Glass composition analysis and identification model based on variance test

Miao Wang, Yifan Zhang *, Maotai Cheng

School of Computer Science, Zhuoyue Honors College, School of Sciences, Hangzhou Dianzi University of China, Hangzhou, China, 310018

* Corresponding Author Email: 20010139@hdu.edu.cn

Abstract. In this study, a qualitative and quantitative study was conducted on three factors: weathering condition and glass type, decoration, and color; bar graphs and line graphs were drawn and described for each factor at different levels of weathering condition; the relationship between the three factors and weathering condition was analyzed using chi-square test, and the correlation between glass type and weathering condition was obtained to be high; the correlation analysis (2) based on APRIORI algorithm (1) was used to The correlation analysis (2) based on APRIORI algorithm (1) was used to quantitatively calculate the relationship between different levels under each factor and the weathering condition, and it was obtained that the levels of ornamentation B, light blue and blue-green, high potassium and lead-barium had more significant effects on the weathering condition, and the results were consistent with the qualitative analysis. For the second sub-question of question 1, the chemical components with high variability between weathered and unweathered samples were selected using the MannWhitney test, and ridge regression was used to establish a model of weathering condition and these chemical components.For question 2, this paper conducted independent sample MannWhitney tests on glass type and each chemical composition to obtain twelve chemical composition indicators with significant differences between high potassium and lead-barium glasses. Specific functions between glass type and each component were fitted by ridge regression on these chemical components for high potassium and lead-barium glasses.

Keywords: Glass; Weathering; APRIORI algorithm; Ridge regression; MannWhitney test.

1. Introduction

As one of the commodities traded between ancient China and the world civilization, glass symbolizes the prosperity and flourishing of Chinese culture for five thousand years [1]. By analyzing the physical properties and chemical composition of glass and tracing its history, it is of great significance to maintain the roots of the Chinese nation, preserve cultural relics and protect history [2-4].

The process of making glass is usually made by firing quartz sand (SiO2) as the main ingredient, grass wood ash, natural bubbling soda, saltpeter or lead ore as the accelerant, and limestone (CaCO3) as the stabilizer [5-7]. The type of glass obtained by firing changes with the change of combustion agent. The reaction process of lead-barium glass and potassium glass, for example, is shown as follows.

Finished glass after firing is subject to environmental weathering, resulting in changes in its physical properties and chemical composition [8-10]. The physical properties and chemical composition ratios of a batch of ancient Chinese glass products (divided into high potassium glass and lead-barium glass) are available, please complete the following questions based on the sample data.

1) Analyze the relationship between the weathering of the glass surface and the three factors of decoration, glass type and color.

2) Analyze the pattern of weathering on the glass surface in relation to the chemical composition content considering the glass type.

2. Model construction and testing

2.1. Problem 1 Model Construction and Testing

2.1.1 Cardinality test to examine the correlation of factors

The surface weathering, ornamentation, color, and type given in Table 1 all belong to a fixed class of data without specific values and without knowing the distribution of the variables, so a non-parametric hypothesis test is conducted on the data using the chi-square test, the basic idea of which is to determine whether there is a correlation between the theoretical frequency and the actual frequency by comparing the degree of agreement. The specific implementation steps are as follows.

First make the hypothesis that

Original hypothesis H0: Surface weathering and its glass type, decoration, and color are independent of each other (not correlated).

Alternative hypothesis H1: There is a correlation between surface weathering and its glass type, decoration, and color.

We analyzed the specific levels of each of the three factors under color, type, and ornamentation by SPSS, and obtained the chi-square, corrected chi-square, and p-values for each factor (SPSS results were placed in the supporting material) (Table 1).

