# Classification of Ancient Glass Based on Logistic Regression and K-means Clustering Algorithm

Ruimeng Shi[1], Shoupeng Chen[2, *], Duanduan Han[3]

[1]School of Management, Shijiazhuang Tiedao University, Shijiazhuang, Hebei, 050043

[2]School of Electrical and Electronic Engineering, Shijiazhuang Tiedao University, Shijiazhuang, Hebei, 050043

[3]Department of Mathematics and Physics, Shijiazhuang Tiedao University, Shijiazhuang, Hebei,050043

*Corresponding author: c903838417@163.com

**Abstract.** Ancient glass is easy to weather, after a large number of internal and external elements exchange, its composition ratio will change, thus affecting the correct judgment of the type of glass. In order to classify the types of glass, this paper first preprocessed the data, analyzed the classification rules of high-potassium glass and lead-barium glass by using logistic regression, and got a classification model that could identify the types of glass according to the content of chemical components. Next, K-means clustering algorithm is adopted to divide lead, barium and high potassium into A and B respectively according to the content of silica. After that, the rationality test was carried out on the clustering results. The test index was the belongingness of the sample to its category, and the clustering results passed the rationality test. Finally, the sensitivity analysis of the clustering results was carried out.

**Keywords:** Classification of glass, Cluster analysis, Reasonableness test.

## 1. Introduction

Glass is an important material evidence of cultural exchanges between China and the West in ancient times along the Silk Road. After learning foreign glass manufacturing techniques, Chinese people began to make their own glass using local materials. Therefore, although domestic and foreign glass have similar appearance, their chemical composition is different. Glass is mainly made of quartz sand, the main component of quartz sand is silicon dioxide (SiO2). Due to the high melting point of pure quartz sand, flux is added during refining to reduce the melting temperature, along with the stabilizing agent limestone. After calcination, limestone will be converted into calcium oxide (CaO), and different fluxes will make different varieties of glass. [1]For example, the glass of Chu culture is mainly lead-barium glass, and its flux in the process of calcination is lead ore, which has a high content of lead oxide (PbO) and barium oxide (BaO). [2]The glass in the south of China, i.e. Southeast Asia and India, is mainly high potassium glass, and its flux in the process of calcination is the material with high potassium content, such as plant ash. Due to the buried environment of cultural relics, ancient glass is easy to weather. When the internal and external elements of glass are exchanged in large quantities, the composition ratio will change, thus affecting the correct judgment of the type of glass. An artifact may be partially weathered, and to varying degrees. In order to better protect cultural relics, some current researches on ancient glass focus on the chemical composition of ancient glass, and classify ancient glass according to the chemical composition [3]. According to the chemical composition, the existing ancient glass mainly consists of high potassium and lead barium two types. Because ancient glass is easily weathered, there may be differences between weathering and unweathering under all kinds of glass. Different maintenance measures should be taken for the ancient glass with different chemical compositions and different degrees of weathering, so as to effectively protect the ancient glass from the secondary damage of the environment. Yin proposed to analyze the composition of ancient glass products by correlation prediction, and studied the classification of high-potassium and lead-barium glass by correlation prediction, effectively classifying ancient glass according to its chemical composition, but did not study the weathering characteristics of glass in detail

[4]. The infrared reflection, micrography and transmittance methods mentioned by Liu et al have been well studied on the degree of glass differentiation, but different types of glass have different regulations on the degree of weathering, and these methods have errors in the discrimination of the degree of weathering of high-potassium glass and lead-barium glass [5]. Zhang et al. proposed a glass classification method based on logistic regression and Fisher linear discrimination, but the classification model of glass weathering degree lacks corresponding rationality test [6]. Based on the advantages of previous glass classification models, the composition characteristics and weathering degree of ancient glass are considered in this paper. Firstly, the chemical composition data of glass were cleaned. Five chemical components, silica, potassium oxide, calcium oxide, lead oxide and barium oxide, were selected as judgment indexes. Logistic regression was used to analyze the two kinds of glass with high potassium and lead barium, and the classification rules were analyzed. Then, appropriate chemical components were selected for each class and subclass classification was carried out. At the same time, specific classification methods and results were given. The K-means algorithm is used to divide the high potassium, lead and barium glass into subclasses. Finally, the rationality assessment method and sensitivity assessment algorithm of cluster analysis are used to analyze the rationality and sensitivity.

