International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-19

https://doi.org/10.70917/ijcisim-2026-0167
Article

Construction and Practical Research on the Dynamic Analysis Model of College Students' Employment Data in Higher Education Institutions under Big Data Environment

Meili Zhao *

Ordos Vocational College, Ordos, Inner Mongolia, 017000, China; bjxzmlbsq@163.com

Abstract: In the context of big data, college students face numerous challenges and opportunities in terms of employment. The processing of employment dynamic data suffers from issues such as poor employment trend prediction and weak precision in employment services. To address these issues, this paper employs big data to quantitatively process employment data. Using the K-means clustering algorithm, the paper determines the number of clusters in the employment data and completes the data clustering process. It primarily calculates the degree centrality parameters of network nodes to identify key points in the employment data required by users, thereby achieving employment data visualization. Based on time series analysis methods, the paper constructs a dynamic analysis model for college student employment data to analyze and predict employment trends among college students. Employment trend prediction practices are conducted using employment data from a certain higher education institution from 2017 to 2024 to predict the employment trends of college students in 2025. The employment index for the coming year in 2025 continues to show cyclical fluctuations, peaking in the fourth week (spring recruitment period), then rapidly declining, and recovering and rising to its highest point after the 30th week (autumn recruitment period). The results of the employment trend prediction generally align with actual social conditions.

Keywords: employment data; data visualization; clustering algorithm; time series analysis

1. Introduction

With the rapid development of the socio-economic landscape, the employment prospects of university graduates have become a focal point of societal concern. However, traditional career guidance models face challenges such as information asymmetry and a lack of personalized services, making it difficult to meet the actual needs of graduates [1-2]. To address these issues, universities should keep pace with the times, implement more targeted career guidance services, drive innovation in university graduate employment systems, and optimize career guidance efforts [3-5].

Most universities have employment management systems, and their graduate information management systems contain a large amount of student personal information and employment data [6-7]. However, the use of this information is limited to routine queries, with low utilization rates. Most of the data remains stored on hard drives without realizing its full potential [8-9]. Additionally, traditional database query technologies struggle to identify correlations between data or analyze latent information within the data [10]. Therefore, big data analysis technology can be utilized to enhance career guidance work, improve the quality of career guidance for university students, and assess and predict future employment trends [11-13]. Through data mining technology, exploring the correlation between student employment and academic performance can not only innovate talent cultivation models but also provide employers with a more precise talent evaluation system [14-15]. It also helps university administrators

understand the employment situation of college students, improve the structure of professional discipline construction, enhance the quality of professional talent cultivation, and thereby improve the quality of employment services [16-17].

Today's society is in the era of big data, and big data analysis technology is a widely used technology, with its application in the field of university employment receiving widespread attention. Literature [18] uses internet recruitment text information as its object, employing natural language processing and machine learning technologies to predict and analyze job positions, thereby identifying employment opportunities highly aligned with students' core skills. Literature [19] investigates the effectiveness of employment prediction supported by big data technology, utilizing an edge fog computing model to perform cluster analysis on individual college students. By enhancing the model's data parameters and quality, the accuracy of employment predictions is significantly improved. Literature [20] uses K-means clustering methods and the Apriori algorithm to analyze behavioral data of college students from enrollment to graduation. By exploring the relationship between student behavior and employment, it provides timely adjustment plans for university talent cultivation, thereby enhancing students' core employment competitiveness. Literature [21] conducts big data analysis of university public information, finding that universities with high employment rates often have a student-centered vision and university specialization, providing insights for schools to improve employment rates. Literature [22] introduces a balance coefficient to improve the traditional decision tree algorithm and constructs a university graduate employment information system based on student sample data analysis, thereby providing education administrators and graduates with high-precision employment prediction results. Literature [23] constructs an ant colony algorithm-based data mining model for college students' employment and entrepreneurship, which demonstrates high computational efficiency in analyzing college students' employment and entrepreneurship intentions, thereby providing college students with employment information aligned with their skills and interests. Literature [24] proposes a feature selection-based college student employment data analysis method, which mines the employment structure in the employment market and performs deep clustering based on college student data features to provide accurate employment trends and current employment status for college students. Based on this, big data analysis is conducted on employment data related to college students to discover potential patterns and identify hidden trends, providing decision-making basis for employment guidance, thereby promoting reforms in graduate employment systems and facilitating college student employment.

This paper proposes a visualization approach for processing employment data of vocational college students based on big data. Big data is utilized to quantify employment data, resulting in a time series of employment data. The K-means clustering algorithm is employed to cluster the employment data. After normalization processing, the clustering results of the employment data are obtained, determining the number of clusters. The employment data of college students is then presented in a visualized format. Based on this, time series analysis methods are combined to analyze and predict the employment time series data, determine the type of time series model, and obtain the corresponding order and parameters, thereby constructing a dynamic analysis model for college students' employment data. Taking 16,871 employment data records of college students from a certain higher education institution from 2017 to 2024 as the research object, the original employment data is subjected to visualization processing and clustering. The dynamic analysis model for college students' employment data proposed in this paper is applied to analyze and predict the employment trends of college students.

2. Visualization of college student employment data based on big data

To achieve dynamic analysis of employment data for college students, this chapter will first visualize the employment data of college students, use big data to process the employment data of college students, store the processed employment data in a big data warehouse, perform clustering processing, calculate the relationships between the data, and complete the data visualization analysis.

2.1. Data processing for college student employment

In order to process the employment data of college students, it is necessary to extract it from the employment data storage point, remove information other than employment information, change the data format and spatial dimensions, and standardize the employment data of college students using big data.

To this end, big data is used to quantify employment data and convert it into a vector $\overline{X}$, where $\overline{X} = (x_1, x_2, \cdots, x_i)$ and x_i represents the vector of the i th piece of employment data. At this point, it is only necessary to normalize the vectors resulting from this data transformation. However, the employment data stored in the network database has uncertain attribute values such as maximum and minimum values. Therefore, assuming that the employment data of higher education institutions in the

network contains a total of n data points, the arithmetic mean $\overline{x}$ and standard deviation s_i are:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$s_i = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2}$$

Based on the arithmetic mean $\overline{x}$ and standard deviation s_i obtained from Equations (1) and (2), the data is standardized to obtain the employment time series y_i:

$$y_i = \frac{x_i - \overline{x}}{s_i} \tag{3}$$

At this point, the time series $y_i = y_1, y_2, \cdots, y_j$ obtained from equation (3) is obtained, where y_i represents the j th employment data series, which has characteristics such as no bias, variance of 1, and arithmetic mean of 0. At this point, the time series y_i obtained from equation (3) is used for data clustering.

2.2. Clustering of employment data

2.2.1. Overview of Cluster Analysis

The principle of cluster analysis is to calculate the distance between all elements according to certain rules, and then divide elements with high similarity into multiple subsets based on distance. Each subset contains distinct elements. This achieves the goal of minimizing the dissimilarity among elements within each subset while maximizing the dissimilarity between different types of subsets, thereby achieving an optimal clustering effect. This process is used to classify data relationships.

The workflow of cluster analysis: First, sample data is selected. Then, these sample data are clustered and grouped using an appropriate distance method, followed by similarity calculations. The results of the clustering are then evaluated and predicted. The workflow diagram is shown in Figure 1.

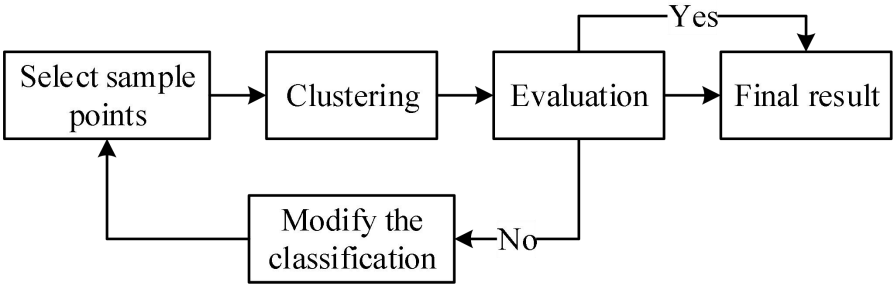


Figure 1. Clustering Flowchart.

2.2.2. Clustering Algorithm Classification

There are three commonly used clustering analysis methods: network-based clustering analysis, hierarchical clustering analysis, and density-based clustering analysis. Each of these methods has its own advantages and disadvantages.

