ISSN 2616-5775 Vol. 5, Issue 13: 30-37, DOI: 10.25236/AJCIS.2022.051305

Identification and Analysis of Glass Components by Fusing K-Means Clustering and Ridge Regression

Jingyu Zhang*

School of Electronic and Information Engineering, Guangdong Ocean University, Zhanjiang, Guangdong, China, 524088
**Corresponding author: szu19876602893@163.com*

Abstract: *The craft production of glass has a long history, and glass has been a valuable physical evidence of trade exchanges during the Silk Road. As traditional value crafts, the study and identification of the chemical composition of glass is of great significance. This paper mainly completes the classification problem by K-means clustering analysis and clustering tree, etc.; the statistical law study and prediction analysis of data are completed by integrating radial basis ridge regression and ARIMA differential autoregression, etc., which solves the problem of identification, classification and prediction of glass components. Firstly, the correlation analysis of the given data is carried out, and the Pearson correlation coefficient is calculated and visualized as the heat map representation, and performed statistical law fitting analysis of the data by ridge regression with fused radial bases, and finally predicted the chemical composition content before and after weathering by Autoregressive Integrated Moving Average (ARIMA) differential autoregression.*

Keywords: *Glass chemical composition; Ridge regression; Pearson correlation coefficient; K-means cluster analysis; ARIMA differential autoregression; regression prediction*

1. Introduction

1.1 Problem Background

The Silk Road was a channel for cultural exchange between China and the West in ancient times, of which glass was a valuable physical evidence of early trade exchanges. Early glass was often made into bead-shaped jewelry in West Asia and Egypt and imported into China, and our ancient glass absorbed its technology and made locally, so the appearance of glass products with foreign, but the chemical composition is not the same.

The main raw material of glass[1] is quartz sand, whose main chemical composition is silicon dioxide (SiO2). Due to the high melting point of pure quartz sand, in order to reduce the melting temperature, fluxes need to be added during refining. The fluxes commonly used in ancient times were grass ash, natural alkali, saltpeter and lead ore, and limestone was added as a stabilizer, which was converted to calcium oxide (CaO) after calcination. The main chemical composition of the fluxes added differed. For example, lead-barium glass is made by adding lead ore as a flux in the firing process, and its content of lead oxide (PbO) and barium oxide (BaO) is high, which is usually regarded as our own invented glass species, and the glass of Chu culture is mainly lead-barium glass. Potassium glass is made by using substances with high potassium content, such as grass wood ash, as a flux.It is mainly popular in Lingnan of China and other regions in Southeast Asia and India.

Ancient glass[2] is highly susceptible to weathering by the environment in which it is buried. During the weathering process, a large number of internal elements were exchanged with environmental elements, resulting in changes in their composition ratios, thus affecting the correct determination of their categories. For example, the artifacts in Figure 1 are marked with no weathering on the surface, and the color and ornamentation of the artifacts are clearly visible on the surface, but lighter weathering is not excluded locally; the artifacts in Figure 2 are marked with weathering on the surface, and large grayish-yellow areas on the surface are weathered layers, which are obviously weathered areas, and purple parts are generally weathered surfaces. In some of the weathered artifacts, the surface also has unweathered areas.

ISSN 2616-5775 Vol. 5, Issue 13: 30-37, DOI: 10.25236/AJCIS.2022.051305

1.2 Problem formulation

Archaeologists have classified these artifacts into two types: high-potassium glass and lead-barium glass, based on their chemical composition and other testing methods. Annex Form 1 gives information on the classification of these objects and Annex Form 2 gives the corresponding percentages of the main components (blanks indicate that the component was not detected). These data are characterized by compositionality, i.e., the sum of the component ratios should be 100%, but the sum of the component ratios may not be 100% due to testing methods and other reasons. In this problem, the data with the sum of components between 85% and 105% are considered as valid data. The following problems need to be solved by mathematical modeling.

(1) The relationship between the surface weathering of these glass artifacts and their glass type, decoration, and color was analyzed; the statistical patterns of the chemical composition content of the artifacts with and without weathering on the surface of the samples were analyzed in relation to the glass type, and the pre-weathering chemical composition content was predicted based on the weathering point detection data.