## 2. Ancient glass classification model based on logistic regression and K-means clustering algorithm

### 2.1. Evaluation of glass weathering degree based on logistic regression

#### 2.1.1 Data cleaning

Due to the different detection techniques of various glass, the proportion of chemical composition may not be 100%, and the chemical composition detection data of individual glass samples may be missing. In order to facilitate the subsequent model training, it is necessary to preprocess the data.

Referring to the relevant data, it is generally believed that the total sum of the detected chemical components is 85~100% of the test samples are qualified. Therefore, samples whose total chemical composition is not in this range are first deleted. At the same time, certain chemicals are detectable only in individual glass samples and not in other glass samples. This feature is of little help to glass classification, and may even cause interference. Therefore, in this paper, chemical components that are absent in more than 50% of glass samples are excluded from the feature set. If there is any missing chemical composition in the remaining glass samples, the missing value will be filled with hot card. Therefore, a relatively perfect data set can be used for model training. In this paper, five chemical components including silicon dioxide, potassium oxide, calcium oxide, lead oxide and barium oxide were finally determined as the subsequent classification and discrimination characteristics.

#### 2.1.2 Feature weight standardization

In this paper, logistic regression is used to classify the data, but it is found in the previous calculation that if the data is not standardized, the results will produce very large errors. However, standardization will also bring about the problem of underestimating the level of indicators. Due to the large difference in the content of some chemical components, the levels reflected by them are also quite different. In order to solve the problems caused by standardization, the corresponding weight of each chemical component should be calculated and the standardized data multiplied by the weight, so as to reflect the level of each chemical component. In this paper, entropy weight method and normalized weight were used to solve the weight from two perspectives of data volatility and component content ratio, and the arithmetic average of the two methods was taken as the weight of each chemical component.

**Step1** Firstly, entropy weight method was used to calculate the weight, $n$ is the number of samples of high potassium or lead barium, $O'$ is the remaining chemical component of high potassium or lead barium. The calculation formula is as follows.

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}, \ (j \in O') \tag{1}$$

$$e_j = -\frac{1}{\ln n} \sum_{i=1}^{n} p_{ij} \ln(p_{ij}), \ (j \in O') \tag{2}$$

$$W_{j1} = (1 - e_j) / \sum_{i=1}^{n} (1 - e_j), \ (j \in O') \tag{3}$$

**Step2** Calculate the normalized weight. The greater the proportion of a chemical component in the glass, the greater the effect on the glass, and therefore the greater the weight should be.

$$m_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \tag{4}$$

$$W_{j2} = \frac{m_j}{\sum_{j \in O'} m_j} \tag{5}$$

**Step3** The arithmetic average of information entropy weight and normalized weight is taken to represent the final weight.

$$W_j = \frac{W_{j1} + W_{j2}}{2} \tag{6}$$

### 2.1.3 Logistic regression modeling

Judging glass weathering degree by chemical composition is a binary problem, so logistic regression is used. In this paper, 1 is used to represent weathering and 0 is used to represent unweathering. The probability of weathering can be judged by the chemical composition of a glass. When the probability is greater than 0.5, it means the glass is weathered; when the probability is less than or equal to 0.5, it means the glass is not weathered.[7].

When the data set was used for logistic regression for many times, it was found that the error was very large. After investigation, it was found that the problem was mainly caused by the content of silica, which caused a large gap with other components, so it was necessary to standardize the data. After that, through the weight of the chemical composition of the high potassium and lead and barium classes obtained above, the data is multiplied with the corresponding weight to improve the level of each chemical composition. Under each glass type, The $j$-th chemical component of the $i$-th glass is $x_{ij}^{old} J(\theta)$, the formula is as follows:

$$a_{ij} = \frac{x_{ij}^{old} - \min x_{\cdot j}^{old}}{\max x_{\cdot j}^{old} - \min x_{\cdot j}^{old}} \tag{7}$$

$$x_{ij} = a_{ij} \cdot W_j \tag{8}$$

$a_{ij}$ stands for standardized data, $x_{ij}$ stands for new data that has been normalized and weighted. After that, the new data was trained by logistic regression. The logistic regression function formats of high potassium and lead barium are as follows:

$$h_\theta(x_i) = g(\theta^T x_i) = \frac{1}{1 + e^{-\theta^T x_i}} \tag{9}$$

Among them, $h$ is the prediction function, $\theta$ is the parameter of each chemical component. The specific content is $\theta=\left[\theta_1,\theta_2,\cdots,\theta_n\right]^T$, $x_i$ is the value of each chemical composition of the $i$-th glass, and the specific content is $x_i=\left[x_{i1},x_{i2},\cdots,x_{in}\right]$.