1) Network-based clustering analysis method: This method primarily divides the application object space into several finite spaces, forms a network structure using these spaces, and completes the clustering analysis process using the network structure.

2) Hierarchical clustering analysis method: This method merges all sampling points from the bottom to the top into a tree or splits them from the top into a tree.

3) Density-based clustering analysis method: This method performs clustering analysis on all sample data based on density. It further clusters based on the density around the object according to the environment, thereby enabling the identification and extraction of clusters of any shape from a spatial database containing noise.

2.2.3. Methods for calculating similarity

The Euclidean distance formula is a relatively simple formula for calculating similarity. The principle is to determine the similarity between two sample data sets by calculating the distance between two points. The formula for Euclidean distance is as follows [25]:

$$d(r,y)=\sqrt{(r_1-y_1)^2+(r_2-y_2)^2+.....+(r_n-y_n)^2}=\sqrt{\sum_{i=1}^{n}(r_i-y_i)^2} \tag{4}$$

The smaller the calculated distance, the closer the distance between two different sample data sets, and the closer the similarity. The Pearson correlation coefficient is a more complex method than Euclidean distance for determining the similarity of user interests and preferences. Its formula is as follows:

For example, given two variables X and Y, the Pearson correlation coefficient between these two variables X and Y can be calculated using the following formula:

$$\rho(x,y)=\frac{cov(x,y)}{\sigma_x\sigma_y}=\frac{E((x-u_x)(y-u_y))}{\sigma_x\sigma_y}=\frac{E(xy)-E(x)E(y)}{\sqrt{E(x^2)-E^2(x)}\sqrt{E(y^2)-E^2(y)}} \tag{5}$$

$$\rho_{x,y}=\frac{N\sum XY-\sum X\sum Y}{\sqrt{N\sum X^2-(\sum X)^2}\sqrt{N\sum Y^2-(\sum Y)^2}} \tag{6}$$

$$\rho_{X,Y}=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}} \tag{7}$$

$$\rho_{X,Y}=\frac{\sum XY-\frac{\sum X\sum Y}{N}}{\sqrt{(\sum X^2-\frac{(\sum X)^2}{N})(\sum Y^2-\frac{(\sum Y)^2}{N})}} \tag{8}$$

From the Pearson formulas listed above, they are computationally equivalent, and all contain the characters E, cov, and N. Among them, E represents the mathematical expectation, cov represents the variance, and N represents the number of variable values. The expectation is the average value, which requires linking the values of $E(xy)$ and X and Y and then calculating the average value.

2.2.4. K-means algorithm

The K-means algorithm is the most commonly used clustering analysis algorithm based on partitioning methods. It first calculates the distances between samples and then uses distance similarity as an evaluation metric, meaning that the closer the distance between two samples, the greater their similarity [26]. Its ultimate goal is to group the closest samples into the same category and divide the sample data into compact and mutually independent data clusters of different categories.

In our daily work, the clustering algorithm analysis we use refers to the process of performing a clustering analysis on the required sample data. The most common clustering analysis process involves simultaneously extracting multiple features from the required sample data and refining them into a multidimensional vector. For example, M vectors are extracted simultaneously and combined into an M-dimensional vector, resulting in a mapping from the original data to the M-dimensional vector. Subsequently, based on certain criteria, the original data is subjected to clustering classification calculations. Under these criteria, data of the same type exhibit the highest similarity.

1) Suppose that the given initial dataset is $X=\{X_m \mid m=1,2,\ldots,total\}$, where the sample data in X is represented by d descriptive attributes $A_2,\cdots,A_d$ (dimensions).

2) Set sample data set $X_i=(X_{i1},X_{i2},X_{i3}\ldots X_{id}), X_j=(X_{j1},X_{j2},X_{j3}\ldots X_{jd})$, where each element in these two data sets represents the d specific values of descriptive attribute $A_1,A_2\ldots A_d$ corresponding to sample sets X_i and X_j.

3) The similarity between samples X_i and X_j is typically represented by distance. The smaller the distance, the closer the similarity between them and the smaller the difference. Conversely, the larger the

distance, the farther apart the similarity between them and the greater the difference. The Euclidean distance formula is as follows:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2} \tag{9}$$

4) Evaluation of the K-means clustering algorithm Typically, the sum of squared errors criterion function is used to verify the standard error. That is, for a given dataset X, only a small number of descriptive feature attributes are included, without including category attributes. Example: The dataset X contains K subsets of clusters $X_1, X_2, \ldots, X_n$, each cluster containing $n_1, n_2 \ldots n_k$: The cluster centers for each cluster subset are $m_1, m_2 \ldots m_k$, and the sum of squared errors criterion function formula is:

$$E = \sum_{i=1}^{k} \sum_{p \in Xi} || \, p - m_i \, || \tag{10}$$

2.2.5. Clustering of employment data

Employment data clustering is a crucial step in the process of visualizing and analyzing employment data. The results of data clustering determine the effectiveness of the visualization analysis. The input clustering data is a normalized n -dimensional vector, and the output is the data cluster. Each cluster contains multiple n -dimensional vectors. Based on the data clustering process, the employment data clustering results are obtained and stored in a big data database, completing the data clustering process and enabling data visualization analysis.

2.3. Data visualization analysis

Since college student employment data is stored in a big data database, there will be network nodes, so it is necessary to calculate the degree centrality parameters of the network nodes. Assuming that the degree centrality of the i th employment data node is $C(i)$ and the number of nodes in the big data database is $|V|$, the normalized degree centrality is:

$$C(i) = \frac{\sum_{j=1}^{|V|} x_{ij}}{|V| - 1} (i \neq j) \tag{11}$$

In this context, x_{ij} represents the direct connection between the i th and j th employment data sequences. At this point, data point scores are assigned based on the relationships between data sequences. Nodes with higher point scores are always higher than those with lower point scores, and lower nodes are connected to nodes with higher point scores. Therefore, x_i represents the employment data vector.

Based on the above calculation process, using equations (1) and (3) from the preceding text, we can determine the changes in the connectivity chain paths of employment data and the shortest paths for all time periods, thereby assessing the familiarity of employment data within the big data database. Using equation (11), we identify the key nodes representing the feature attributes of the data. At this point, by mapping the network node feature parameters, communication protocols, and other basic data attributes, we obtain the key points of employment data required by users, thereby presenting the employment data in a visual format.

3. Dynamic Analysis Model of College Student Employment Data

In the previous chapter, this paper used big data to visualize and cluster employment data for college students. In this chapter, based on the data processing, we will use time series analysis methods to construct a dynamic analysis model for college student employment data, enabling analysis and prediction of employment trends for college students.

3.1. Time Series Analysis Methods

Time series analysis is a quantitative method for predicting the future values of an object of interest. By analyzing time series data, it leverages the continuous patterns of development in phenomena to make predictions through statistical analysis or by establishing mathematical models for trend extrapolation

[27]. Time series analysis is also known as time series forecasting, historical extrapolation, or extrapolation. A time series, also referred to as a dynamic sequence, is a sequence of observations of economic variables arranged in order. Currently, the most well-developed and precise algorithm for analyzing and forecasting time series data is the Box-Jenkins method, whose commonly used models include: AR models, MA models, ARMA models, and ARIMA models.

3.2. Smooth Time Series Analysis

In predictive analytics, methods used for stationary time series include autoregressive models (AR models), moving average models (MA models), and autoregressive moving average models (ARMA models).

1) Autoregressive model AR(p)

A P-order autoregressive model is denoted as AR(p) and satisfies the following equation:

$$y_t = \lambda_1 y_{t-1} + \lambda_2 y_{t-2} + \cdots + \lambda_p y_{t-p} + \varepsilon_t \quad t = 1, 2, \cdots, T \tag{12}$$

where $\lambda_1, \lambda_2, \cdots \lambda_p$ are the coefficients of the autoregressive model; p is the order of the autoregressive model; ε_t is a white noise sequence with mean 0 and variance σ^2.

2) Moving Average Model MA(q)

The qth-order moving average model is denoted as MA(q) and satisfies the following equation:

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{13}$$

The parameters $\theta_1, \theta_2, \cdots, \theta_q$ are the coefficients of the qth-order moving average model; ε_t is a white noise sequence with mean 0 and variance σ^2.

