Analytical Chemistry: A Journal (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/analc.2022.010110
ISSN 2616-1958 Vol. 1 Num. 1

Component Analysis and Identification Model of Ancient Glass Products Based on Correlation Analysis

Yilei Wang[1], Wenxuan Liu[2], Ziqiang Lin[3]

[1]*School of Civil and Transportation Engineering, Beijing University of Civil Engineering and Architecture, Beijing, 102616, China*
[2]*School of Architecture and Urban Planning, Beijing University of Civil Engineering and Architecture, Beijing, 100044, China*
[3]*School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, 102616, China*

Keywords: Ancient glass products, Composition analysis and identification, K-Means, PCA, Sensitivity analysis

Abstract: To help ancient glass products to analyze and identify their components, this paper establishes a comprehensive evaluation model to help identify and analyze ancient glass products and their components, and classifies them according to the data, so as to clarify the correlation and sensitivity between their chemical elements. First, this paper makes a simple classification of the data, and then calculates several factors that account for a large proportion of the weight through principal component analysis (PCA). By reducing the dimension of data, the variables are reduced, making the classification basis more intuitive; At the same time, the factors that account for a large proportion of the main factors can be used as the intuitive basis for the division of subcategories. Finally, K-Means is used to further confirm the rationality and sensitivity of the relationship between specific factors and cultural relics.

1. Introduction

There are a number of relevant data on ancient glass products. Archaeologists have divided these cultural relics into two types: high potassium glass and lead barium glass according to their chemical composition and other detection methods. Ancient glass is easily weathered by the influence of burial environment [1]. In the process of weathering, a large number of internal elements will exchange with environmental elements, resulting in changes in the composition proportion of the glass products found, thus affecting the correct judgment of archaeologists on their categories [2].

Therefore, this paper needs to find out the classification rules of high potassium glass and lead barium glass according to the data. In the data, the table of detected components of most cultural relics is given. In this paper, a large number of irregular data are initially processed [3]. Then the principal component analysis method is used to reduce the dimension, so that it is easier to find the corresponding laws. Then, this paper uses the factors with high principal component weight as the basis for classification of subcategories. Finally, K-Means is used to analyze the rationality and sensitivity of the relationship between specific factors and cultural relics [4].

2. Model establishment and solution

2.1 Analysis of classification rules of high potassium and lead barium glasses based on PCA

2.1.1 Data preprocessing

(1) Abnormal data elimination

In this paper, the data between 85%~105% and the cumulative proportion of components are regarded as valid data, and the data that do not meet this range are excluded.

(2) Missing value data processing

In this paper, the high potassium is coded as 1, the lead barium is coded as 2, and the missing value is filled as 0 [5].

2.1.2 KMO and Bartlett's inspection

KMO test: 0.8 is very suitable for principal component analysis, 0.7-0.8 is generally suitable, and less than 0.6 is not suitable.

Bartlett test: if P is less than 0.05 and the original hypothesis is rejected, it means that principal component analysis can be done. If the original hypothesis is not rejected, it means that these variables may provide some information independently and are not suitable for principal component analysis.

Table 1: KMO Test and Bartlett Test of High Potassium Glass

KMO value		0.351
Bartlett's sphericity test	Approximate chi square	243.112
	df	91
	P	0.000***

Note: * * *, * * and * represent the significance level of 1%, 5% and 10% respectively

As shown in Table 1, the results of the KMO test show that the value of KMO is 0.351. At the same time, the results of Bartlett's spherical test show that the significance P value is 0.000 * * *, which is significant at the level. The original hypothesis is rejected. There is a correlation between variables. The principal component analysis is effective, and the degree is not appropriate.

Table 2: KMO test and Bartlett test of lead barium glass

KMO value		0.346
Bartlett's sphericity test	Approximate chi square	420.851
	df	91
	P	0.000***

Note: * * *, * * and * represent the significance level of 1%, 5% and 10% respectively

As shown in Table 2, the results of the KMO test show that the value of KMO is 0.346. At the same time, the results of Bartlett's spherical test show that the significance P value is 0.000 * * *, which is significant at the level. The original hypothesis is rejected. There is a correlation between variables. The principal component analysis is effective, and the degree is not appropriate.