(2) Analyze the classification rules of high potassium glass and lead-barium glass based on the attached data; for each category, select the appropriate chemical composition to classify them into subcategories, give specific classification methods and classification results, and analyze the reasonableness and sensitivity of the classification results.

2. Problem Analysis and Model Construction

2.1 Analysis of Question

Firstly, we correlate the surface weathering of glass artifacts with their related properties, and we can calculate Pearson correlation matrix and heat map; then for high-dimensional data, we can fuse radial basis functions for multivariate ridge regression to obtain the statistical law of chemical content; finally, we can predict the chemical content before and after weathering based on ARIMA differential autoregression.

2.2 Model Assumptions

(1) Assume that the smooth data can be ARIMA differential autoregressed.

(2) Assume that the two indicators of strong correlation can be analyzed in place of each other.

(3) Assume that all study variables can be transformed from nonlinear space to linear space.

(4) Assume that the prior distribution of the transformed ridge regression parameters is normally distributed with mean 0.

2.3 Development and solution of a weathering model for the surface of glass artifacts

2.3.1 Model establishment

2.3.1.1 Surface weathering thermal map of glass relics based on Pearson correlation coefficient

First need on the surface of the glass weathering and glass type, grain and color were analyzed, and the relationship between the type of glass, decorative pattern and color of digital content, first of all to its digital dimension, high potassium and lead to type barium Numbers for 1 and 2, decorative design A corresponding number 1-3 - C, blue green, shallow blue, and purple three colors correspond to 1-3, At the same time whether weathering corresponds to 0 or 1.

Therefore, two of them are selected as an example. The sample data of the glass type is $X:\{X_1, X_2, \ldots, X_n\}$, Sample data for surface weathering of glass artifacts is $Y:\{Y_1, Y_2, \ldots, Y_n\}$, and the average sample value is

$$\bar{\mathrm{X}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X_i}\,,\ \bar{\mathrm{Y}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{Y_i} \tag{1}$$

Sample collaborative difference is

ISSN 2616-5775 Vol. 5, Issue 13: 30-37, DOI: 10.25236/AJCIS.2022.051305

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

Therefore, its Pearson related coefficient is

$$a_{XY} = \frac{Cov(X,Y)}{S_X S_Y} \tag{3}$$

Among them, SX is the sample standard deviation of X, $S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}$, In the same way $S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}}$

Based on this, we can establish a correlation matrix model for the relationship between the surface weathering of glass artifacts and their glass types, ornamentation and color, namely

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \tag{4}$$

Where a_{ij} represents the surface weathering of glass artifacts, glass type, ornamentation and color of the four pairwise correlation where i = 1, 2, 3, 4 ; j = 1,2,3,4.

2.3.1.2 Statistical Rule of Chemical Composition Content Based on Multiple Ridge Regression with Fusion Radial Basis Function

In this problem, we face the high-dimensional data of chemical composition content, including silicon dioxide, sodium oxide, potassium oxide and other chemical composition. In order to increase the credibility of the model and reduce the variance of the model, we decided to use ridge regression method to establish the model.

Ridge regression is a biased estimation, which is an improvement of least squares estimation. When the design matrix X is ill-posed, there is a strong linear correlation between the column vectors of X. We use regression diagnosis and independent variable selection to address this issue.

Let Xij be the value of the j-th variable in the i-th cultural relic sampling point sample, and y_i be the degree of weathering in the i-th sample, expressed as 0 or 1. In ridge regression, it adds the L2 penalty term to shrink the parameter values in the model, so as to achieve the purpose of parameter selection. The parameters selected by the model need to minimize the following equation:

$$\hat{\beta}_{ridge} = \underset{\beta}{argmin}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}{\beta_j}^2\right) \tag{5}$$

Among them, $\lambda \geq 0$, We call it the adjustment coefficient to control the degree of shrinkage.

When$\sum_{j=1}^{p}{\beta_j}^2 = 0$, $\lambda \to +\infty$, Using the least square method, we can get the best estimate of ridge regression:

$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1}X'Y \tag{6}$$

The ridge regression model of the above process can effectively control the model parameters and improve the reliability of the model due to the addition of L2 penalty term, but the ridge regression regards this problem as a linear problem. In order to solve this problem, we decided to introduce the radial basis function based on Gaussian kernel as the basic variable.