3) Autoregressive-Moving Average Model ARMA(p, q)

The autoregressive moving average model is a combination of an autoregressive process and a moving average process, so it can be expressed as:

$$y_t = \lambda_1 y_{t-1} + \lambda_2 y_{t-2} + \cdots + \lambda_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{14}$$

where p is the autoregressive order, q is the moving average order, $\lambda_1, \lambda_2, ...\lambda_p, \theta_1, \theta_2, ..., \theta_q$ are the parameters of the model, and $\lambda_p \neq 0, \theta_q \neq 0$, ε_t is a white noise sequence with mean 0 and variance σ^2. This model is abbreviated as ARMA(p, q).

3.3. Non-stationary time series analysis

The three models described above are only applicable to describing the autocorrelation of a stationary sequence. However, most time series in real life are non-stationary. There are two methods for describing such non-stationary time series. One method involves a deterministic time trend, while the other method involves obtaining a stationary sequence through differentiation of the non-stationary sequence. Consider the following equation:

$$y_t = c + y_{t-1} + u_t \quad t = 1, 2, \cdots, T \tag{15}$$

where c is a constant and u_t is a stationary sequence. If the sequence y_t becomes a stationary sequence after d differences, but the sequence is not stationary after $d-1$ differences, then the sequence y_t is called a d th-order integral sequence, denoted by $y_t \sim I(d)$, then:

$$\omega_t = \Delta^d y_t = (1-L)^d y_t \tag{16}$$

Since ω_t is a stationary sequence, an ARMA(p, q) model can be established for ω_t:

$$\omega_t = \lambda_1 \omega_{t-1} + \cdots + \lambda_p \omega_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{17}$$

Expressed using a lag operator, this is:

$$\Phi(L)\omega_t = \Theta(L)\varepsilon_t \tag{18}$$

Among them:

$$\Phi(L) = 1 - \lambda_1 L - \lambda_2 L^2 - \cdots - \lambda_p L^p \tag{19}$$

$$\Theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q \tag{20}$$

After undergoing a d -order difference transformation, the $ARMA(p, q)$ model is referred to as

the $ARIMA(p,d,q)$ model, which is equivalent to the following equation:

$$\Phi(L)(1-L)^d y_t = \varepsilon + \Theta(L)\varepsilon_t \tag{21}$$

3.4. Steps for establishing an ARIMA model

1) Stationarity Test

Testing the stationarity of data is an important step in time series analysis. Stationarity is typically tested using time series plots and autocorrelation and partial correlation coefficient plots obtained through graphical testing methods. In this paper, we first make an intuitive judgment based on the time series plot, then use autocorrelation and partial correlation time series plots for further testing. If the data is non-stationary, the time series can be appropriately differenced and retested until stationarity is achieved. The order of differencing is the order d of the ARIMA(p, d, q) model.

2) Model identification and order determination

Model identification involves preliminarily determining the appropriate model type based on the autocorrelation function and partial autocorrelation function of the time series, as shown in the table below:

After determining the model based on the characteristics exhibited by the sequence, it is necessary to further determine the order of the model. In this paper, after identifying the model using Eviews software, the AIC statistic, SC statistic, and HC statistic are obtained. By comparing the sizes of the three statistics, the model fit is evaluated based on the minimum information criterion to determine the order of the model (p and q).

3) Parameter estimation and testing of the model

Parameter estimation for the model typically involves methods such as least squares estimation and maximum likelihood estimation. In this paper, the least squares estimation method is used to estimate the parameters, and Eviews software is used to obtain the estimated values, standard deviations, t-test statistics, and corresponding p-values for each parameter.

Residual testing for the model primarily involves verifying whether the residual sequence of the model estimation results meets the randomness requirement and whether it falls within the confidence interval, i.e., determining whether the residual sequence passes the white noise sequence test. If it passes, the model can be used for future predictions. If the model fails the test, it must be refitted until it passes the white noise test.

4. Practical Predictions on Employment Trends for College Students

This chapter will select 16,871 pieces of college student employment data from a certain institution of higher learning from 2017 to 2024 as the research object. Through the college student employment data visualization method proposed in this paper, the original employment data will be visualized and clustered, and the college student employment data dynamic analysis model constructed in this paper will be applied to analyze and predict the employment trends of college students.

4.1. Determining the number of clusters

In the actual data used in this study, 580 test data points were used to calculate the BIC and AIC for different numbers of clusters n, as shown in Figure 2. By analyzing the inflection points of the curve in the BIC-number of clusters coordinate system, the point where the overall slope of the curve changes significantly is identified as the optimal number of clusters. At n = 8, the slope of the curve transitions from a steep downward trend to a stable fluctuating trend, indicating this as the optimal number of clusters.

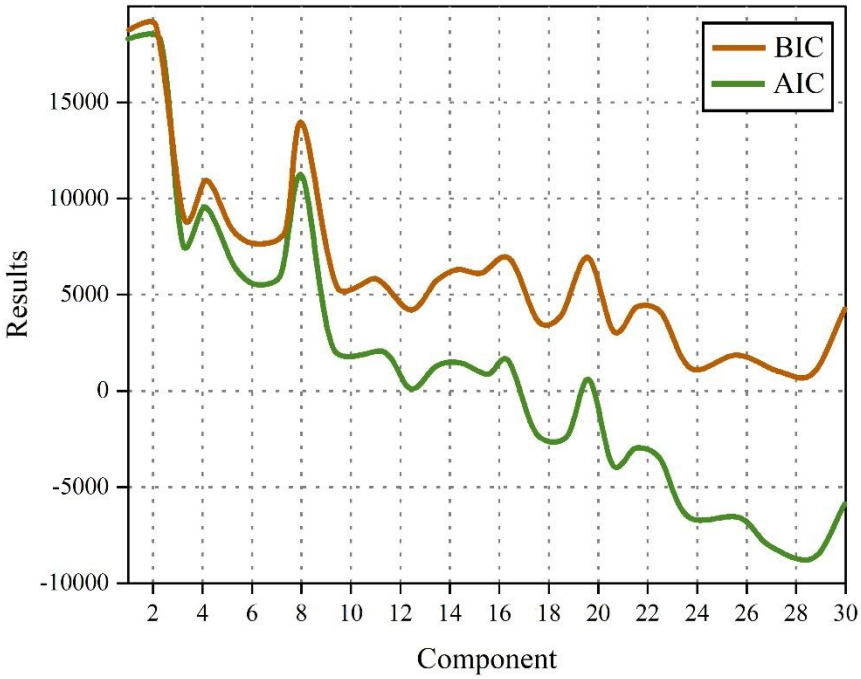
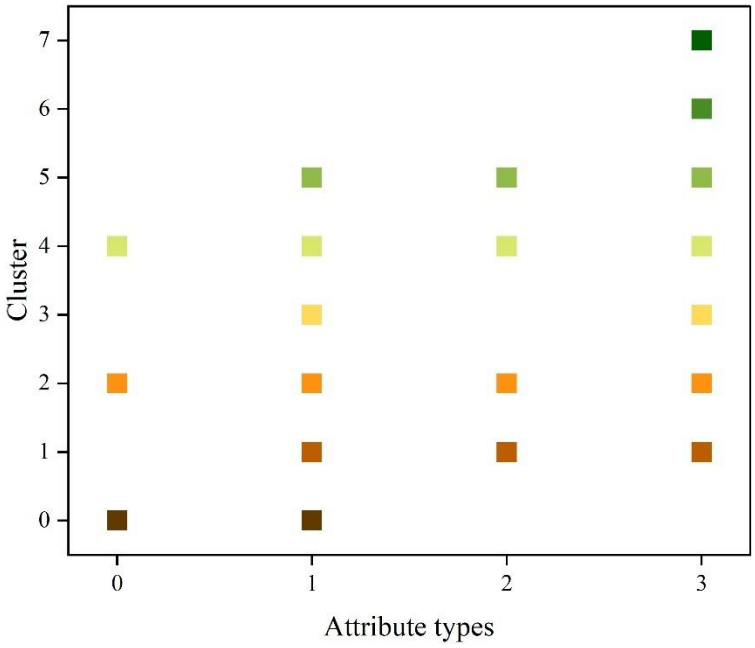
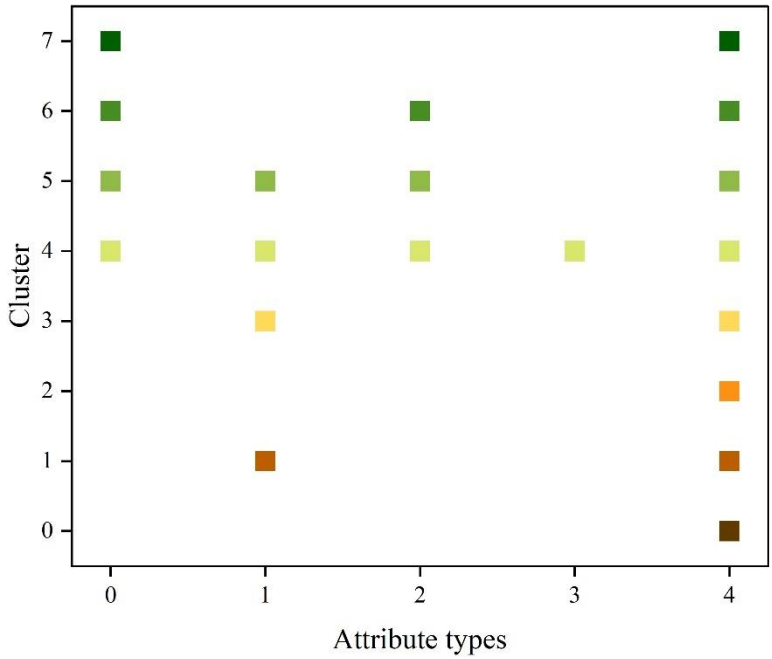