2.1.3 Analysis variance interpretation table and gravel diagram

The variance interpretation table mainly looks at the contribution rate of principal components to variable interpretation. The explanation of total variance of high potassium glass and lead barium glass is shown in Table 3 and Table 4.

Table 3: Explanation of total variance of high potassium glass

Component	Eigenvalue		
	Eigenvalue	Variance interpretation rate（%)	Cumulative variance interpretation rate（%)
1	5.229	37.347	37.347
2	2.492	17.799	55.146
3	1.739	12.423	67.569
4	1.656	11.832	79.401
5	0.964	6.887	86.288
6	0.664	4.744	91.033
7	0.540	3.856	94.888
8	0.310	2.213	97.101
9	0.200	1.429	98.531
10	0.113	0.804	99.335
11	0.063	0.447	99.782
12	0.019	0.133	99.915
13	0.007	0.052	99.967
14	0.005	0.033	100

Table 4: Explanation of total variance of lead barium glass

Component	Eigenvalue		
	Eigenvalue	Variance interpretation rate（%)	Cumulative variance interpretation（%)
1	3.539	25.275	25.275
2	2.954	21.097	46.372
3	1.617	11.55	57.922
4	1.145	8.179	66.1
5	0.883	6.304	72.404
6	0.828	5.917	78.321
7	0.743	5.308	83.629
8	0.622	4.445	88.074
9	0.563	4.02	92.094
10	0.369	2.635	94.729
11	0.344	2.456	97.185
12	0.268	1.911	99.097
13	0.122	0.874	99.971
14	0.004	0.029	100

The function of the gravel map is to confirm the number of principal components to be selected according to the gradient of the characteristic value. The combination of the two can be used to confirm or adjust the number of principal components. The crushed stone diagram of high potassium glass and lead barium glass are shown in Figure 1 and Figure 2. Through analysis, this paper considers that it is appropriate to take two principal variables for analysis.

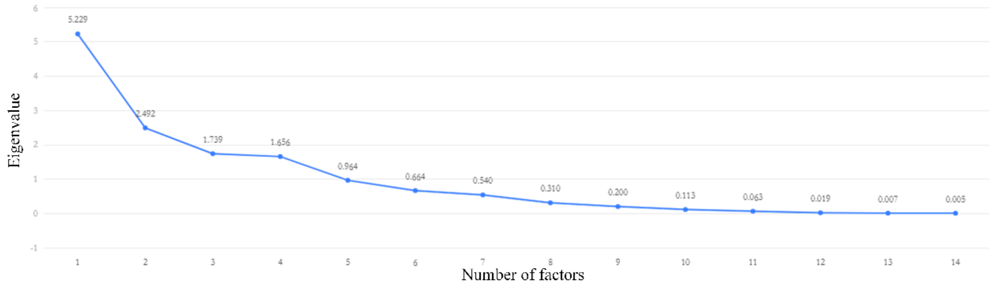

Figure 1: Crushed stone diagram of high potassium glass

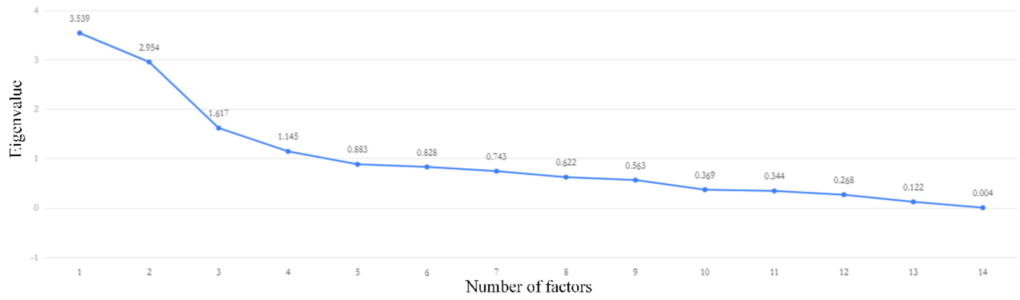

Figure 2: Gravel diagram of lead barium glass

2.1.4 Principal component load coefficient and thermodynamic diagram

The importance of hidden variables in each principal component can be analyzed by analyzing the load coefficient of the principal component and the thermodynamic diagram.

