

基于主成分和聚类分析的我国档案行业人员变动影响因素分析

纪文迪¹, 潘月君^{1*}, 潘 瑛²

¹山东理工大学数学与统计学院, 山东 淄博

²山西北方恒通动力系统有限公司人事行政部, 山西 大同

收稿日期: 2025年5月28日; 录用日期: 2025年7月16日; 发布日期: 2025年7月28日

摘 要

本文运用统计学方法, 对影响我国档案行业人员变动情况的因素进行了深入剖析, 运用主成分分析、聚类分析等方法确定各影响因素与人员变动之间的定量关系并加以验证。结果表明, 三十四岁及以下人员流入流出率、工资福利、档案学本科专业率、接待人次、馆藏量等因素在不同时期对我国档案行业人员的变动有着不同影响。本研究为各级各类档案馆(室)制定合理的人才政策和发展规划以及人才队伍培养提供了科学的依据和指导。

关键词

档案行业, 人员变动, 主成分分析, 聚类分析, 影响因素模型

Analysis of Factors Influencing Personnel Changes in Archival Industry of China Based on Principal Component and Cluster Analysis

Wendi Ji¹, Yuejun Pan^{1*}, Ying Pan²

¹School of Mathematics and Statistics, Shandong University of Technology, Zibo Shandong

²HR & Admin Department, Shanxi North Hengtong Power System Co., Ltd., Datong Shanxi

Received: May 28th, 2025; accepted: Jul. 16th, 2025; published: Jul. 28th, 2025

Abstract

The influencing factors of personnel changes in China's archival industry is analyzed by statistical methods in this paper. Principal component analysis, cluster analysis and other methods were used

*通讯作者。

文章引用: 纪文迪, 潘月君, 潘瑛. 基于主成分和聚类分析的我国档案行业人员变动影响因素分析[J]. 社会科学前沿, 2025, 14(7): 705-713. DOI: 10.12677/ass.2025.147665

to determine the quantitative relationship between each influencing factor and personnel change and verify it. The results show that factors such as the inflow and outflow rate of personnel aged 34 and below, wages and benefits, the rate of undergraduate majors in archives, the number of receptions, and the volume of library collections have different impacts on the changes of personnel in our country's archives industry in different periods. This study provides a scientific basis and guidance for the formulation of reasonable talent policies and development plans and the cultivation of talent teams in various archives (offices) at all levels.

Keywords

Archival Industry, Personnel Changes, Principal Component Analysis, Cluster Analysis, Influencing Factor Model

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着我国经济的快速发展和产业结构的不断调整，各行业人员变动现象日益突出。深入分析我国大行业人员变动情况及影响因素，对于把握劳动力市场动态、制定合理政策等具有重要价值，能为行业发展提供有力支撑[1]。本文通过对大量数据的分析，能够为政策制定者提供客观、准确的信息，使他们在制定人才政策、规划行业发展时更加科学合理，可以避免盲目决策，提高决策的科学性和有效性。

本文数据主要来源于国家档案局官网：考虑到国家档案局官网则是获取官方动态数据的重要渠道，因此由此调取了我国国家档案局官网 2014~2023 年的年度报告以及基本情况摘要、财政摘要等，获得了真实有力的官方数据。

通过对国家档案局关于各级各类档案馆人员数据的整理和分析，绘制如下表格，从表中可以看出，我国档案行业人员数量在 2018 年前呈现下降趋势，2018 年成为分水岭，推测大概率与 2018 年我国党政机构改革，档案局馆分立有关，2018 年后则呈现出缓慢增长的趋势。随着社会对档案工作的重视程度和档案行业的发展，人员数量逐渐增长。近年来，增长速度趋于平稳，这可能与档案行业发展进入相对成熟阶段以及人员结构优化调整有关。