Figure 2. BIC AND AIC results under different clustering numbers.

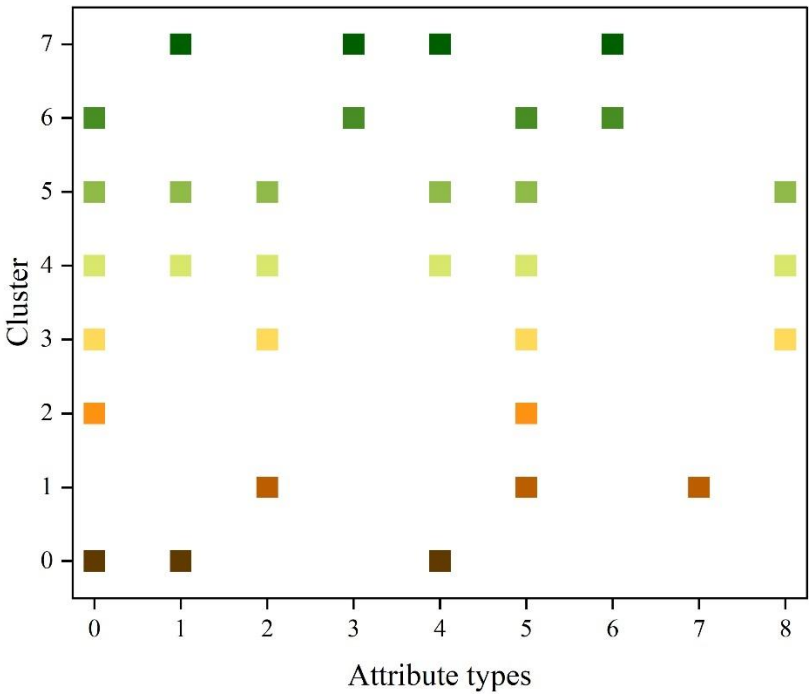
The clustering results for the following categories in the data—"Graduation Year," "Candidate Category," "Major," "Current Work Location," "Personal Strengths," "Hobbies and Interests," "Whether Current Job is First Job," "Method of Finding First Job," and "Industry of Current Job"—are displayed and analyzed separately. The specific relationships between different attributes and their corresponding clustering clusters are shown in Figure 3. Figures (a) to (i) correspond to the different attributes of "Graduation Year," "Candidate Category," "Major," "Current Work Location," "Personal Strengths," "Hobbies and Interests," "Whether Current Job is First Job," "How First Job was Found," and "Industry of Current Job," respectively.