Table 1. Chi-square test

factor	level	Observations			Chi-square	P
		weathering	Unweathered	total		
color	black	2	0	2	6.287	0.507
	blue-green	9	6	15		
	green	0	1	1		
	light blue	12	8	20		
	light green	1	2	3		
	dark blue	0	2	2		
	dark green	4	3	7		
	pruple	2	2	4		
type	lead barium	24	12	36	5.4	0.020**
	high-k	6	12	18		
ornamentation	A	9	11	20	5.747	0.056*
	B	6	0	6		
	C	15	13	28		

Analyzing the test table, it can be seen that at $\alpha = 0.05$, i.e., the confidence level of 95%, $p > \alpha$ corresponding to color and decoration, so the original hypothesis H0 is accepted, i.e., there is no correlation between surface weathering and color and decoration (independent of each other); similarly, $p < \alpha$ corresponding to type, so the original hypothesis H0 is rejected, i.e., there is a correlation between surface weathering and glass type.

In summary, there is a correlation between the factor of glass type and weathering, while the correlation between the factors of color and decoration and weathering is not obvious. This result is the same as the results of the qualitative analysis of the previous graphs, so it can be considered to have a high degree of confidence.

2.1.2 Association analysis based on APRIORI algorithm

We investigated whether there is a correlation between surface weathering and decoration, color, and glass type by means of a chi-square test.

The correlation between surface weathering and grain, color, and glass type was investigated by the chi-square test and qualitatively analyzed by cross-plot. In order to study more deeply the influence of different levels on surface weathering under each factor, we choose the correlation analysis method, firstly, we calculate the support and confidence degree between levels, and then we select the frequent items set (i.e., the valid set) by APRIORI algorithm, so as to calculate the enhancement degree of different levels on surface weathering, and get the calculated influence of each level on surface weathering. The specific steps of its implementation are as follows.

Step 1: Calculate support, confidence, and elevation

The support represents the probability of the occurrence of an item set, that is, the ratio of the occurrence of X and Y at the same time to all cases, which is calculated as follows.

$$Support(\{X,Y\}) = P(X,Y) = \frac{num(x,y)}{num(allsamples)} \tag{1}$$

For example, if there are 11 pieces of glass with weathered surface and A decoration, and 58 pieces of glass in total, the support of the glass with weathered surface and A decoration is $support(x,y) = \frac{11}{58}$.

The confidence level, which indicates the probability that if X occurs, Y also occurs at the same time, is calculated as follows.

$$Confidence(X \to Y) = P(x|Y) = \frac{P(xy)}{p(y)} \tag{2}$$

For example, if there are 25 glasses weathered on the surface and 11 glasses weathered on the surface with the pattern A, the confidence level is $\frac{support(x,y)}{p(y)} = \frac{11/58}{25/58}$.

The degree of elevation represents the ratio of the probability of containing Y at the same time under the condition of containing X, to the probability of seeing only Y occur. It is calculated as follows.

$$Lift(X \to Y) = \frac{Cinfidence(X \to Y)}{Support(Y)} \tag{3}$$

Step 2: APRIORI algorithm to filter the frequent items set

In the form given in the question, due to the small data set, there may be chance cases that affect the analysis, for example, when the color is black, the surface is weathered, thus obtaining a very large lift, but there are only two samples in this case, there is a large chance.

In order to overcome the above problem, it is necessary to prune the original data, and the data with reference left after pruning is called the set of frequent items; the pruning process is implemented by the APRIORI algorithm, whose basic steps are as follows.

(1) List the itemsets that contain only one element, such as lead-barium and blue-green.

(2) Calculate the support of each itemset, retain the itemsets that satisfy the minimum support threshold (here 5%), and prune the itemsets that do not satisfy it.

(3) Add elements to the remaining itemsets, using the remaining itemsets to generate all possible combinations.

(4) Repeat Step 2 and Step 3 to determine the support for larger and larger itemsets, knowing that no new itemsets are generated.

Step 3: Result Analysis

The lifting degree of frequent itemsets is calculated, and if the lifting degree is greater than 1, it can be considered as a strong association, and the single-factor strong association analysis table and the multi-factor strong association analysis table are obtained (there are more cases of multi-factor combinations, which will be shown in the appendix) (Table 2).