Table 5: Factor load coefficient of high potassium glass

	Factor load factor		Common degree
	Principal component 1	Principal component 2	
SiO2	-0.799	0.436	0.828
Na2O	0.032	-0.773	0.599
K2O	0.647	-0.525	0.695
CaO	0.595	-0.628	0.748
MgO	0.723	0.355	0.648
Al2O3	0.850	-0.009	0.722
Fe2O3	0.835	0.147	0.719
CuO	0.530	-0.092	0.289
PbO	0.402	-0.233	0.216
BaO	0.582	0.343	0.456
P2O5	0.686	0.551	0.775
SrO	0.731	0.499	0.783
SnO2	-0.089	0.339	0.123
SO2	0.287	-0.195	0.120

Based on the above research, two principal components are determined in this paper. As shown in Table 5, the factor load coefficient of Al2O3 in principal component 1 is large, so principal component 1 can be defined as recessive aluminum. The factor load coefficient of P2O5 in principal component 2 is large, so principal component 2 can be defined as recessive phosphorus.

Table 6: Factor load coefficient of lead barium glass

	Factor load factor		Common degree
	Principal component 1	Principal component 2	
SiO2	-0.878	-0.325	0.877
Na2O	-0.351	-0.409	0.290
K2O	-0.322	0.201	0.144
CaO	0.291	0.783	0.698
MgO	-0.292	0.686	0.556
Al2O3	-0.582	0.362	0.470
Fe2O3	-0.237	0.575	0.387
CuO	0.546	-0.395	0.455
PbO	0.601	0.421	0.538
BaO	0.626	-0.463	0.605
P2O5	0.504	0.569	0.578
SrO	0.558	0.231	0.365
SnO2	-0.285	0.362	0.212
SO2	0.528	-0.198	0.318

As shown in Table 6, the factor load coefficient of principal component 1 BaO is large, so principal component 1 can be defined as recessive barium. As shown in Figure 3 and Figure 4, the factor load coefficient of CaO in principal component 2 is large, so principal component 2 can be defined as recessive calcium.

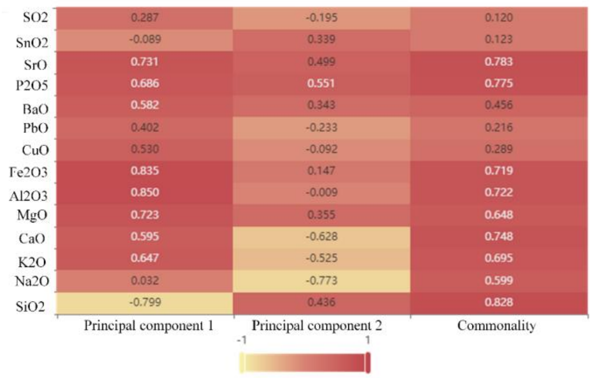

Figure 3: Principal component heat map

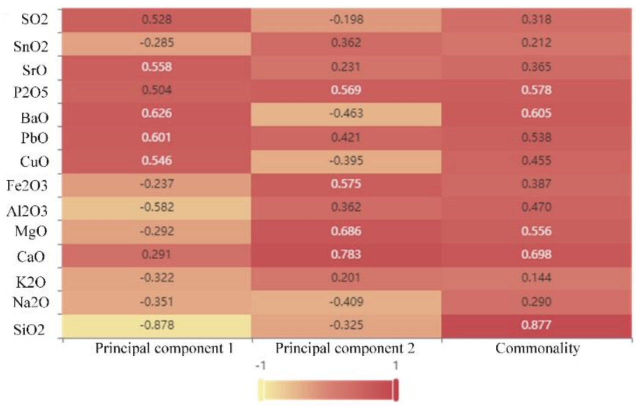

Figure 4: Principal component heat map

2.1.5 Dimension reduction analysis of related variables

Based on the principal component load diagram, the spatial distribution of principal components is presented through quadrant diagram by reducing the dimensions of multiple principal components into double principal components or three principal components.

In conclusion, the classification of high potassium glass and lead barium glass mainly depends on calcium oxide (CaO), barium oxide (BaO) and aluminum oxide (Al2O3) as the main components.

2.2 Classification of subclasses based on K-Means

2.2.1 Data processing

1) Take variables: {SiO2, Na2O, K2O, CaO, MgO, Al2O3, Fe2O3, CuO, PbO, BaO, P2O5, SrO, SnO2, SO2}.

