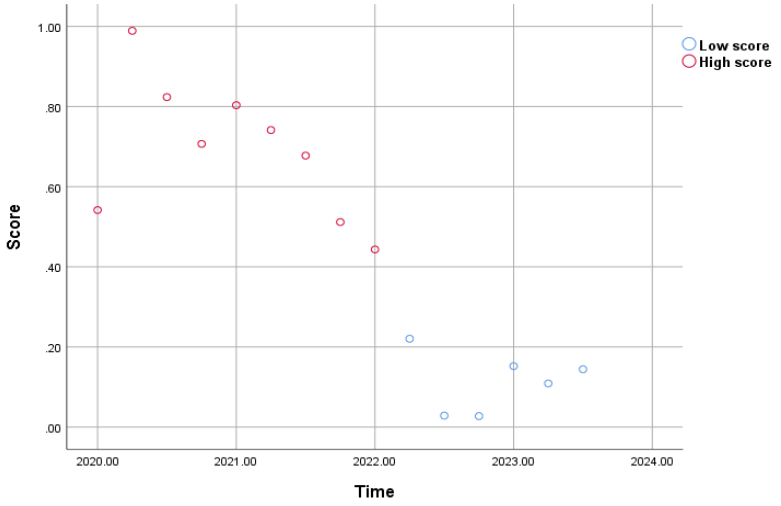

Fig. 6. Visualization of clustering result

From Fig 6, It can be seen that COVID-19, which began at the end of 2019, will have an impact on the British macro-economy from 2020 Q1 to 2022Q1. After 2022 Q2, the British economy will return to normal.

Next, we will conduct a linear regression on common macroeconomic quantities in the UK. And compare the effects of direct time series modelling and models that did not participate in the epidemic abnormal interval.

6 Fitting macroeconomic quantities using linear regression methods

First, we directly use the linear regression method of OLS to model the UK's GDP data from 2010 to 2023 and establish two models respectively. One is the linear regression model based on the original time data(model 1), and the other is the linear regression model based on deleting the abnormal interval of COVID-19 pneumonia(model 2).

Table 3. Regression results

	Model 1	Model 2
MSE	405094340	50620586
R-Squared	0.6763	0.9572
Adjust R-Squared	0.6702	0.9562

The visualization results of fitting different linear regression models are as follows:

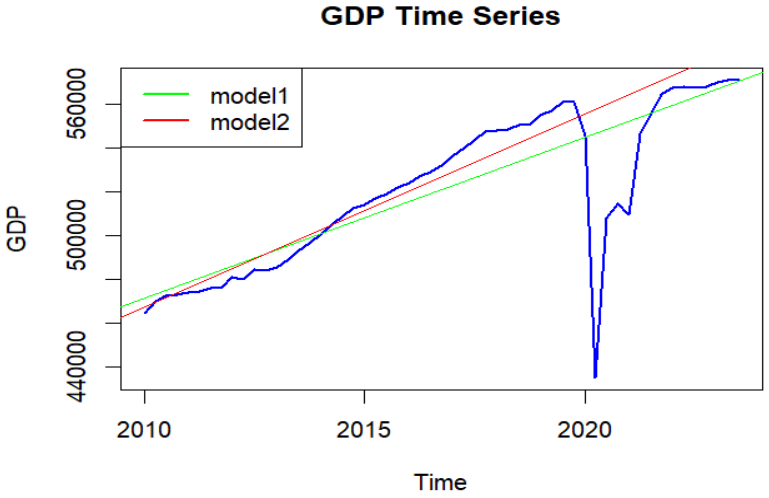

Fig. 7. Fitting results of different linear regression models

From Table 3 and Fig 7, It can be seen that after removing the abnormal interval, the MSE of Model 2 is much smaller than that of Model 1, and the R-squared reaches 95, which is nearly 30% higher than the R-squared of Model 1.

Because R-squared reaches 95, We believe that the model better reflects the laws of macroeconomic GDP:

$$y = 24.18x_i + 11400 + \delta \tag{9}$$

For the sake of rigour, we use the Consumer Price Index(CPI) [10] as the macroeconomic quantity to be predicted. Perform the same OLS-based linear regression modelling:

Table 4. Regression results 2

	Model 1	Model 2
MSE	9.915283	8.759817
R-Squared	0.8927	0.9116
Adjust R-Squared	0.8907	0.9096

The visualization results of fitting different linear regression models are as follows:

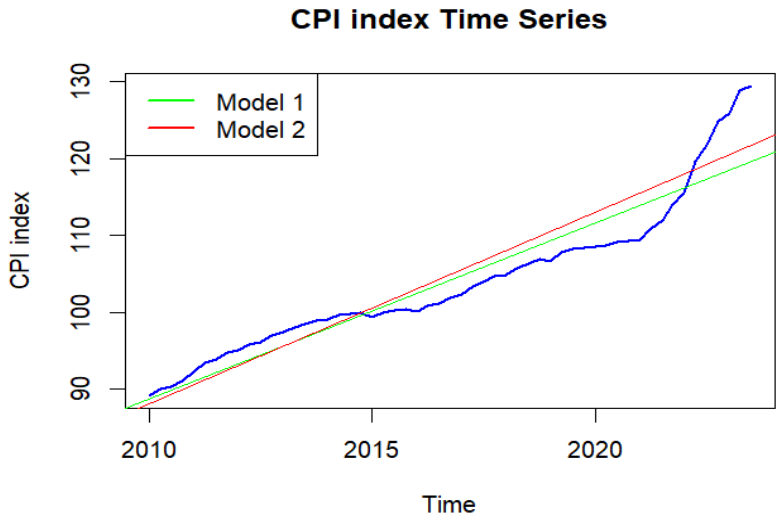

Fig. 8. Fitting results of different linear regression models 2

From Table 4 and Fig 8, the abnormality of the CPI index during the epidemic is not significant. However, after removing the data from the epidemic, the model's fitting degree still needs to be improved. Because the impact of the epidemic on CPI is reflected in the CPI rate[11].

$$CPI\ rate = \frac{value(year_{current})}{value(year_{prior})} 100\% \tag{10}$$

Due to the economic downturn caused by the epidemic, the CPI rate is close to 0, which is different from the normal healthy economic state of a country. So the degree of model fitting has increased.

So we believe that after excluding the abnormal range of the epidemic, the effectiveness of the macroeconomic forecasting model is improved. For the macroeconomic volume that has been greatly impacted by the epidemic, the degree of improvement is high. In terms of macroeconomic quantities that are not directly impacted by the epidemic, there has also been a certain increase.

7 Conclusion

The article first uses the IQR method to determine outliers, but due to its low sensitivity, the IQR method cannot accurately determine the range of the epidemic, and the IQR method cannot comprehensively judge the comprehensive value of a country's international trade and tourism vitality in the economic model.

Therefore, this paper proposes a TOPSIS-based scoring method for the impact degree of the COVID-19 epidemic, quantifying the impact degree of the epidemic in different periods into a number with an interval of [0,1].

After quantifying the impact of the epidemic, we classified different levels of impact using clustering algorithms. It is divided into the high-impact zone of the COVID-19 epidemic and the low-impact zone of the new epidemic. We found that from 2020Q1 to 2022Q1 is the high-impact range of COVID-19. And the data after Q2 2022 is in the low-impact area, which is the post-pandemic period.

Finally, we compare the fitting effect of the model in the complete time series dataset and the dataset excluding the COVID-19 epidemic interval. It is found that the fitting degree of the macroeconomic quantity that is greatly affected by the epidemic has been greatly improved after removing the impact range of the COVID-19 epidemic. For GDP, the fitting degree has been improved by nearly 30% after removing the impact range. For the macroeconomic quantity that is not directly affected by the epidemic, the fitting degree still has some improvement after removing the impact range. For the CPI index, after removing the influence interval, the goodness of fit increases by about 2%.

Therefore, the article believes that since we are already in the post-epidemic era, in the training process of the model for forecasting macroeconomic quantities, the data during the COVID-19 epidemic period is noise, and the data from 2020Q1 to 2022Q1 should be deleted before forecasting, which can better reflect the change cycle rate of macroeconomic quantities.

Reference

1. Xiao, D., (2023). Analysis and prediction of GDP in Sichuan Province based on the ARIMA model. Productivity Research. 2023(10):62-66.
2. Wang, Z. (2021). Prediction and Analysis of GDP in Guangdong Province Based on ARIMA Model. Modern Business. 2021(36):69-71.
3. Zhang, Z. H.; Duan; S. Q. China's Economic Trend and Future GDP Forecast. ContemporaryEconomics. 2023(06) :10~17.
4. Gross Domestic Product (2023). office for national statistics(UK government). https: //www.ons.gov.uk/economy/grossdomesticproductgdp/timeseries/abmi/pn2.
5. Yuan, L. M. Y.; Impact of COVID-19 on economic globalization. Modern Business Trade Industry. 2021(21) .7-9.
6. Total Trade Import(2023).office for national statistics(UK government). https: // www. ons. gov.uk/economy/nationalaccounts/balanceofpayments.
7. Total Trade Export(2023).office for national statistics(UK government). https: //www. ons. gov.uk/economy/nationalaccounts/balanceofpayments.
8. OS visits to UK(2023).office for national statistics(UK government). https: // www. ons. gov.uk/peoplepopulationandcommunity.
9. Robert Soczewica(2020). What is TOPSIS? A simple but powerful decision method. https: //robertsoczewica.medium.com/what-is-topsis-b05c50b3cd05.
10. CPI annual rate(2023).office for national statistics(UK government). https: // www. ons. gov. uk/economy/inflationandpriceindices/timeseries/l55o/mm23.
11. JASON FERNANDO. Consumer Price Index (CPI) Explained(2023). https://www.investopedia.com/terms/c/consumerpriceindex.asp#toc-consumer-price-index-cpi-formulas.