Table 2. Factor correlation analysis

Factor	Relation	Rule support	Confidence	Support of afterparts	Lifting degree
ornamentation	B->weathering	0.103	1	0.586	1.706
color	light blue->weathering	0.207	0.6	0.586	1.024
color	blue-green->weathering	0.155	0.6	0.586	1.024
type	lead barium->weathering	0.483	0.7	0.586	1.194
type	high-k->unweathering	0.207	0.667	0.414	1.611

According to the single-factor strong correlation analysis table, we can find that the weathering results are more likely to occur when the ornamentation is B, when the color is light blue or blue-green, when the glass type is lead-barium, and when the glass type is high potassium, it is more likely to show no weathering results.

In summary, the B level under the decoration factor, the two levels of light blue and blue-green under the color factor, and the two levels of lead-barium and high potassium under the glass type factor, these levels have a greater correlation with weathering. It is also consistent with the results of our previous graphical qualitative analysis and has a high degree of confidence.

2.2. Problem 2 Model Construction and Testing

2.2.1 Shapiro-Wilk test for data normality

To determine the appropriate method, we need to test the normality of the data first, and due to the small amount of data, the Shapiro-Wilk test is used in this paper.

Step 1: Making assumptions

H0: the sample obeys normal distribution; H1: the sample does not obey normal distribution.

Step 2: Calculate the statistic

The statistic of this test is calculated as follows

$$W=\frac{(\sum_{i=1}^{n}a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{4}$$

where x(i) denotes the i-th smallest number of the sample and $\bar{x}=\frac{(x_1+x_2+\cdots+x_n)}{n}$ denotes the sample mean.

The constant ai is calculated by the following equation.

$$(a_1,\dots,a_n)=\frac{m^T V^{-1}}{\sqrt{(m^T V^{-1} V^{-1} m)}} \tag{5}$$

where $m = (m_1, \dots m_n)^T$, denotes the expected value of an ordered independent identically distributed statistic sampled on a standard normally distributed random variable, and V denotes the covariance of the ordered statistic.

Step 3: Test and Analysis

The S-W test was performed using SPSS, and two representative test tables (Table 3) and histograms (Figure. 1) were selected for analysis.

Table 3. S-W normaily test

variable	samples	mean-value	standard deviation	skewness	kurtosis	S-W test	K-S test
Na2O	49	0.904	1.813	2.299	5.069	0.58(0.000***)	0.405(0.000***)
Fe2O3	49	0.656	0.948	2.067	5.316	0.732(0.000***)	0.245(0.005***)

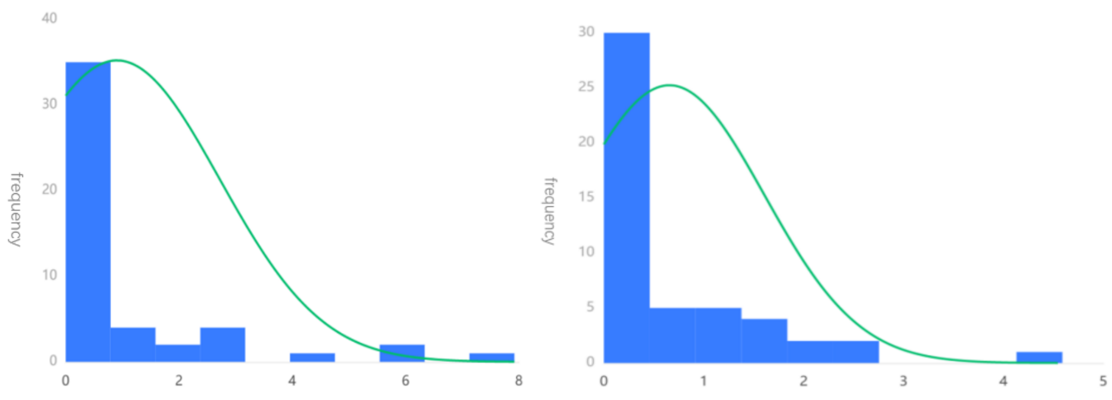

Figure 1. Na2O, Fe2O3 normaily test histogram

Taking the above test results as an example, the p-value corresponding to the S-W test results is much smaller than α(0.05), so the original hypothesis H0 is rejected, and combined with the histogram of normality test, it is found that most of the chemical composition data data are obviously not normally distributed, so the T-test cannot be used, and the independent sample MannWhitney test will be used in the subsequent test.