为了全面剖析人员流动的影响因素，本研究选取了多个关键变量进行考察。薪资福利作为影响人员流动的重要因素之一，关系到档案行业人员的职业满意度，因此，本研究将薪资福利作为核心变量之一，旨在探究其对人员流动的具体影响。同时，人员的年龄结构和学历结构对于其职业发展路径和流动趋势具有不可忽视的作用，笔者也将年龄与学历水平指标纳入考察范围，用来表示其对人员流动的影响机制。接待人次作为衡量的关键指标，它能够清晰地展示各级各类档案馆在社会服务中的角色和影响力。馆藏量作为衡量档案馆资源丰富程度的关键指标，代表着档案馆的档案材料对公众的满足程度。丰富的馆藏资源能够吸引更多的研究人员、学者以及普通公众前来查阅和利用，从而间接影响人员的流动。因此，本研究旨在通过综合考察这些关键变量，更深入地揭示档案行业人员流动的内在规律。为后续的深入研究奠定基础。

我国 2014~2023 年各级档案馆人员变动数据如表 1 所示。

2. 影响因素的相关性分析与回归分析

2.1. 相关性分析

为了更好地反映人员流动情况，选用人员流入率和流出率代替简单的人员总数作为统计指标，避免

了采用总数所造成的单一片面的结论，得出的结果会更具代表性。人员流入流出率具体计算方法如下：

Table 1. Data on the annual change of archivists at all levels and types of archives
表 1. 各级各类档案馆档案人员年度变动数据

年度	人员总数	中央级	省级(副省级)	地级	县级
2014	48,495	628	3526	10,845	33,496
2015	49,190	651	3800	10,710	34,029
2016	48,871	642	3447	11,028	33,754
2017	48,505	647	3455	10,682	33,721
2018	46,412	653	3511	9598	31,672
2019	41,495	647	4044	8763	27,741
2020	40,496	634	4089	8645	27,378
2021	41,393	642	4338	8804	27,609
2022	42,134	651	4395	8992	28,096
2023	42,732	647	4423	9169	28,493

人员流动人数 = 当年(某指标)统计人数 - 上年末(某指标)统计人数

人员流入流出率 = 当年(某指标)人员流动人数 ÷ 期初(某指标)人员总人数

补充：正值代表该年新入职人员多出的数量，负值代表该年离职人员多出的数量。结果如表 2 (从上往下每行数据依次为 2014~2023 年)。

Table 2. Variable data summary
表 2. 变量数据汇总

总体人员流入流出率	本科及以上学历人员流入流出率	三十四岁及以下人员流入流出率	工资福利
0.014	0.051	-0.002	6001.60
-0.006	0.019	-0.053	6342.19
-0.007	0.021	-0.027	7171.79
-0.043	0.017	-0.092	11009.69
-0.106	-0.067	-0.095	14882.19
-0.024	0.015	0.042	15417.92
0.022	0.041	0.020	15737.29
0.018	-0.105	0.065	14867.97
0.014	0.221	0.105	15698.67

Table 3. Correlation analysis
表 3. 相关性分析

			总体人员流入 流出率	本科及以上学历 人员流入 流出率	三十四岁及以 下人员流入 流出率	工资 福利
斯皮尔曼 Rho	总体人员流入 流出率	相关系数	1.000	0.410	0.686*	0.192
		显著性(双尾)	.	0.273	0.041	0.620

续表

	N	9	9	9	9
本科及以上人员 流入流出率	相关系数	0.410	1.000	0.267	0.000
	显著性(双尾)	0.273	.	0.488	1.000
	N	9	9	9	9
三十四岁及以下 人员流入流出率	相关系数	0.686*	0.267	1.000	0.450
	显著性(双尾)	0.041	0.488	.	0.224
	N	9	9	9	9
工资福利	相关系数	0.192	0.000	0.450	1.000
	显著性(双尾)	0.620	1.000	0.224	.
	N	9	9	9	10