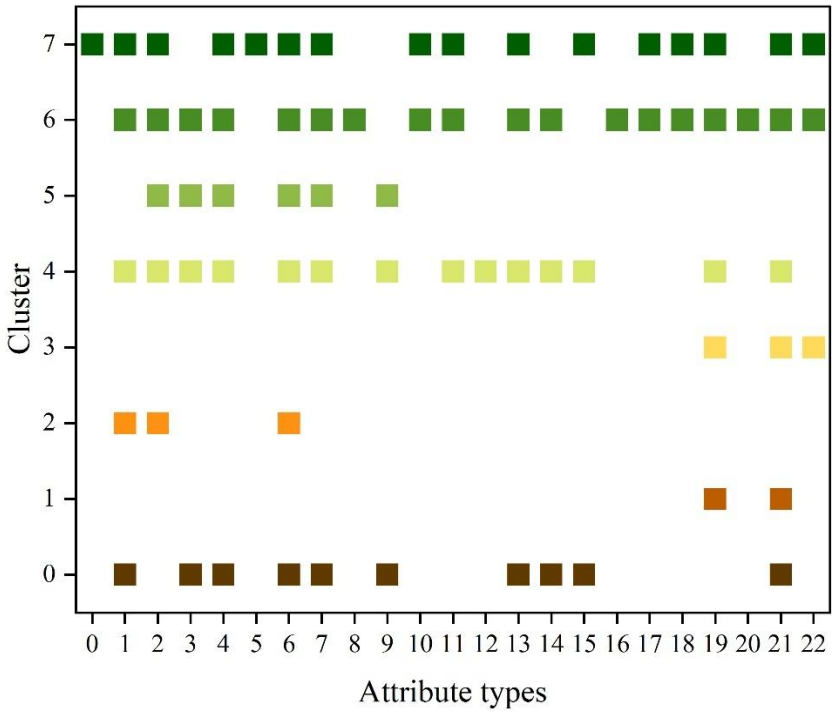
(a) Year of graduation



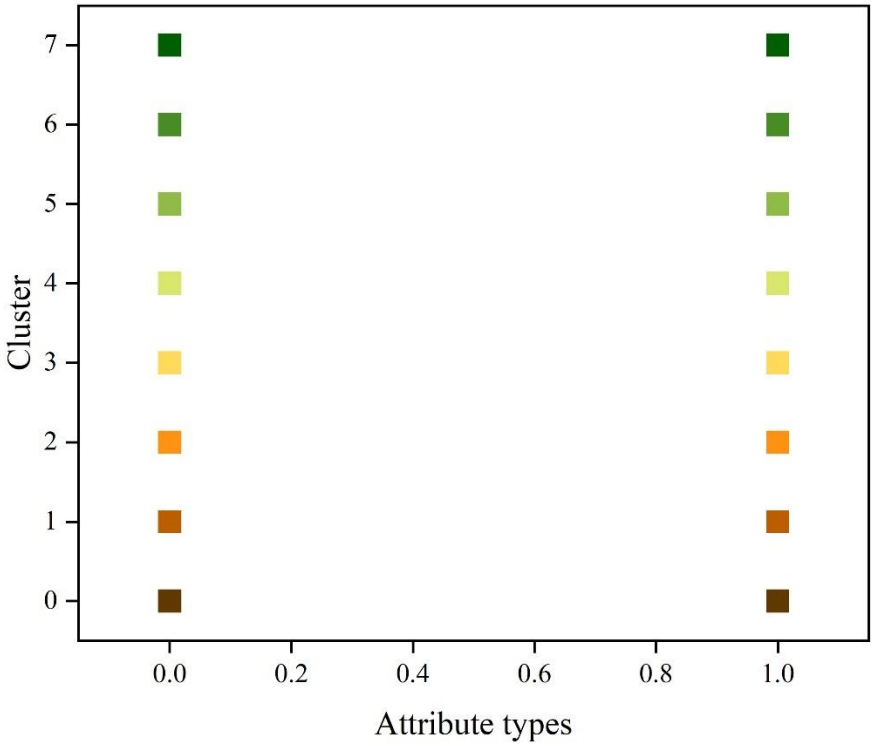
(b) Candidate category



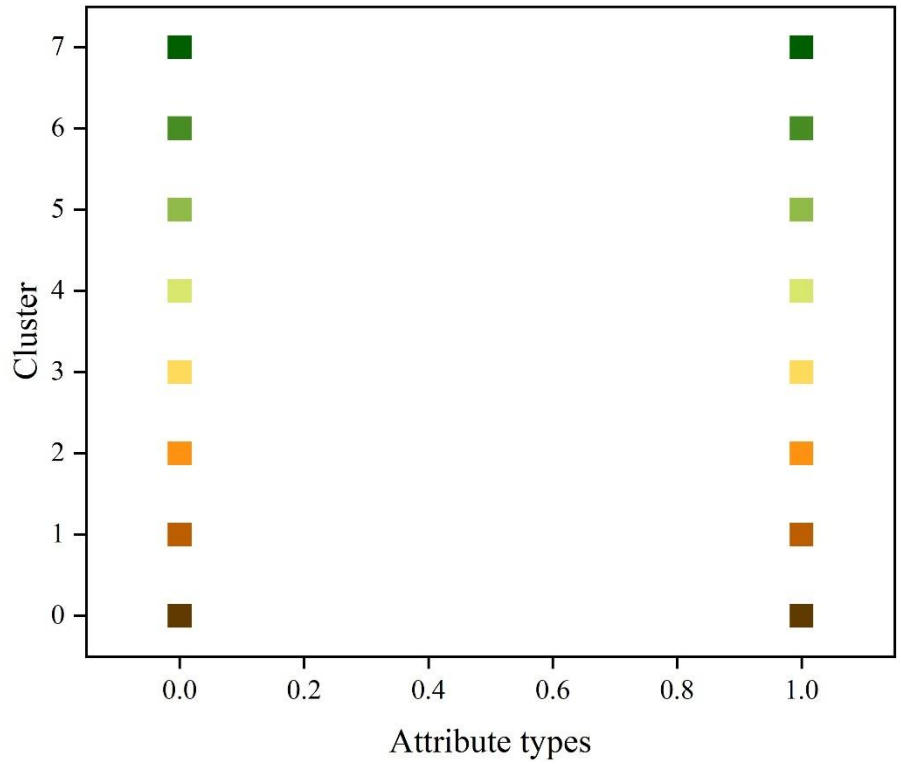
(c) Profession



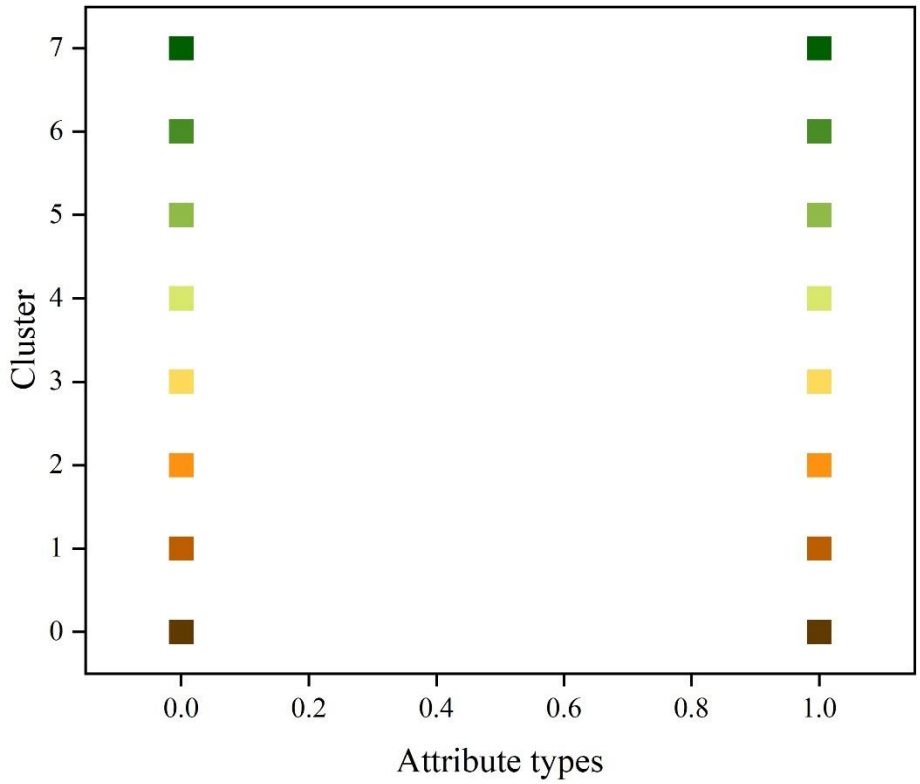
(d) Current place of work



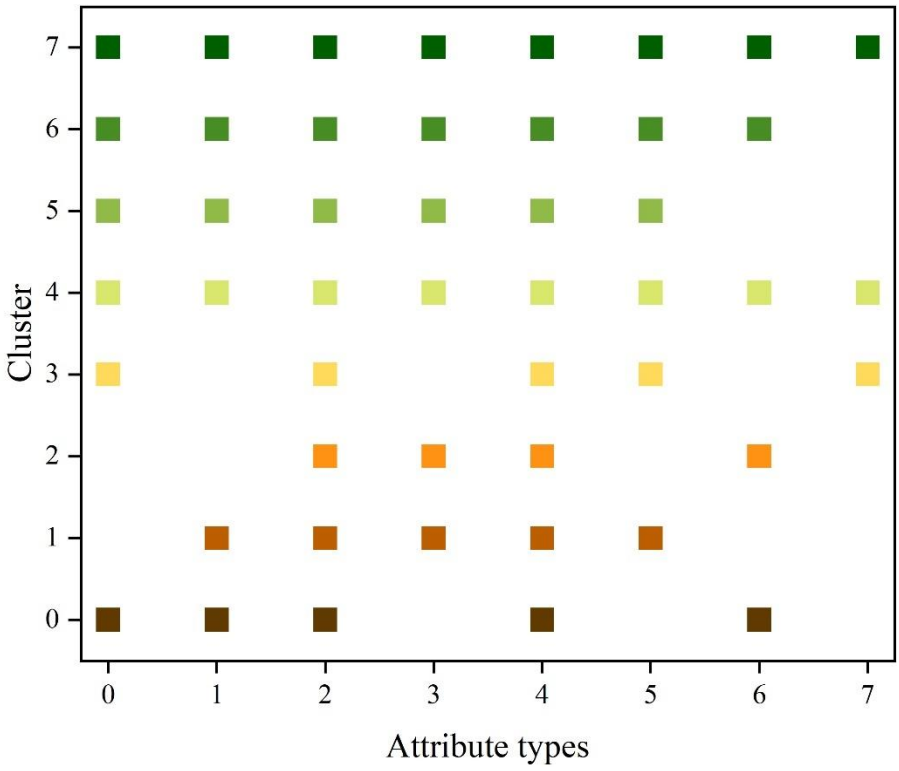
(e) Personal specialty



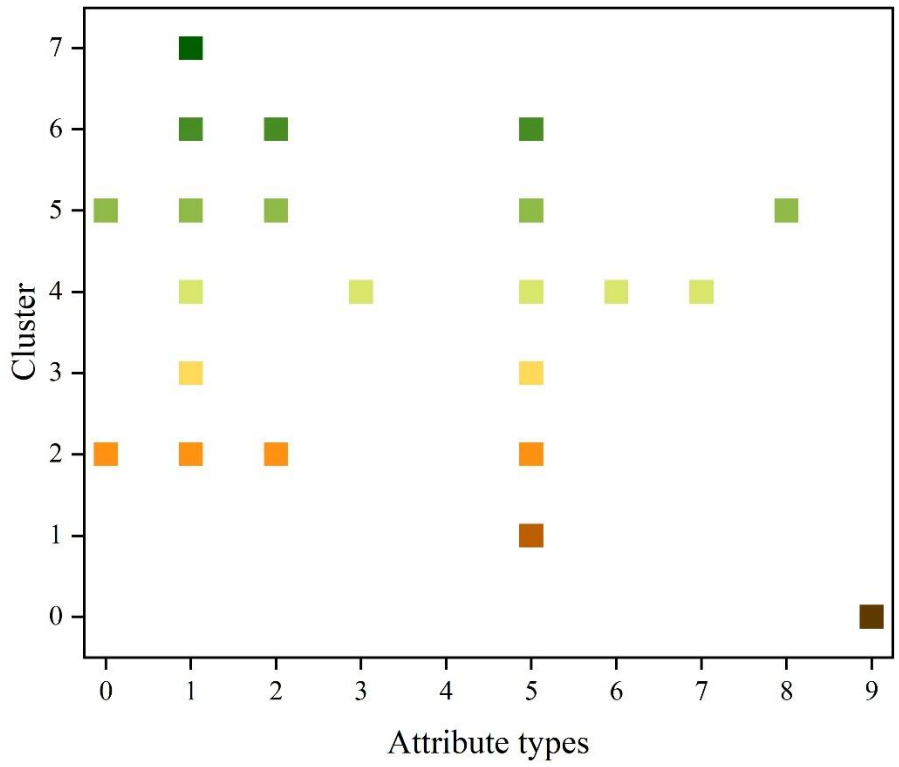
(f) Hobbies



(g) Whether the current job is the first job



(h)The way to find the first job



(i) The industry in which you are currently working

Figure 3. Cluster relationship corresponding to different attributes.