The cost function is called $Cost\left(h_\theta(x),y\right)$, the loss function is denoted as $J(\theta)$, the formula is as follows:

$$Cost\left(h_\theta(x),y\right)=\begin{cases}-\log\left(h_\theta\left(x_i\right)\right) & if\ y=1\\ -\log\left(1-h_\theta\left(x_i\right)\right) & if\ y=0\end{cases} \tag{10}$$

$$\begin{aligned}J(\theta)&=\frac{1}{m}\sum_{i=1}^{m}Cost\left(h_\theta\left(x_i\right),y_i\right)\\&=-\frac{1}{m}\left[\sum_{i=1}^{m}\left(y_i\log h_\theta\left(x_i\right)+\left(1-y_i\right)\log\left(1-h_\theta\left(x_i\right)\right)\right)\right]\end{aligned} \tag{11}$$

Gradient descent algorithm is adopted to update parameters, and the formula is as follows:

$$\theta_j{}'=\theta_j-\frac{\alpha}{n}\sum_{i=1}^{n}\left(h_\theta(x_i)-y_i\right)x_{ij}-\frac{\lambda}{n}\theta_j,\ (j\in O') \tag{12}$$

Among them, $\alpha$ is the learning rate parameter, $\lambda$ is the regular term argument, $n$ is the number of high potassium or lead barium samples respectively.

### 2.2. Subclassification of glass based on K-means clustering

For each type of glass, it is necessary to subdivide it according to its chemical composition, and take the result after cleaning as the data set. Since the clustering standard is unknown, the K-means clustering algorithm is used to divide the data.

K-means clustering algorithm is an iterative solution algorithm. By specifying the number of clusters to be divided, the final cluster number and clustering center are adjusted iteratively. Because ancient glass can be divided into high potassium type and lead barium type according to chemical composition, and into weathered and unweathered type according to weathering degree, the initial clustering number of the two types of glass in this paper is set as 2, and the maximum number of iterations is both 10. [8] The specific algorithm flow chart is shown in Figure 1.