*. 在 0.05 级别(双尾), 相关性显著。

从表 3 中 Spearman 相关系数的数据中,我们可以深入了解各级各类档案馆 2014~2023 年间的总体人员流入流出率与本科及以上人员流入流出率、三十四岁及以下人员流入流出率以及工资福利这三个变量间的相关关系,为了更精准地衡量这些变量之间的关联性。结果显示,各级各类档案馆 2014~2023 年间三十四岁及以下人员流入流出率这个变量与总体人员流入流出率呈现出了显著性相关,这表明它们之间存在统计上可识别的关联。具体来说,相关系数值为 0.686,所有值大于 0,表明总体人员流入流出率与这个变量之间存在正相关关系,正相关意味着当一个变量的值增加时,另一个变量的值倾向于增加,反之亦然。因此,通过相关性分析可以看出,人员变动与队伍的年轻程度呈现正相关关系,这意味着当前档案行业的人才队伍结构在发生着变化,逐步呈年轻化态势,可能与年轻人活力更高,干劲更足且更具创新力有关。然而,值得注意的是,这个数值在统计学意义上相对较低,表明这种正相关关系在强度上并不是很高,无法归为高度相关,可能受到其他因素的干扰。另一方面,总体人员流入流出率与本科及以上人员流入流出率和工资福利之间的相关系数值分别为 0.410 和 0.192,小于 0.5,则不具备很强的相关性。这表示近年来人员流入流出情况与本科及以上人员流入流出率和工资福利在统计上没有显著的关联,或者说它们之间的关联性非常微弱,以至于在当前的样本量和统计显著性水平下无法被识别,因此无法直接判断高学历以及福利待遇与人员的流动变化有无显著影响,或许需要其他统计方法进行进一步讨论。

2.2. 回归分析

基于相关性分析的结果,使用 SPSS 软件对该案例尝试构建多元线性回归模型,以进一步确定各影响因素对档案行业人员数量变动率的影响程度。

Table 4. ANOVA^a
表 4. 方差分析表^a

	模型	平方和	自由度	均方	F	显著性
1	回归	0.012	4	0.003	7.064	0.042 ^b
	残差	0.002	4	0.000		
	总计	0.013	8			

a. 因变量: 总体人员流入流出率; b. 预测变量: (常量), 接待人次, 工资福利, 三十四岁及以下人员流入流出率, 档案学本科专业率。

Table 5. Coefficient^a
表 5. 系数^a

模型	未标准化系数		标准化系数	显著性		共线性统计	
1	B	标准错误	Beta	t		容差	VIF
三十四岁及以下人员流入流出率	1.087	0.287	1.838	3.794	0.019	0.132	7.570
工资福利	-7.709E-6	0.000	-0.810	-1.543	0.198	0.112	8.903
档案学本科专业率	-0.510	0.680	-0.380	-0.750	0.495	0.121	8.290
接待人次	0.001	0.001	0.900	1.994	0.117	0.152	6.562

a. 因变量：总体人员流入流出率。

对回归模型进行拟合优度检验、显著性检验和多重共线性检验。通过分析上表的结果，R 方以及调整后的 R 方均>0.7，说明模型拟合度较高，自变量能够较好的预测因变量的变化，模型对数据的拟合效果好，能够较好地解释档案行业人员数量变动率的变化；表 4 中 F 检验结果显著，说明回归方程整体是显著的。各变量的方差膨胀因子(VIF)均小于 10，自变量之间不存在严重的多重共线性问题，意味着模型中自变量相对独立，能够被清晰的区分和估计。但各自变量的显著性并不理想，表 5 中只有三十四岁以下人员的流入流出率显著性 < 0.05，进一步证实了相关性分析的结果，肯定了档案行业人才队伍的年轻化态势，其他变量并不显著，因此多元线性统计模型并不能较好的反映影响因素变化，考虑使用主成分分析来进行进一步的研究。

3. 影响因素的主成分分析

主成分分析(PCA)是一种基于数据降维和特征提取的统计方法，其核心思想是将众多具有相关性的原始变量，通过线性变换的方式，转化为一组互不相关的综合指标，即主成分，并使这些主成分尽可能多地保留原始数据的信息[2]。

3.1. 主成分分析过程

3.1.1. 对数据标准化处理及因子分析适用性检验

首先，选取 2015~2023 年的数据，变量包括三十四岁及以下人员流入流出率、工资福利、档案学本科专业率、接待人次、馆藏量。利用 SPSS 软件将原始数据进行标准化处理，得到新的数据[3]。

Table 6. Correlation matrix
表 6. 相关性矩阵

	Zscore (档案学本科专业率)	Zscore (接待人次)	Zscore (馆藏量)	Zscore (工资福利)	Zscore (三十四岁及以下人员流动率)
Zscore (档案学本科专业率)	1.000	0.266	0.874	0.933	0.416
Zscore (接待人次)	0.266	1.000	-0.091	0.247	-0.681
相关性 Zscore (馆藏量)	0.874	-0.091	1.000	0.853	0.743
Zscore (工资福利)	0.933	0.247	0.853	1.000	0.441
Zscore (三十四岁及以下人员流动率)	0.416	-0.681	0.743	0.441	1.000