The essence of radial basis function RBF is to transform the original nonlinear space into a linear space, and then regress on the new linear space. This function is a kind of feedforward artificial neural network with good performance. It has a good nonlinear mapping ability and can solve the problem that there is not a linear relationship between different chemical composition combinations.

Set the original sample space to

$$X \subset \boldsymbol{R}^n, x = (x_1, x_2, \ldots, x_n)^T \in X, \tag{7}$$

The new sample space is

$$Z \subset \boldsymbol{R}^n, z = (z_1, z_2, \ldots, z_n)^T \in Z \tag{8}$$

ISSN 2616-5775 Vol. 5, Issue 13: 30-37, DOI: 10.25236/AJCIS.2022.051305

Consider a mapping from X to Z

$$\phi(x): X \to Z \tag{9}$$

Such that for all $x, z \in X$the function $K(x, z)$ satisfies the condition.

$$K(x, z) = \phi(x) \cdot \phi(z) \tag{10}$$

We do not define the mapping function $\phi(x)$ explicitly, I'm just going to define the function $K(x, z)$, that's our radial basis function.

In this paper, we choose the Gaussian kernel function as the radial basis function, that is

$$K(x, z) = exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right) \tag{11}$$

Based on this, the chemical composition of ridge regression with radial basis function can be solved.

2.3.1.3 Prediction of chemical composition content based on ARIMA differential autoregressive

Based on the previous paper, we have completed the summary of the chemical composition pattern by ridge regression, and then we can further predict its chemical composition content before weathering based on the regression results.

Pre-weathering data prediction by ARIMA based on weathering point detection data is first considered briefly for the bank's time series model ARIMA[3] $(p, 1, \mathrm{d})$。

Here make $W_t = Y_t - Y_{t-1}$, got:

$$\begin{aligned} W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + \cdots + \phi_{t-p} W_{t-p} \\ +e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_{t-q} e_{t-q}\ t = \pm 1, \pm 2, \ldots \end{aligned} \tag{12}$$

i.e.:

$$\begin{aligned} Y_t - Y_{t-1} = \quad & \phi_1(Y_{t-1} - Y_{t-2}) + \phi_2(Y_{t-2} - Y_{t-3}) + \cdots + \phi_{t-p}\left(Y_{t-p} - Y_{t-p-1}\right) \\ & +e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_{t-q} e_{t-q}\ t = \pm 1, \pm 2, \ldots \end{aligned} \tag{13}$$

Collated by:

$$\begin{aligned} Y_t = (1 + \phi_1) Y_{t-1} + (\phi_2 - \phi_1) Y_{t-2} + (\phi_2 - \phi_3) Y_{t-3} + \cdots + \left(\phi_p - \phi_{p-1}\right) Y_{t-p} - \phi_p Y_{t-p-1} \\ +e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_{t-q} e_{t-q}\ t = \pm 1, \pm 2, \ldots \end{aligned} \tag{14}$$

The specific steps of the whole algorithm are as follows.

1) Plotting the data and observing whether it is a smooth time series, and for non-smooth time series to be transformed into a smooth time series by first performing a d-order difference operation.

2) After the second step of processing, the smooth time series has been obtained. To find its autocorrelation coefficient ACF and partial autocorrelation coefficient PACF respectively for the smooth time series, the optimal stratum p and order q are obtained by analyzing the autocorrelation and partial autocorrelation plots.

3) From d, q and p obtained above, the ARIMA model is obtained.

4) The prediction inverse solution based on ARIMA[4] model was performed to obtain the prediction of preweathering data.

2.3.2 Model solving and analysis

2.3.2.1 Solution and analysis of surface weathering heat map of glass artifacts based on Pearson correlation coefficient

Based on the analysis, the Pearson correlation matrix can be built and solved to obtain the results corresponding to Figure 1 below.