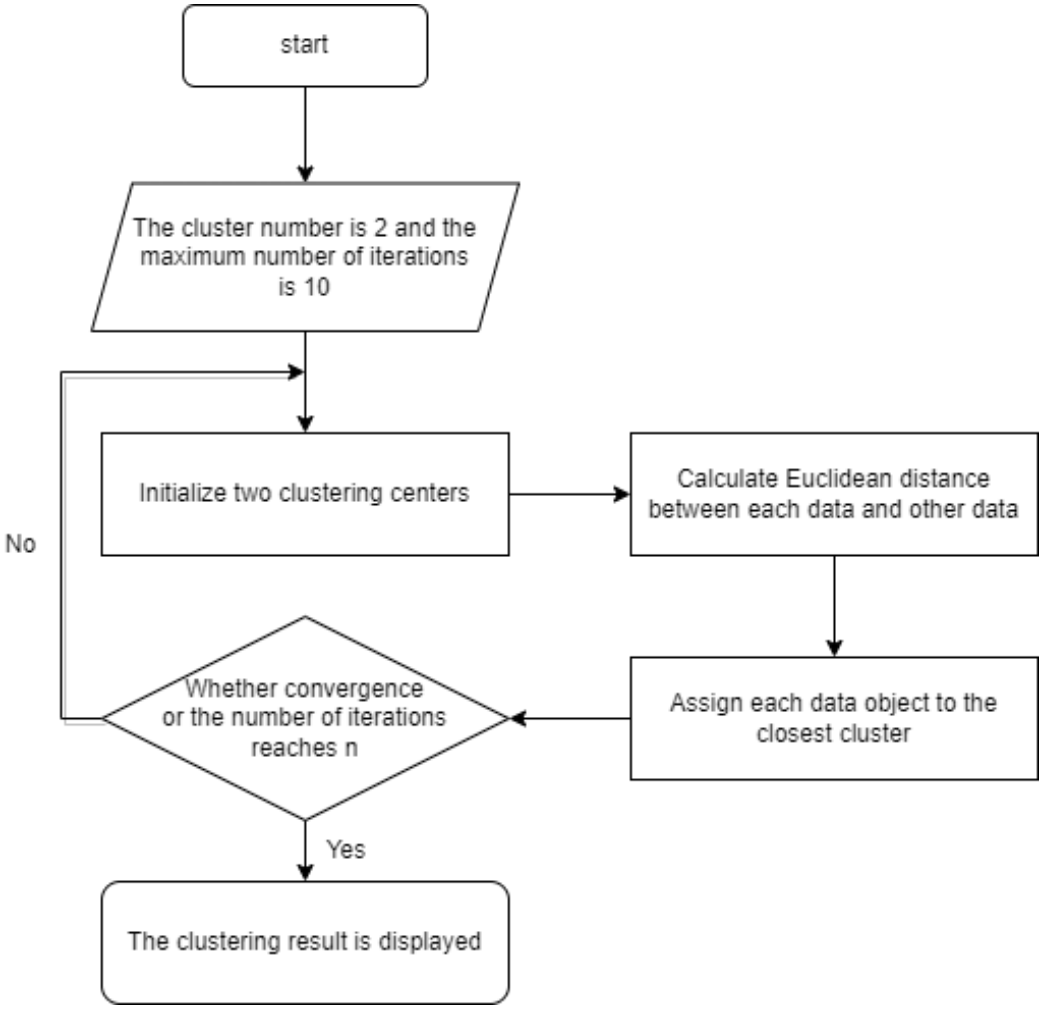



**Figure 1.** Flowchart of K-means clustering algorithm for this question

## 3. Experimental result

The data collected in this paper are from Question C of the 2022 Higher Education Community Cup National College Students Mathematical Contest in Modeling. A total of 69 cultural relics data are collected and 14 chemical components are selected. After processing, there are 65 effective data samples and 5 features. In this paper, 80% of the samples are taken as the training set and 20% as the test set.

### 3.1. Logistic regression and clustering results

#### 3.1.1 Logistic regression result

The main components of the two kinds of glass are silica, high potassium because the flux in the manufacturing process is the material with high potassium content such as wood ash, so the content of potassium oxide and calcium oxide is high; The flux in the manufacturing process of lead and barium is lead ore, so the content of lead oxide and barium oxide is high. Therefore, the division of the two types of glass is mainly based on the five chemical components of silicon dioxide, potassium oxide, calcium oxide, lead oxide and barium oxide. The calculated logistic regression iterative graph and confusion matrix are shown in Figure 2. and Figure 3. respectively.