Table 7. KMO and Bartlett test
表 7. KMO 和巴特利特检验

KMO 取样适切性量数。		0.626
巴特利特球形度检验	近似卡方	38.979
	自由度	10
	显著性	0.000

然后通过 KOM 检验法和 Bartlett 球体检验法进行因子分析的适用性检验。KOM 值等于 $0.626 > 0.6$ ，说明指标间存在一定的相关性。Bartlett 球体检验结果为 38.979，Sig 值小于 0.001，说明拒绝相关系数是单位阵，即各个指标是相关的。表 6，表 7 两项结果检验都表明了该数据适用因子分析法[4] [5]。

3.1.2. 主成分分析过程

Table 8. Common factor variance
表 8. 公因子方差

	初始	提取
Zscore (档案学本科专业率)	1.000	0.960
Zscore (接待人次)	1.000	0.971
Zscore (馆藏量)	1.000	0.964
Zscore (工资福利)	1.000	0.945
Zscore (三十四岁及以下人员流动率)	1.000	0.975

Table 9. Total variance explained
表 9. 总方差解释

成分	初始特征值			提取载荷平方和		
	总计	方差百分比	累积%	总计	方差百分比	累积%
1	3.170	63.401	63.401	3.170	63.401	63.401
2	1.646	32.910	96.312	1.646	32.910	96.312
3	0.088	1.768	98.080			
4	0.070	1.404	99.484			
5	0.026	0.516	100.000			

由表 8，表 9 可知，前两个主成了解释了全部方差的 96.312%，说明提取的 2 个主成分能够代表原来各指标信息的 96.312%，所提取的主成分评价档案行业人员变动原因已有一定的把握。因此，提取 2 个主成分，分别为 Y_1 和 Y_2 。

Table 10. Ingredient matrix
表 10. 成分矩阵

	成分	
	1	2
Zscore (馆藏量) x_3	0.980	-0.061
Zscore (档案学本科专业率) x_1	0.922	0.333

续表

Zscore (工资福利) x ₄	0.921	0.311
Zscore (三十四岁及以下人员流动率) x ₅	0.714	-0.682
Zscore (接待人次) x ₂	-0.047	0.984

根据表 10 中 2 个主成分系数，首先分别对其开根号，得到 1.780 和 1.282 两个数值，继续计算得到 Y₁、Y₂ 的线性组合：

$$Y_1 = 0.517x_1 - 0.026x_2 + 0.55x_3 + 0.517x_4 + 0.401x_5$$
$$Y_2 = 0.259x_1 + 0.767x_2 - 0.047x_3 + 0.242x_4 - 0.531x_5$$
$$y = 0.63401*y_1 + 0.3291*y_2$$

由上式可知，在主成分 Y₁ 中，档案学本科专业率(x₁)、馆藏量(x₃)、工资福利(x₄)的系数绝对值大于其他变量的系数绝对值，所以主成分 Y₁ 是三个影响指标的综合反映，说明用这三个指标来考察各级各类档案馆的人员变动情况是必不可少的，三者都在人员变动中起到重要的作用。

在主成分 Y₂ 中，接待人次(x₂)、三十四岁及以下人员流入流出率(x₅)的系数绝对值大于其他变量的系数绝对值，所以主成分 Y₂ 主要是由这两个影响指标来综合反映，说明用这两个指标来考察各级各类档案馆的人员变动情况也是必不可少的，二者在人员变动中起到重要的作用[6]。

3.2. 综合得分与结果

根据上述所得的线性组合，计算 2015 至 2023 九年间的综合得分与排名表。

Table 11. Comprehensive score & ranking table
表 11. 综合得分与排名表

样本	Y ₁ 得分	Y ₂ 得分	综合得分	排名
2015	-1.92	-0.89	-1.55	9
2016	-1.52	-0.72	-1.23	8
2017	-1.56	-0.74	-1.26	7
2018	0.21	0.98	0.45	5
2019	0.34	1.16	0.61	4
2020	0.68	-0.32	0.35	6
2021	0.95	0.64	0.84	2
2022	1.42	-0.25	0.81	3
2023	1.81	-0.10	1.12	1