ISSN 2616-5775 Vol. 5, Issue 13: 30-37, DOI: 10.25236/AJCIS.2022.051305

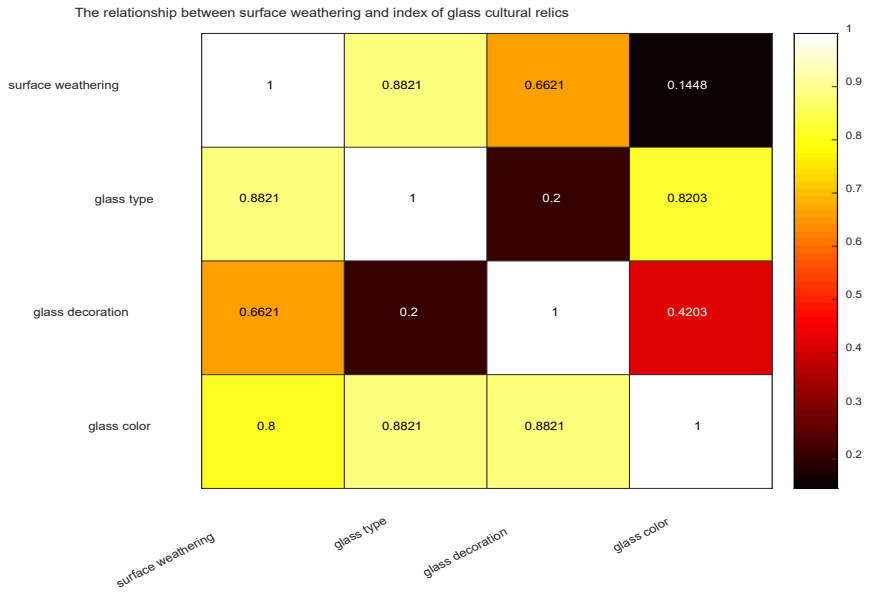

Figure 1: Thermogram of surface weathering of glass artifacts in relation to indicators

As shown above, where the lighter the color represents its stronger correlation, it can be seen that surface weathering has the greatest correlation with glass type, followed by glass ornamentation, while there is almost no relationship with glass color; also the correlation between glass type, glass ornamentation and glass color is smaller.

2.3.2.2 Solving and analyzing the statistical law of chemical content by multivariate ridge regression based on fused radial basis functions

The algorithmic steps for solving the radial basis function model for the statistical law of chemical composition content are as follows.

Input: Samples $\boldsymbol{x}$ requiring nonlinear transformation

Output: Sample $\boldsymbol{x}'$ after performing nonlinear transformation

Step1: Mapping the original data points x_j of sample $\boldsymbol{x}_i$ into a new feature vector x_j', i.e.

$$x_j' = exp\left(-\frac{\|x_j - z_j\|^2}{2\sigma^2}\right) \tag{15}$$

Step2: Calculate the dot product of $\boldsymbol{x}_i'$ and $\boldsymbol{x}_{i+1}'$, i.e.

$$\boldsymbol{x}_i' \cdot \boldsymbol{x}_{i+1}' = \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_{i+1}) \tag{16}$$

Step3: Stop if all samples have been traversed.

Otherwise, $i = i + 1$, Jump to Step1.

The solution of the multiple ridge regression model is similar to the description in Section 1.1.2 only when $\sum_{j=1}^{p} \beta_j{}^2 = 0$, $\lambda \to \infty$, and then using the least squares method, which leads to the best estimate of the ridge regression as

$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1} X'Y \tag{17}$$

In addition, the best estimate of $\hat{\beta}_{ridge}$ that we obtain at this point is the model parameter with the radial basis function $\phi(x_j)$ as the independent variable.

The final kernel function coefficient matrix can be solved as follows

[-0.66 0.22 0.47 -8.05 0.35]

The constant term is -78.9.

ISSN 2616-5775 Vol. 5, Issue 13: 30-37, DOI: 10.25236/AJCIS.2022.051305

The final analytical formula (with coefficients rounded to two decimal places) can be derived as

$$y=-78.97-0.72\phi(x_1)+0.11\ \phi(x_2)+0.46\phi(x_3)-8.02\phi(x_4)+\text{+}\ldots.\text{+}0.27\ \ \phi(x_{14})$$

where x_1-x_{14} denote silicon dioxide (SiO2), sodium oxide (Na2O), potassium oxide (K2O), calcium oxide (CaO), magnesium oxide (MgO), aluminum oxide (Al2O3), iron oxide (Fe2O3), copper oxide (CuO), lead oxide (PbO), barium oxide (BaO), phosphorus pentoxide (P2O5), strontium oxide (SrO), tin oxide (SnO2), and sulfur dioxide (SO2), in that order.