As shown in the figure, the first, second, and third items under “Candidate Category” correspond to the 3+ Certificate Program, independent enrollment, and others, respectively, and are all concentrated in the upper cluster. The first item, five-year higher education institutions, is distributed in the middle cluster. The fourth item, college entrance examination, is distributed across all clusters. This suggests that different tutoring and counseling strategies may need to be developed for members within these three clusters.

For the “Major” categories 3, 6, and 7, which correspond to Mechatronics Technology, Computer Application Technology, and Computer Network Technology, respectively, their clusters are concentrated in a small range of clusters toward the top. Category 8, Chinese Language Education, is concentrated in the middle three clusters. Category 5, English Education, has a relatively wide distribution of cluster clusters. Categories 1 and 4, corresponding to Business English and E-Commerce, have very similar cluster distribution patterns. Categories 0 and 2, corresponding to Business Management and Preschool Education, also have similar distributions, with slight differences in the lower clusters. Their distribution patterns are generally consistent with expectations. Categories 3, 6, and 7, which belong to the science and engineering category, exhibit potential internal similarities while also distinguishing themselves from other majors. The remaining categories have relatively broad distributions overall, with only minor differences in individual clusters.

For the “current industry” categories 3, 4, 6, 7, 8, and 9, which correspond to resident services, repairs, or other services; marketing; e-commerce; computer application technology; computer network technology; and finance, respectively, they are each concentrated within a single cluster. Among these, resident services, repairs, or other services; e-commerce; and computer application technology all belong to the same cluster. Marketing and computer network technology belong to the same cluster. During career counseling, different employment guidance plans may need to be developed for industries with significant differences in cluster classifications.

Since students within the same cluster may share potential similarities, intra-cluster communication can more efficiently help students obtain valuable reference information, making it a necessary step in employment data processing.

4.2. Employment Trend Forecast

Employment data is seasonal, indicating that it exhibits time series characteristics and can be used as a source of dynamic data.

4.2.1. Determining Model Parameters and Types

There are many time series models, and to confirm the model category and parameters, it is necessary to use the autocorrelation function and partial autocorrelation function. The autocorrelation function data is shown in Table 1. Sig is less than 0.05, indicating that there are autocorrelation factors and that not all are white noise sequences. The autocorrelation coefficient has a transmissibility issue, so to shield its transmissibility, it is necessary to calculate its partial autocorrelation coefficient.

Table 1. Autocorrelation coefficient.

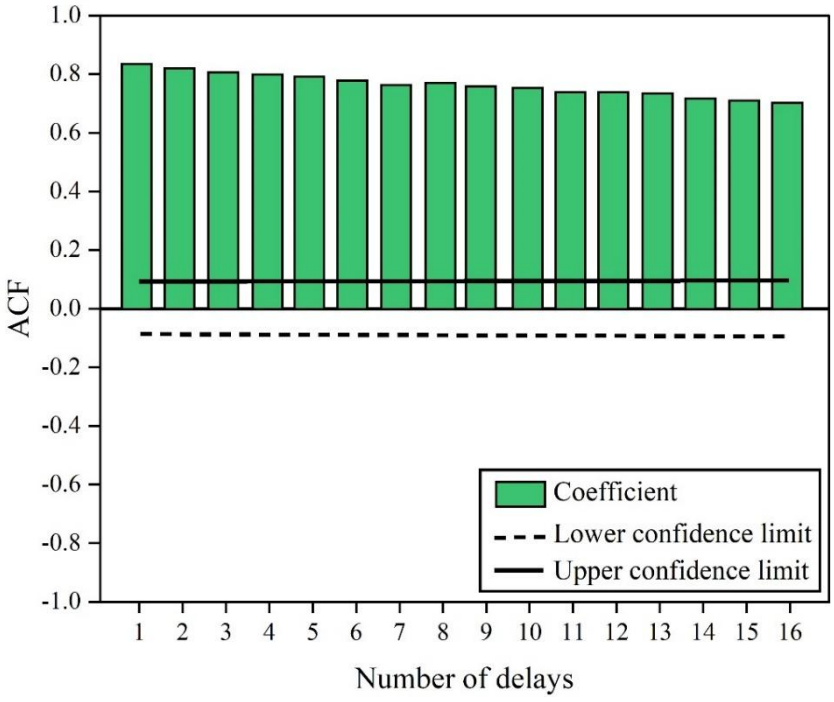
Lag	Autocorrelation	Standard error	Box-Ljung statistic		
			value	Df	Sig
1	0.934	0.42	373.793	1	0
2	0.963	0.42	745.054	2	0
3	0.968	0.42	1094.093	3	0
4	0.92	0.42	1445.981	4	0
5	0.931	0.42	1787.86	5	0
6	0.924	0.41	2119.924	6	0
7	0.957	0.41	2445.031	7	0
8	0.92	0.41	2768.179	8	0
9	0.863	0.41	3083.832	9	0
10	0.896	0.41	3392.484	10	0
11	0.85	0.41	3688.073	11	0
12	0.878	0.41	3992.058	12	0
13	0.852	0.41	4281.844	13	0
14	0.898	0.41	4567.864	14	0
15	0.811	0.41	4582.558	15	0
16	0.849	0.41	5131.079	16	0

The partial autocorrelation coefficients are shown in Table 2. From the partial autocorrelation coefficient table, we can see that the standard error is 0.42. To determine which time model it belongs to, we also need to know its truncation and tailing conditions.

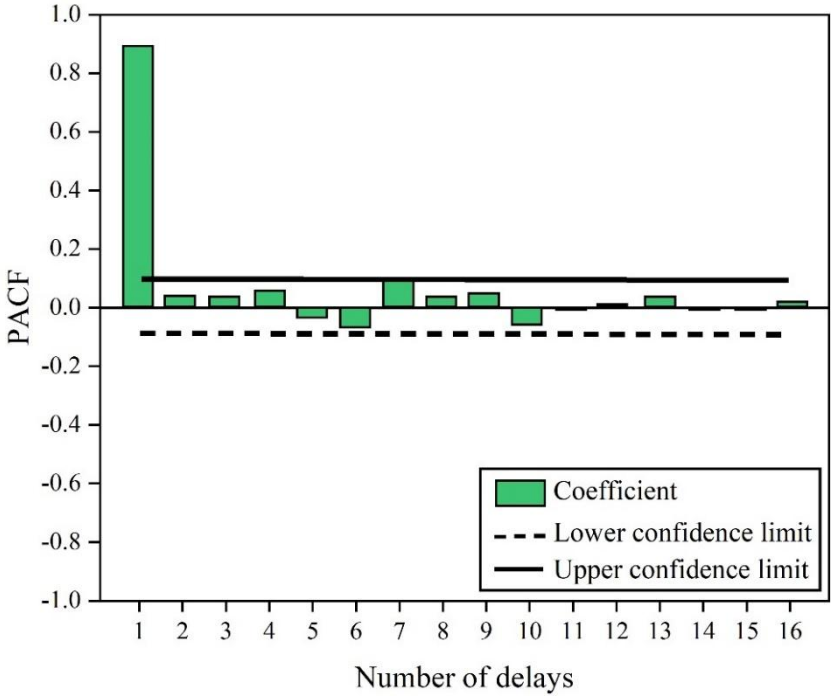
Table 2. Partial autocorrelation coefficient.