表 11 是根据各个主成分方差贡献率占两个主成分总方差贡献率的比率为权重计算的综合得分及排名。它代表了每个观测对象在所有原始变量综合影响下的相对位置或水平。综合得分越高，往往说明该观测对象在这些变量所代表的综合特征上表现越突出。对于年份变化来说，如果年份的综合得分排名逐年上升，说明该时间段内所研究的对象在综合表现上呈现出不断改善和发展的趋势。

针对上述 2015~2023 年的综合得分及排名变化表，我们可以看出主成分分析后的各年份综合得分及排名大致呈现逐年上升的趋势，这表明最近几年来影响我国档案行业变动的各因素综合作用越来越明显，

且处于不断的发展完善中。2015~2017 年处于低分阶段，存在馆藏资源不足、工资福利低(Y_1 极低)，导致专业人才流失的问题(Y_2 中 x_5 流动率高)。2018~2020 年处于转型阶段，馆藏数字化投入增加(x_3 提升)，但接待量(x_2)受疫情影响下降，人员流动率短暂降低(Y_2 波动)。2021~2023 年处于高分阶段，面临新的机遇与挑战，尽管资源投入(Y_1)持续优化，但接待量回升与年轻人才流动率(x_5)同步上升，反映工作压力与职业吸引力不匹配，需注意“人才粘性”问题”。

4. 聚类分析人员变动的影响因素

聚类分析是一种无监督的统计分析方法，它根据“物以类聚”的原理，将一组物理的或抽象的对象分组成为由类似的对象组成的多个类(簇)的过程，是通过数据内在的相似性将数据集中的样本划分为若干个互不相交的子集(称为簇或类)，使得同一簇内的样本相似度高，而不同簇之间的样本相似度低。

4.1. 聚类分析的过程

使用 SPSS 软件对 2015~2023 年间我国各级各类档案馆三十四岁及以下人员流入流出率、档案学本科专业率、工资福利、接待人次及馆藏量这五个变量使用系统聚类分析法进行聚类，使用平方欧氏距离，绘制谱系图[7]。

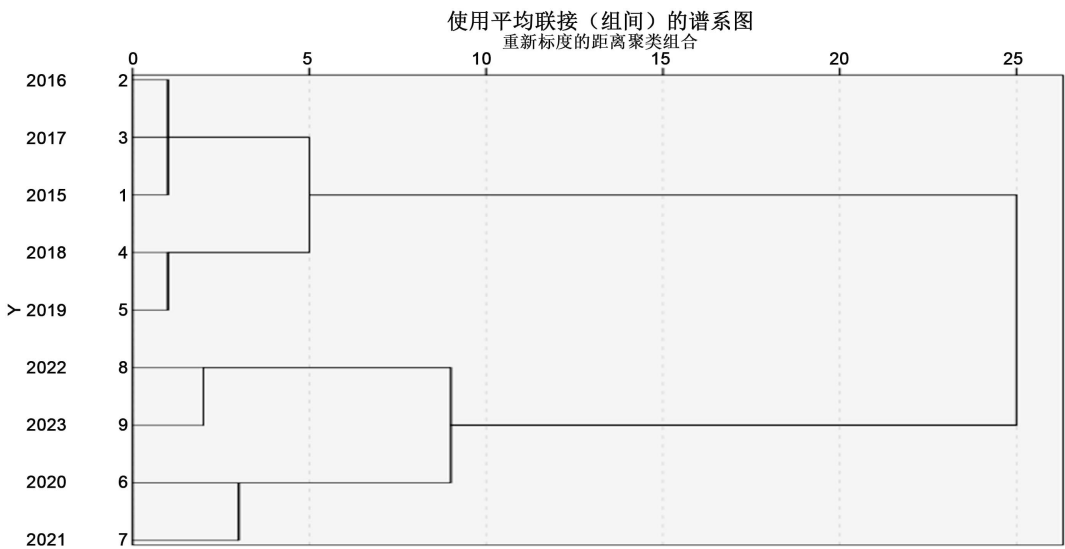


Figure 1. Pedigree chart
图 1. 谱系图

首先对图 1 所提供的数据进行描述性分析：

聚类 1：我们将其特征总结为低薪资、低专业率、人员稳定。此时 34 岁以下人员流动率接近 0 或轻微负值(-0.002 至 -0.053)。工资福利从 6001.60 元升至 7171.79 元/月。档案学本科专业率为 6.58%至 7.23%。接待人次为 638.3 至 659.4 人次/日。馆藏量从 58, 641.7 至 65, 371.1 卷。