2.3.2.3 Solving and analysis of ARIMA-based differential autoregression for chemical content prediction

Through the ARIMA[5] differential autoregressive method, firstly, its smoothness was verified and followed by the prediction of the past chemical component content. The smoothness was verified as shown in Figure 2 below, containing ACF and PACF, while the component changes to potassium oxide as an example, and the results are shown in Figure 3 below. Similarly, the prediction results of other components can be obtained, and the differential results are shown in Figure 4.

Then, according to the weathering point test data, the content of its chemical composition before weathering can be predicted, for example, 54 severely weathered store, will get 16.77 of silicon dioxide (SiO2), 1.23 of sodium oxide (Na2O), 0.38 of potassium oxide (K2O), 2.88 of calcium oxide (CaO), 0.98 of magnesium oxide (MgO), 6.79 of aluminum oxide (Al2O3), Iron oxide 2.01 (Fe2O3), copper oxide (CuO) 0.81, lead oxide (PbO) 14.22, barium oxide (BaO) 7.88, phosphorus pentoxide (P2O5) 7.49, strontium oxide (SrO) 0.62, tin oxide (SnO2) 1.23, sulfur dioxide (SO2) 3.77.

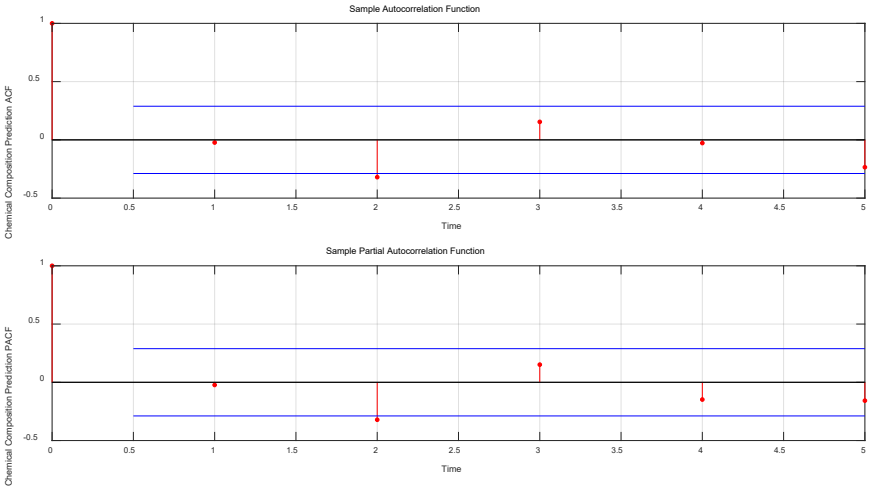

Figure 2: Chemical composition of ACF and PACF

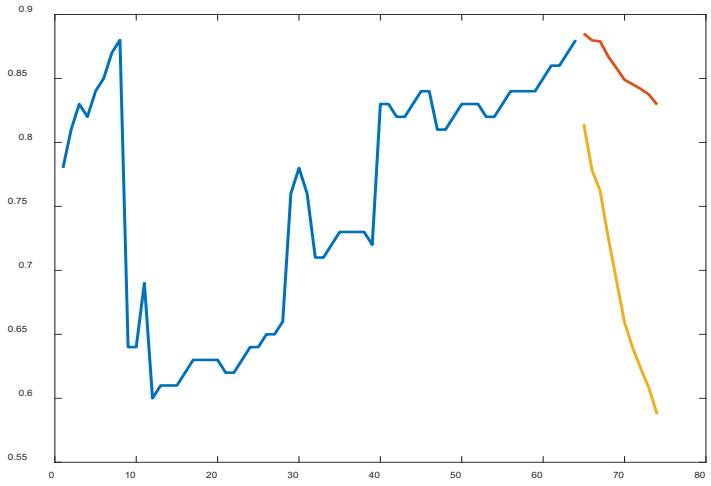

Figure 3: Changes in potassium oxide composition

ISSN 2616-5775 Vol. 5, Issue 13: 30-37, DOI: 10.25236/AJCIS.2022.051305

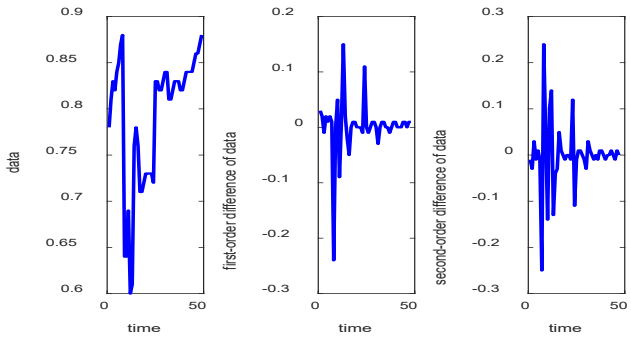

Figure 4: Differential results

3. Model improvement and extension

3.1 Advantages of the model

(1) The visualization methods such as heat map and clustering tree are used in the paper, which make the results more clear and visible, and at the same time, the model robustness, good adaptability and superior noise resistance can be found in the sensitivity test for some models.

(2) Correlation analysis can select several indicators with strong correlation to avoid repetitive and redundant analysis, thus reducing the 14 chemical composition indicators to 3, which reduces the workload of analysis and operation.

(3) The multivariate nonlinear ridge regression model with radial basis functions converts the nonlinear space to linear space by basis functions, and the final parameters are obtained in the ridge regression by adding a bias factor and using a great posteriori estimation to exclude the influence of high covariance, and adding an L2 regular term to greatly reduce the variance, which makes the final fitting results more accurate and reliable. The model takes into account a variety of factors, and is a strong guide for realistic data analysis.