Lag	Partial autocorrelation	S tandard error
1	0.986	0.42
2	0.057	0.42
3	0.024	0.42
4	0.061	0.42
5	-0.017	0.42
6	-0.039	0.42
7	0.09	0.42
8	0.029	0.42
9	0.063	0.42
10	-0.057	0.42
11	0.005	0.42
12	0.025	0.42
13	0.037	0.42
14	-0.014	0.42
15	-0.003	0.42
16	0.024	0.42

The ACF and PACF test results are shown in Figure 4. Figures (a) and (b) correspond to the ACF plot and PACF plot, respectively. From the figures, it can be seen that it is a first-order difference, and it can also be seen that it is first-order truncated and first-order tailed, which is consistent with the ARIMA (1,1,1) model.



(a) ACF



(b) PACF

Figure 4. ACF and PACF.

4.2.2. Forecasting and Testing of Time Series Models

After determining that the model is an ARIMA (1,1,1) model, you can model and plot the graph based on the parameters. Using the expert modeling function in SPSS, you can plot the results as shown in Figure 5. As can be seen from the figure, there is a significant node change in week 242, so the data is divided into two parts.

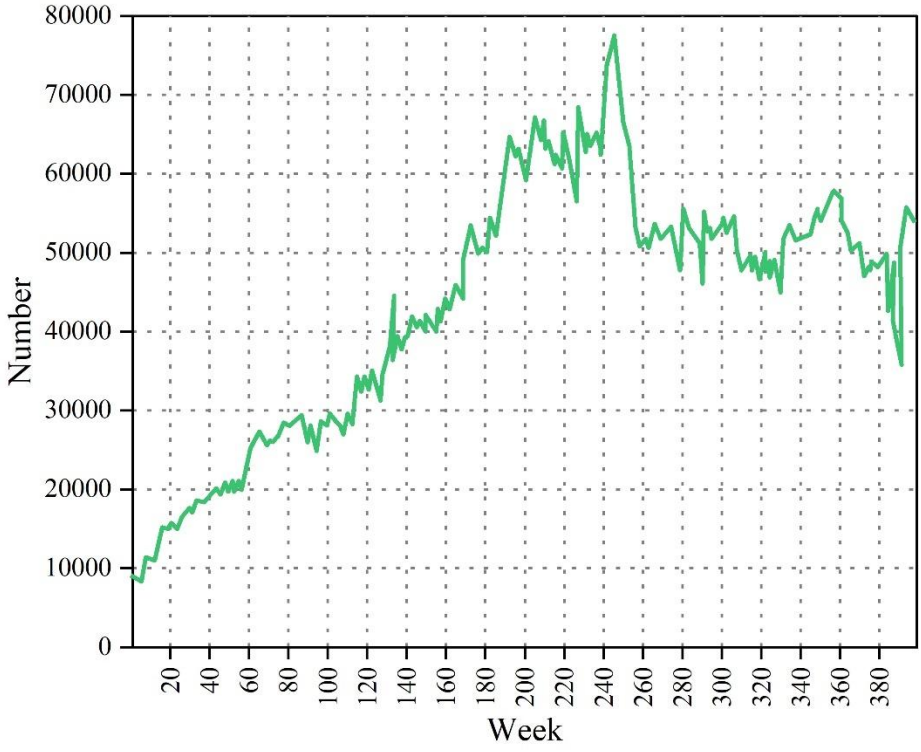


Figure 5. Expert modeling diagram.

To verify the effectiveness of the simulation model, we can next test it against future employment trends. Save and apply the model, make predictions in the model bar, and display the actual data in the same chart. Using a regression prediction model established with college graduate employment data from 2017 to 2023, we can test the model's performance by predicting data for 2024. The model test results are shown in Figure 6. As can be seen, the predicted employment trends for 2024 exhibit stable cyclical changes, aligning well with actual data, indicating good model fit.

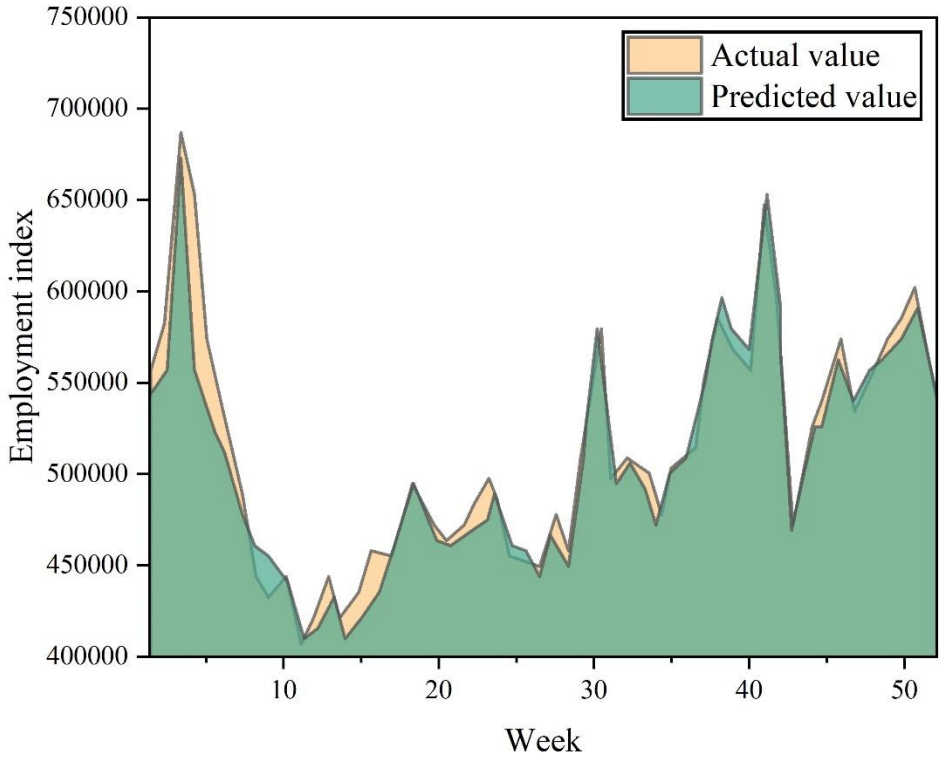


Figure 6. Model calibration.

A forecast of employment trends for the year 2025 is presented, with the specific results shown in Figure 7. The employment index for the coming year continues to exhibit cyclical fluctuations. It reaches its peak in Week 4, which marks the height of the spring recruitment season. Graduates who miss this window will find it difficult to secure satisfactory employment, while companies will continue to hire to replace employees who have resigned or left during the fall recruitment period. The index then declines rapidly, indicating that companies will not engage in large-scale hiring once their recruitment processes conclude, leading to a downward trend in employment. The trend begins to recover again after six months, or Week 30. During this phase, job seekers begin preparing for the fall recruitment season, and companies gradually release relevant employment and recruitment information. The employment index then rises rapidly over the next several weeks until it reaches its peak. The fall recruitment season is larger in scale and longer in duration than the spring recruitment season, with companies offering more positions. College students tend to seek employment during this period to secure satisfactory jobs. After the fall recruitment season concludes, the year comes to a close, and employment attention returns to normal levels. The curve aligns well with real-world social conditions.

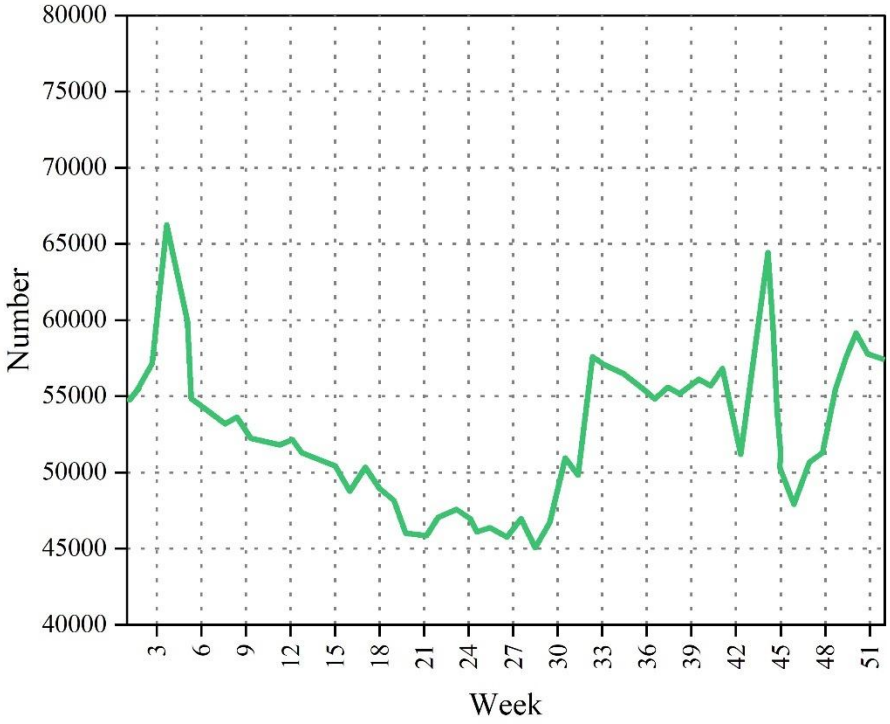


Figure 7. Employment index forecast.