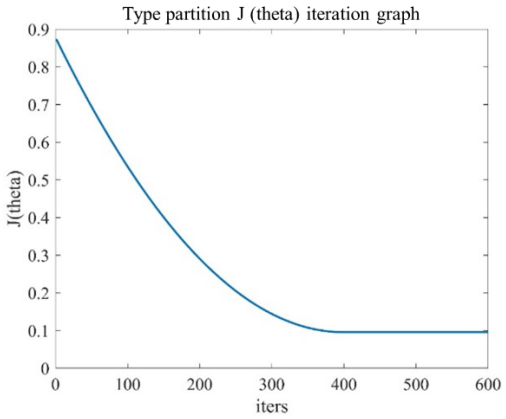



**Figure 2.** Logistic regression iteration diagram

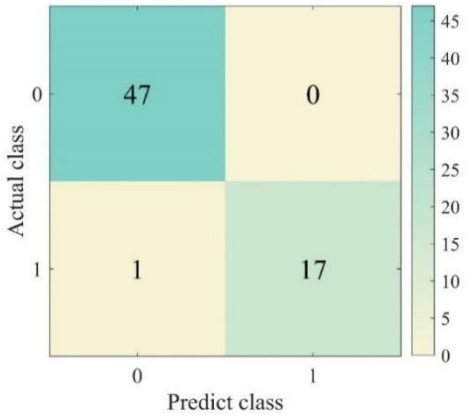



**Figure 3.** Confusion matrix

According to Figure 2, when the number of iterations of logistic regression reaches about 400, it converges and the loss value is about 0.1. According to Figure 3, only one glass with high potassium was predicted to be lead barium, and all the other predicted results were correct. Thus, the accuracy rate of training results was 98.5%, which was relatively high. The above results show that the classification of glass based on logistic regression proposed in this paper is good, and the two types of glass are mainly classified according to the content of five chemical components: silicon dioxide, potassium oxide, calcium oxide, lead oxide and barium oxide. When the content of silica, potassium oxide and calcium oxide is large, the glass can be judged as high potassium class; When the content of silica, lead oxide and barium oxide is large, the glass can be judged to be lead-barium, and the content of high-potassium silica is generally greater than that of lead-barium silica.

### 3.1.2 Clustering result

The clustering centers of high potassium and lead barium obtained by K-means clustering are shown in Table 1, and the sub-class names are both class A and Class B.

**Table 1.** Cluster center

| High-potassium | A | B | Lead-barium | A | B |
|---|---|---|---|---|---|
| $SiO_2$ | 64.93 | 91.29 | $SiO_2$ | 26.51 | 54.92 |
| $K_2O$ | 10.68 | 2.42 | PbO | 2.90 | 6.18 |
| CaO | 6.91 | 1.58 | BaO | 2.73 | 5.29 |
| MgO | 1.25 | 0.74 | CaO | 2.62 | 1.74 |
| $Al_2O_3$ | 7.23 | 2.34 | $Al_2O_3$ | 43.51 | 17.11 |
| $Fe_2O_3$ | 2.32 | 0.46 | CuO | 11.92 | 6.88 |
| CuO | 2.87 | 1.54 | $P_2O_5$ | 5.38 | 2.14 |
| $P_2O_5$ | 1.50 | 0.51 | | | |

It can be seen from Table 1 that high-potassium class A has higher content of silica and iron oxide, and low content of high-potassium class B, while high-potassium Class B has higher content of potassium, calcium oxide, magnesium oxide and alumina. According to statistics, there are 10 high-potassium class A, all of which are unweathered, showing the characteristics of not easy weathering. There are 8 high potassium class B, and the ratio of weathered to unweathered is 6:2, showing the characteristics of easier weathering. The content of silica in lead-barium A class is lower than that in lead-barium B class, while the content of alumina, copper oxide and phosphorus pentoxide is higher than that in lead-barium B class. According to statistics, there are 25 lead-barium A types, the proportion of weathering: unweathered is 22:3, showing the characteristics of easy weathering; there are 21 lead-barium B types, the proportion of weathering: unweathered is 1:20. It shows that the classification characteristics of the model are obvious and the effect is good.

### 3.2. Rationality test and sensitivity analysis of clustering

**Reasonableness test** Inspired by the Rationality Assessment Method of Classification Results of Cluster Analysis written by Wang et al., we adopt rationality assessment method to conduct rationality analysis of the cluster analysis used in this paper. [9] Since this method has achieved a good effect on the assessment results of the rationality of the cluster analysis of leukemia patients, this method is used to evaluate the rationality of the cluster analysis in this paper. This method defines the concept of median $cen$ of category attribution degree, and the range of $cen$ is 0~1. When $cen$ >0.95, the classification results have practical significance, which can reflect the characteristics of high potassium or lead barium glass under different subcategories, and the classification is reasonable; Otherwise, the classification is unreasonable.

According to the calculation, the median belonging degree of all samples in the subclassification of high potassium and lead barium were 0.9998 and 0.9609, respectively, which were both greater than 95%, indicating that the classification was reasonable.

Sensitivity analysis Inspired by the "Cluster Analysis and Geometry Based Target Feature Sensitivity Evaluation Algorithm" written by Wu Xia et al., [10]to measure the sensitivity of a cluster, the "closeness" and "separation" can be compared. The smaller the "tightness" of the intra-class feature distance of the same class, and the larger the "separation" of the inter-class feature distance of different classes, the more obvious the classification effect of this feature is considered, and the more sensitive it is to classification. Different from the traditional separation judgment based on the distance between two points, this method adopts the intra-class distance model and makes use of the global information so as to obtain better geometric meaning and make the results more accurate. After calculation, the sensitive factors of each chemical component under high potassium, lead and barium are listed in descending order as shown in Table 2.