2) Parameter setting: number of high potassium glass clusters: {4}; Number of lead-barium glass clusters: {3}.

2.2.2 Analysis of Cluster Category Differences

1) Cluster category difference analysis.

Table 7 and Table 8 show the results of quantitative field difference analysis, including the results of mean ±standard deviation, F test results, and significant P value.

Analyze whether the P value of each analysis item is significant (P<0.05).

If it is significant, reject the original hypothesis, which indicates that there is a significant difference between the two groups of data. The difference can be analyzed in the way of mean ± standard deviation, otherwise, it indicates that the data does not show differences.

Table 7: Cluster Analysis

	Cluster classification of high potassium glass (mean ±standard deviation)				*F*	*P*
	Category 2(n=9)	Category 1(n=6)	Category 4(n=2)	Category 3(n=2)		
SiO2	63.624±3.558	92.635±3.197	61.29±0.82	78.07±1.966	108.595	0.000***
Na2O	0.927±1.427	0.0±0.0	2.665±0.771	0.0±0.0	3.583	0.039**
K2O	10.818±2.37	1.31±1.952	6.575±1.223	4.71±6.661	15.241	0.000***
CaO	6.363±2.64	1.102±0.649	0.0±0.0	2.355±3.33	9.801	0.001***
MgO	1.133±0.672	0.197±0.306	0.935±0.12	1.375±0.219	4.64	0.017**
Al2O3	7.349±2.346	2.387±1.23	1.575±2.227	4.62±2.22	9.283	0.001***
Fe2O3	2.312±1.643	0.212±0.122	1.04±0.0	1.185±1.676	3.332	0.048**
CuO	2.819±1.565	1.433±0.988	1.19±0.141	1.64±2.319	1.571	0.238
PbO	0.41±0.64	0.042±0.102	0.19±0.0	0.5±0.707	0.792	0.517
BaO	0.579±1.001	0.0±0.0	0.0±0.0	0.985±1.393	1.127	0.370
P2O5	1.523±1.652	0.332±0.262	0.22±0.057	1.23±0.184	1.453	0.267
SrO	0.048±0.051	0.0±0.0	0.0±0.0	0.035±0.049	2.084	0.145
SnO2	0.0±0.0	0.0±0.0	0.0±0.0	1.18±1.669	4.474	0.020**
SO2	0.136±0.205	0.0±0.0	0.0±0.0	0.0±0.0	1.291	0.314

Note: * * *, * * and * represent the significance level of 1%, 5% and 10% respectively

3. Analysis of variance

For SiO_2, the significance P value is 0.000 * * *, showing significance at the level. The original hypothesis is rejected, indicating that there is a significant difference between the categories of SiO_2 classified by cluster analysis;

For Na_2O, the significance P value is 0.039 * *, showing significance at the level. The original hypothesis is rejected, indicating that there is a significant difference between the categories classified by cluster analysis for Na_2O;

For K_2O, the significance P value is 0.000 * * *, which is significant at the level. The original hypothesis is rejected, indicating that there is a significant difference between the categories of K_2O classified by cluster analysis;

For CaO, the significance P value is 0.001 * * *, showing significance at the level, and rejecting the original hypothesis, indicating that CaO has significant differences between the categories classified by cluster analysis;

For MgO, the significance P value is 0.017 * *, showing significance at the level, rejecting the original hypothesis, indicating that there is a significant difference between the categories classified by cluster analysis;

For Al_2O_3, the significance P value is 0.001 * * *, which is significant at the level. The original hypothesis is rejected, indicating that there is a significant difference between Al_2O_3 categories divided by cluster analysis;

For Fe_2O_3, the significance P value is 0.048 * *, which is significant at the level. The original hypothesis is rejected, indicating that Fe_2O_3 has significant differences among the categories classified by cluster analysis;

For CuO, the significance P value is 0.238, which is not significant at the level, and the original hypothesis cannot be rejected, indicating that there is no significant difference between the categories of CuO classified by cluster analysis;

For PbO, the significance P value is 0.517, which is not significant at the level, and the original hypothesis cannot be rejected, indicating that there is no significant difference between the categories classified by PbO cluster analysis;