3.2 Disadvantages of the model

(1) The ridge regression fitting process is carried out when the amount of data is too small and the number of polynomials of the established model is high may appear rank loss phenomenon, so the model has certain requirements for the amount of data.

(2) The amount of data data such as glass components of the problem are small, lacking certain test data, and the specific generalizability of the model is unknown.

3.3 Further improvement and extension of the model

The multivariate nonlinear ridge regression model incorporating radial basis functions can be simpler and more convenient for some multivariate complex and tedious nonlinear problems in linear space, while the model can be flexibly transformed to different scenarios, which can be analyzed in nonlinear problems by converting basis functions to linear space, and in linear problems directly by ridge regression model, while the model is more accurate and can be extended to many different scenarios for application, not only limited to the chemical field in this paper, but also to evaluation and classification problems in energy, environment, finance, etc.

References

[1] Prabhu Nimitha S, Sharmila K, Karunakara N, Almousa Nouf, Sayyed M I, Kamath Sudha D. Thermoluminescence Dosimetric Attributes of Yb3+ Doped BaO-ZnO-LiF-B2 O3 Glass Material After Er3+ Co-doping.[J]. Luminescence: the journal of biological and chemical luminescence, 2022, 37 (5).
[2] Ma Han, Liang Shunlin. Development of the GLASS 250-m leaf area index product (version 6) from MODIS data using the bidirectional LSTM deep learning model [J]. Remote Sensing of Environment, 2022, 273.

ISSN 2616-5775 Vol. 5, Issue 13: 30-37, DOI: 10.25236/AJCIS.2022.051305

[3] Gilmar Veriato Fluzer Santos, Lucas Gamalel Cordeiro, Claudio Antonio Rojo, Edison Luiz Leismann. A Review of the Anthropogenic Global Warming Consensus: An Econometric Forecast Based on the ARIMA Model of Paleoclimate Series [J]. Applied Economics and Finance, 2022, 9(3).
[4] Wenbo Zhang, Zhenyang Li. Research on Prediction Method of Photovoltaic Building Integration Plate Index Based on ARIMA Model [J]. Environment, Resource and Ecology Journal, 2022, 6(3).
[5] Spyrou Evangelos D., Tsoulos Ioannis, Stylios Chrysostomos. Applying and Comparing LSTM and ARIMA to Predict CO Levels for a Time-Series Measurements in a Port Area [J]. Signals, 2022, 3(2).