**Table 2.** High potassium, lead barium chemical components sensitive factors

| Lead-barium | | High-potassium | |
|---|---|---|---|
| Chemical composition | Sensitive factor | Chemical composition | Sensitive factor |
| CuO | 3.36 | $Fe_2O_3$ | 4.76 |
| $P_2O_5$ | 2.56 | $SiO_2$ | 4.35 |
| $Al_2O_3$ | 2.44 | MgO | 4.28 |
| CaO | 2.32 | CaO | 3.64 |
| BaO | 2.30 | $K_2O$ | 3.19 |
| $SiO_2$ | 1.99 | CuO | 2.37 |
| PbO | 1.73 | $P_2O_5$ | 1.94 |
| | | $Al_2O_3$ | 1.28 |

Table 2 shows that in the lead barium subclassification, copper oxide has the largest sensitivity factor, followed by phosphorus pentoxide and alumina, indicating that copper oxide, phosphorus pentoxide and alumina have the largest positive significance for the lead barium subclassification. In the high potassium class, the sensitive factor of iron oxide is the largest, followed by silica and magnesium oxide, indicating that the content of iron oxide, silica and magnesium oxide has the greatest positive significance to the high potassium class subclassification. And for each kind of glass, this rule basically conforms to the characteristic rule of two categories of cluster analysis, which shows that the model in this paper has a good effect.

## 4. Conclusion

In this paper, a classification model of ancient glass based on logistic regression and K-means clustering algorithm is proposed. First, the data are preprocessed, and the classification rules of high-potassium glass and lead-barium glass are analyzed by logistic regression, and a classification model that can identify the type of glass according to the content of chemical composition is obtained. Then,

K-means clustering algorithm was used to divide lead-barium glass and high-potassium glass into class A and Class B respectively according to the content of silica. After that, the rationality test was carried out on the clustering results. The test index was the belongingness of the sample to its category, and the clustering results passed the rationality test. Finally, the sensitivity analysis of the clustering results was carried out, and the sensitivity factor of each feature to the clustering effect was obtained. The test results show that the proposed model is effective.

## References

[1] Moretti C , Hreglich S . Raw Materials, Recipes and Procedures Used for Glass Making[M]. John Wiley & Sons, Ltd, 2013.

[2] Qinghui Li, Jiaozhen Huang, Fei Li, et al. Detection of chemical constituents of a batch of ancient glass samples from the Warring States Period unearthed in China [J]. Cultural Relics Protection and Archaeological Science,2006(02): 8-13.DOI:10.16334/j.cnki.cn31-1652/k.2006.02.002.

[3] Jiajia Lu. Classification model of ancient glass based on integrated feature selection and random forest [J]. Journal of Ceramics,2023,51(04): 1060-1065.DOI:10.14062/j.issn.0454-5648.20220790.

[4] Yulong Yin. Composition Analysis of ancient glass Products by Association Prediction [J]. Contemporary Chemical Industry Research, 2023, No.126(01):122-126.

[5] Yuanxin Liu, Quan Jiang. Selection of Weathering Evaluation Method for flat Glass [J]. China Building Materials Science and Technology,1992(06):29-33.

[6] Shuaifan Zhang, Ruijuan Sun, Hanzhuo Ren. Chemical composition analysis and correlation study of ancient glass products [J]. Chemical Management, 2023, No.659 (08):138-141+164.DOI:10.19900/j.cnki.ISSN1008-4800.2023.08.041.

[7] Jichuan Wang, Zhigang Guo. Logistic Regression Model: Method and Application [M]. Higher Education Press, 2001.

[8] Suhui Wu, Ying Cheng, Yanning Zheng, et al. A review of K-means algorithm research [J]. Modern Library and Information Technology, 2011, 27(5):28-35.P29-34.

[9] Xiaodong Wang, Jun Tian. The reasonableness test method of cluster analysis classification results [J]. Mathematical practice and cognition, 2008, 38(20):4.P2-3.

[10] Xia Wu, Yan Zhou, Longpo Yang, et al. Target Feature Sensitivity Evaluation Algorithm Based on Cluster Analysis and Geometry [J]. Control System and decision, 2012, 27(6):5.P2-3.