For BaO, the significance P value is 0.370, which is not significant at the level, and the original hypothesis cannot be rejected, indicating that there is no significant difference in BaO among the categories classified by cluster analysis;

For P_2O_5, the significance P value is 0.267, which is not significant at the level, and the original hypothesis cannot be rejected, indicating that there is no significant difference between the categories of P_2O_5 classified by cluster analysis;

For SrO, the significance P value is 0.145, which is not significant at the level, and the original hypothesis cannot be rejected, indicating that there is no significant difference between the categories classified by cluster analysis;

For SnO_2, the significance P value is 0.020 * *, showing significance at the level, and rejecting the original hypothesis, indicating that there is a significant difference between the categories of SnO_2 classified by cluster analysis;

For SO_2, the significance P value is 0.314, which is not significant at the level, and the original hypothesis cannot be rejected, indicating that there is no significant difference in SO_2 among the categories classified by cluster analysis.

Table 8: Cluster classification of lead barium glass

	Cluster classification of lead barium glass (mean ± standard deviation)			*F*	*P*
	Category 1(n=22)	Category 2(n=21)	Category 3(n=7)		
SiO2	26.537±7.54	60.236±11.21	21.593±13.619	72.93	0.000***
Na2O	0.219±0.588	1.881±2.4	0.0±0.0	6.921	0.002***
K2O	0.18±0.37	0.182±0.155	0.189±0.277	0.003	0.997
CaO	2.88±1.821	1.172±0.937	1.871±1.386	7.505	0.001***
MgO	0.734±0.689	0.704±0.581	0.101±0.268	3.159	0.052*
Al2O3	2.944±1.505	4.867±3.971	2.009±1.64	3.775	0.030**
Fe2O3	0.865±0.898	0.568±1.047	0.216±0.571	1.428	0.250
CuO	1.205±1.279	1.011±0.975	6.556±3.183	36.822	0.000***
PbO	47.338±8.854	19.441±7.465	26.344±7.799	65.328	0.000***
BaO	8.007±4.576	6.854±3.329	27.706±6.982	61.24	0.000***
P2O5	4.901±4.538	0.87±1.592	5.084±3.067	8.815	0.001***
SrO	0.415±0.282	0.223±0.2	0.464±0.277	4.154	0.022**
SnO2	0.06±0.154	0.073±0.288	0.0±0.0	0.311	0.734
SO2	0.0±0.0	0.174±0.799	5.074±7.194	10.915	0.000***

Note: * * *, * * and * represent the significance level of 1%, 5% and 10% respectively

4. Model evaluation

4.1 Model advantages

1) The model established in this paper can be closely linked with real life, and can solve the problems raised in combination with the actual situation. In addition, the model proposed in this paper is closer to reality, with strong universality and popularization.

2) The PCA method used in this paper can find out the correlation between different factors, and clarify whether the factors are positive, negative or irrelevant.

3) The model designed in this paper has strong operability, wide application scope, high reliability of factor weights, and can be widely used in other fields.

4.2 Model Disadvantages

1) The model proposed in this paper can only be used for qualitative analysis, not quantitative analysis.

2) The model proposed in this paper uses the Pearson correlation coefficient, which must ensure that both variables are continuous variables. In addition, the Pearson correlation coefficient is susceptible to outliers.

4.3 Model promotion

The model proposed in this paper can not only be applied to the classification analysis of ancient glass components, but also be applied to the analysis of a wide range of archaeological relics, such as bronze, which has extensive promotion value.

References

[1] Yu D, Xu H, Chen C L P, et al. Dynamic coverage control based on k-means[J]. IEEE Transactions on Industrial Electronics, 2021, 69(5): 5333-5341.
[2] Li T, Kou G, Peng Y, et al. An integrated cluster detection, optimization, and interpretation approach for financial data[J]. IEEE Transactions on Cybernetics, 2021.
[3] Wu C, Chen H, Du B, et al. Unsupervised change detection in multitemporal VHR images based on deep kernel PCA convolutional mapping network[J]. IEEE Transactions on Cybernetics, 2021.
[4] Shou Y, Meng T, Ai W, et al. Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis[J]. Neurocomputing, 2022, 501: 629-639.
[5] Zhang X, Jiang X, Jiang J, et al. Spectral–spatial and superpixelwise PCA for unsupervised feature extraction of hyperspectral imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-10.