5. Conclusion

This paper stores the post-processing employment data of college students from higher education institutions into a big data repository, performs clustering analysis, calculates the relationships between the data, and completes data visualization. Using time series analysis methods, a dynamic analysis model for college student employment data is constructed to predict employment trends among college students. A dataset of 16,871 college student employment records from a certain higher education institution spanning the years 2017 to 2024 is selected as the research subject. A practical study on predicting employment trends among college students is conducted to validate the effectiveness of the dynamic analysis model for college student employment data proposed in this paper.

Using 580 test data points, we calculated the BIC and AIC values for different numbers of clusters n and determined that the optimal number of clusters for the employment data is 8. An analysis of the clustering patterns for different attributes in the employment data revealed that the "candidate category" attribute, specifically items 0, 2, and 3 (corresponding to 3+ certificates, independent admissions, and others), are concentrated in the upper clusters, while item 4 (college entrance exam) is distributed across all clusters. The clustering distributions of different professional categories under the "Major" attribute and different industry categories under the "Current Industry" attribute show significant differences, necessitating the development of distinct employment guidance plans. Additionally, intra-cluster communication can facilitate the acquisition of more valuable employment reference information.

After determining the number of clusters in the employment data and completing the preprocessing of the data, the parameters and type of the dynamic analysis model for college graduate employment data were further determined. Based on the results of the ACF and PACF tests, the ARIMA (1,1,1) model was selected. A regression prediction model was established using employment data for college students from 2017 to 2023, and the model was tested on data for the year 2024 to confirm its fitting effectiveness. The results showed that the predicted employment trends for 2024 exhibited stable cyclical changes, which were generally consistent with actual data, indicating good fitting performance. The model was applied to predict the employment trends for college students in 2025. The employment index for the entire year of 2025 continues to exhibit cyclical fluctuations, specifically: the employment index reaches its peak during the fourth week (spring recruitment period) and then rapidly declines, recovering during the autumn recruitment period after the 30th week and then rising sharply over the next several weeks until it reaches its highest point. After the fall recruitment period, the employment index returns to normal levels. The employment trend forecasts align with actual social employment conditions, validating the effectiveness of the model presented in this paper.

References

[1] Hooley, T., & Rice, S. (2019). Ensuring quality in career guidance: A critical review. British Journal of Guidance & Counselling, 47(4), 472-486.

[2] Carson, R. D., & Reed, P. A. (2015). Pre-college career guidance on student persistence and performance at a small private university. Career and Technical Education Research, 40(2), 99-112.

[3] Guo, L., Sangsawang, T., Vipahasna, P. P., Pigultong, M., Punyayodhin, S., & Darboth, K. (2024). Statistical approach to evaluating the efficacy of career guidance programs on university graduate employability in China. Journal of Applied Data Sciences, 5(1), 279-293.

[4] Yoon, Y. R., & Yang, A. K. (2017). Career Guidance through Analysis of the Effects of College Counseling Program. International Information Institute (Tokyo). Information, 20(9B), 6843-6850.

[5] Gati, I., Levin, N., & Landman-Tal, S. (2019). Decision-making models and career guidance. International handbook of career guidance, 115-145.

[6] Li, J., & Xu, Z. (2021, October). Design and realization of college student employment management system based on intelligent optimization algorithm. In 2021 2nd Artificial Intelligence and Complex Systems Conference (pp. 7-11).

[7] Li, J., & Ma, Y. (2024). Employment management system for universities based on improved decision tree. Journal of Intelligent Systems, 33(1), 20230138.

[8] Han, I., & Shin, W. S. (2016). The use of a mobile learning management system and academic achievement of online students. Computers & Education, 102, 79-89.

[9] Xia, P., Li, T., Gao, T., & Wang, Y. (2016). Design and Implementation of Employment Management System Based on B/S. International Journal of Advanced Pervasive and Ubiquitous Computing (IJAPUC), 8(4), 75-104.

[10] Richins, G., Stapleton, A., Stratopoulos, T. C., & Wong, C. (2017). Big data analytics: opportunity or threat for the accounting profession?. Journal of information systems, 31(3), 63-79.

[11] Cobelli, N., Bonfanti, A., Cubico, S., & Favretto, G. (2019). Quality and perceived value in career guidance e-services. International Journal of Quality and Service Sciences, 11(1), 53-68.

[12] Zhou, F., Xue, L., Yan, Z., & Wen, Y. (2020). Research on college graduates employment prediction model based on C4. 5 algorithm. In Journal of Physics: Conference Series (Vol. 1453, No. 1, p. 012033). IOP Publishing.

[13] He, S., Li, X., & Chen, J. (2021, May). Application of data mining in predicting college graduates employment. In 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 65-69). IEEE.

[14] Vicente, M. R., López-Menéndez, A. J., & Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing?. Technological Forecasting and Social Change, 92, 132-139.

[15] Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. Applied Sciences, 11(1), 237.

[16] Ansari, G. A. (2017). Career guidance through multilevel expert system using data mining technique. International Journal of Information Technology and Computer Science, 9(8), 22-29.

[17] Liu, J., & Wang, N. (2023). Research on precise employment data analysis and practice of university graduates based on web system implementation. International Journal of Data Science, 8(2), 138-151.

[18] Wei, Y., Zheng, Y., & Li, N. (2023). Big Data Analysis and Forecast of Employment Position Requirements for College Students. International Journal of Emerging Technologies in Learning, 18(4).

[19] Dong, X. (2021). Prediction of college employment rate based on big data analysis. Mathematical Problems in Engineering, 2021(1), 1421356.

[20] Liu, Z. (2024, August). Research on Behavior Analysis and Employment of College Students Based on Big Data. In Proceedings of the 2024 International Conference on Big Data and Digital Management (pp. 160-168).

[21] Lim, H. W., & Kim, S. J. (2021). A study on ways to make employment improve through Big Data analysis of university information public. International Journal of Advanced Culture Technology, 9(3), 174-180.

[22] Fan, H. (2020, October). A Prediction Model of College Students' Employment Based on Data Mining. In 2020 13th International Conference on Intelligent Computation Technology and Automation (ICICTA) (pp. 549-552). IEEE.

[23] Zhang, Y. (2023). Application of data mining based on improved ant colony algorithm in college students' employment and entrepreneurship education. Soft Computing, 1-10.

[24] Qi, M. B. (2021, October). Clustering mining method of college students' employment data based on feature selection. In International Conference on Advanced Hybrid Information Processing (pp. 105-115). Cham: Springer International Publishing.

[25] Otgonbayar Agvaan,Gordon Cichon,Uuganbaatar Dulamragchaa,Hyun chul Kim,Seonuck Paek & Tseren Onolt Ishdorj. (2025). Optimal approximate computation of Euclidean distance in spiking neural P systems framework. Scientific Reports,15(1),18047-18047.

[26] Diego Germán Guamán M. & Juan Carlos Herrera M. (2025). Heuristic Approach Based on a K-Means Algorithm to Reduce the Cost of Macroscopic Fundamental Diagram Estimation. Transportation Research Record,2679(4),337-352.

[27] Dong-Gu Lee,Je-Doo Ryu,Keon-Seok Nam & Kyoung-Nam Ha. (2019). Time Series Analysis of Outcomes for Small and Medium Enterprises' Support of Regional Industry. Proceedings of Engineering and Technology Innovation,12,35-38.