2.2.2 MannWhitney test

Step 1: Mix the i-th chemical composition in weathered and unweathered data sets, and arrange the ranks in order of size, with the smallest data rank being 1, the second smallest data rank being 2, and so on. (If multiple data are equal, the average of these

(If multiple data are equal, the average of the data ranking is taken as its rank).

Step 2: Find the rank sums R1 and R2 of the i-th chemical composition in the weathered and unweathered data sets, respectively.

Step 3: Calculate the U-value of each of the two data sets.

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1+1)}{2} - R_1 \tag{6}$$

$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2+1)}{2} - R_2 \tag{7}$$

Step 4: Select the smaller Un value and check the table to get its corresponding p value at $\alpha = 0.05$. Compare the p value with α. If $p < \alpha$, it means there is a significant difference.

We used SPSS to test and obtained the MannWhitney test table (Table 4) for the respective chemical composition of high potassium and lead-barium glass (only the chemical composition of lead-barium type with very significant difference is shown here).

Table 4. Mann-Whitney Significant difference test

Variable	value	sample size	median	standard deviation	statistic	P	median difference
SiO2	unweathering	25	53.79	11.72	594	0.000***	30.345
	weathering	24	23.445	8.628			
	total	49	35.78	18.646			
Na2O	unweathering	25	0	2.28	427	0.001***	0
	weathering	24	0	0.368			
	total	49	0	1.813			
CaO	unweathering	25	0.84	1.282	133	0.001***	2.13
	weathering	24	2.97	1.658			
	total	49	1.48	1.635			
PbO	unweathering	25	20.12	8.897	37	0.000***	24.315
	weathering	24	44.435	11.304			
	total	49	31.9	14.947			
P2O5	unweathering	25	0.19	1.779	90.5	0.000***	5.685
	weathering	24	5.875	4.129			
	total	49	1.41	3.909			

Analysis of the table shows that in the case of lead-barium type, there is a close relationship between the degree of weathering and the 5 chemical components of silica, sodium oxide, calcium oxide, lead oxide, and phosphorus pentoxide, that is, there is a very obvious difference between the 5 chemical components of weathered and unweathered glass in the lead-barium type, and a judgment can be made on the weathering of the lead-barium type based on these 5 chemical components.

Similarly, the degree of weathering in the high potassium type is closely related to the nine components of silica, potassium oxide, calcium oxide, magnesium oxide, aluminum oxide, iron oxide, lead oxide, phosphorus pentoxide, and strontium oxide, and the weathering of the high potassium type can be judged by these nine chemical components.

2.2.3 Ridge regression calculates statistical laws

By significant differences we selected the significant chemical components of high potassium and lead-barium class data respectively for weathering conditions, and the statistical laws of these chemical components with weathering conditions to date will be analyzed in detail by ridge regression below.

(1) Determination of ridge regression K values

Firstly, the corresponding ridge traces of high potassium glass and lead-barium glass were plotted as follows (Figure 2). The ridge plot with variance expansion factor method can determine K1 = 0.188 for lead-barium glass and K2 = 0.184 for high-potassium glass.

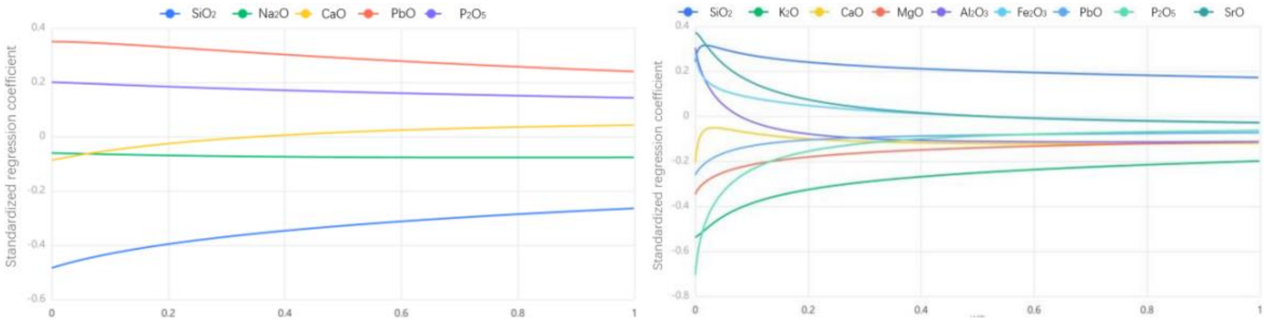

Figure 2. Ridge regression ridge trace map

(2) Ridge regression modeling and solving

In order to analyze the dependent variable (weathering) and to ensure the magnitude of the regression coefficients, we expressed its weathering using 0 and 100 as follows (Table 5).

$$Weathering\ condition = \begin{cases} 0 & \text{unweathering} \\ 100 & \text{weathering} \end{cases} \tag{8}$$

Table 5. Lead barium glass ridge regression analysis

K=0.188	Denormalization coefficient	Standardization coefficient		t	P	R^2	adjustment R^2	F
	B	Standard error	Beta					
constant	51.523	14.846	-	3.47	0.001***	0.766	0.739	28.22(0.000***)
SiO2	-1.084	0.197	-0.4	-5.511	0.000***			
Na2O	-1.923	1.846	-0.069	-1.042	0.303			
CaO	-0.884	2.123	-0.029	-0.417	0.679			
PbO	1.119	0.236	0.331	4.75	0.000***			
P2O5	2.39	0.897	0.185	2.664	0.011**			

The results of the lead-barium glass ridge regression were analyzed: based on the F-test significance P-value of 0.000 ＊＊＊, the level of proof showed significance and there was a regression relationship between the surface independent variable chemical composition and the dependent variable weathering condition. Meanwhile, the coefficient of determination of the goodness of fit of the model is 0.766, which indicates that the model performs well. So the equation between the weathering condition of lead barium glass and chemical composition is

$$y_1 = 51.523 - 1.084x_1 - 1.923x_2 - 0.884x_3 + 1.119x_4 + 2.39x_5 \tag{9}$$

Where, y1 denotes the value corresponding to the weathering of lead-barium glass, x1, x2, . . . , x5 denote: silica, sodium oxide, calcium oxide, lead oxide, and phosphorus pentoxide, respectively.

Similarly, we performed the ridge regression analysis again, and the results of the ridge regression analysis corresponding to high potassium glass are tabulated (Table 6).

Table 6. high-K glass ridge regression analysis

K=0.184	Denormalization coefficient		Standardization coefficient	t	P	R^2	adjustment R^2	F
	B	Standard error	Beta					
constant	16.67	25.046	-	0.666	0.524	0.806	0.587	3.683(0.040**)
SiO2	0.81	0.264	0.244	3.065	0.015**			
K2O	-3.016	1.139	-0.333	-2.647	0.029**			
CaO	-1.415	1.561	-0.097	-0.907	0.391			
MgO	-12.487	10.676	-0.185	-1.17	0.276			
Al2O3	-1.145	2.422	-0.073	-0.473	0.649			
Fe2O3	1.602	4.165	0.052	0.385	0.71			
PbO	-10.07	12.803	-0.108	-0.786	0.454			
P2O5	-6.16	4.915	-0.164	-1.253	0.245			
SrO	90.251	163.099	0.082	0.553	0.595			

The results of the high potassium glass ridge regression were analyzed: based on the F-test significance P-value of 0.040 ＊＊, the level showed significance, indicating a regression relationship between the independent variable chemical composition and the dependent variable weathering. The coefficient of determination of the goodness of fit of the model was 0.806, indicating that the model performed relatively well.

The equation between the weathering condition of lead barium glass and chemical composition was $y_2 = 16.67 + 0.81x_1 - 3.016x_2 - 1.415x_3 - 12.487x_4 - 1.145x_5 + 1.602x_6 - 10.07x_7 - 6.16x_8 + 90.251x_9$.

2.2.4 Analysis of the results of Question 2

The predictions are made by the model and compared with the true values to obtain the following Figure 3.

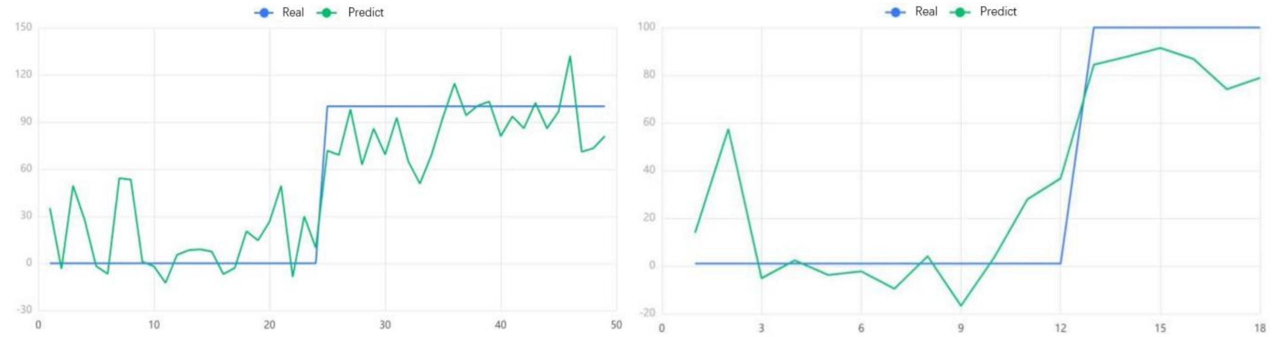

Figure 3. Comparison of actual and predicted values of lead-barium and high-potassium glass ridge regression. Lead barium glass (left), high potassium glass (right).

3. Conclusions

In this paper, we use the ridge regression method for regression analysis, which is able to consider the effects of error and variance, making the accuracy of regression results improve compared with the traditional least squares method. Ridge regression is a supplement to least squares regression, which loses unbiasedness in exchange for high numerical stability, thus obtaining higher computational accuracy. The significance of ridge regression coefficients is significantly higher than that of ordinary regression, which has a greater practical value in studies with covariance problems and pathological data bias.

References

[1] Mahieu Benjamin et al. A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data[J]. Food Quality and Preference, 2021, 93

[2] Zhang Guixin and Wang Zhenlei. Correlation Degree and Clustering Analysis-Based Alarm Threshold Optimization[J]. Processes, 2022, 10(2):224-224.

[3] González Estrada, Elizabeth and Villaseñor, José A. and Acosta Pech, Rocío. Shapiro-Wilk test for multivariate skew-normality[J]. Computational Statistics, 2022,: 1-17.

[4] Zhuang Cunbo, Liu Ziwen, Liu Jianhua, Ma Hailong, Zhai Sikuan, Wu Ying. Digital Twin-based Quality Management Method for the Assembly Process of Aerospace Products with the Grey-Markov Model and Apriori Algorithm[J]. Chinese Journal of Mechanical Engineering,2022,35(1).

[5] Fan Bingqian,Yao Jianting,Lei Dachen,Tong Ruipeng. Representation, mining and analysis of unsafe be haviour based on pan-scene data.[J]. Journal of thermal analysis and calorimetry,2022.

[6] AlNajjar D,AlNajjar H,AlRousan N. CoVID19 symptoms analysis of deceased and recovered cases usin g Chisquare test.[J]. European review for medical and pharmacological sciences,2020,24(21).

[7] Edwin H. Chen. The Power of the Shapiro-Wilk W Test for Normality in Samples from Contaminated Normal Distributions[J]. Journal of the Ame rican Statistical Association,2012,66(336).

[8] Tai Kah Yee, Dhaliwal Jasbir, Balasubramaniam Vinod. Leveraging Mann–Whitney U test on large-scale genetic variation data for analysing malaria genetic markers[J]. Malaria Journal,2022,21(1).

[9] Nie Ruichao. Analysis of influencing factors of Fiscal revenue in Beijing based on Ridge regression and Lasso regression model[J]. International Journal of New Developments in Engineering and Society, 20 22, 6.0(2.0)

[10] Wenyan Wang et al. Prediction Method of Core Dead Stock Column Temperature Based on PCA and Ridge Regression: Instrumentation, Control and System Engineering[J]. ISIJ International, 2021, 61(11): 2785-2791.