聚类 2：我们将其特征总结为高薪资、高专业率、人才净流入。此时 34 岁以下人员流动率为正值(0.020 至 0.105)。工资福利从 14867.97 元升至 15737.29 元/月。档案学本科专业率从 12.02%升至 14.61%。接待人次从 645.9 至 697.3 人次/日。馆藏量为 104671.1 至 126, 846.5 卷。

聚类 3：可以将其特征总结为中薪资、中专业率、人员流出风险。此时 34 岁以下人员流动率为负值 (-0.092 至 -0.027)。工资福利从 11009.69 元升至 15482.19 元/月。档案学本科专业率从 11.80%提升至

12.73%。接待人次为 716.4 至 724.9 人次/日。馆藏量为 75051.1 至 91789.8 卷[6]。

4.2. 聚类分析的结论

根据聚类特征, 可以将年份与类别对应起来:

1) 2015~2019 年: 低薪资、低专业率、人员稳定。

在这一阶段, 34 岁以下人员流动率接近 0 或轻微负值 $[(-0.002)\sim(-0.053)]$, 说明青年人才流动极少, 可能因薪资低(6000~7100 元)、职业发展空间有限。档案学本科专业率为 7%左右, 显示专业化程度较低, 可能依赖非科班人员或老员工。馆藏量增长从 5.8 万增至 8.2 万卷, 增速平缓。推测与 2014 年《关于加强和改进新形势下档案工作的意见》出台有关, 但基层落实较慢。

2) 2020~2021 年: 薪资跃升、专业率提高、人员流出风险。

在这一阶段, 可以看出薪资有大幅增长, 2020 年达到 15,482 元, 可能与机构改革后财政投入增加有关。专业率提升: 从 7%升至 12.7%, 但仍低于 2023 年水平。人员流动率: 2018 年出现明显负值 (-0.092) , 可能因机构合并导致编制调整, 青年人员主动离职。推测是因为自 2018 年机构改革以来, 部分档案局与一些党史办、地方志合并, 在未来几年间引发短期人事动荡。

3) 2022~2023 年: 高薪资、高专业率、人才净流入。

在这一阶段, 薪资峰值 2022 年达 15,737 元, 2023 年略降, 但仍保持高位(15,698 元)。专业率突破 14%: 显示出行业对科班人才的需求增强。2023 年人员净流入率 $+0.105$, 青年人才吸引力显著提升。馆藏量突破 12 万卷, 但接待人次稳定, 说明档案数字化减少了线下压力。推测是由于《“十四五”全国档案事业发展规划》(2021 年)推动了专业化与数字化。

5. 结论

本文从五个方面构建了档案行业人员变动的分析框架, 突破了以往单因素分析的局限, 为了解档案行业人才发展状况提供了更全面的理论视角。验证了“资源 - 需求 - 人才”的动态平衡理论, 发现当服务需求增长超过资源供给时, 会加剧人才流失, 这一发现为档案行业可持续发展理论提供了新的实证支持。提出了“专业人才粘性”概念, 发现专业素质与福利待遇的匹配度是影响人才留存的关键因素, 当两者失衡时, 如高专业率匹配低待遇时, 反而会加速人才流失[1]。

本文虽然运用多种统计方法对我国档案行业人员变动情况及影响因素进行了较为深入的分析, 但仍存在一些不足之处。数据的收集比较有限, 只有近十年以来的数据, 通过之后几年更为详尽的数据, 期待能够做出更为准确的研究, 未来的研究可以进一步扩大数据来源, 完善影响因素的量化指标, 采用更先进的统计方法和模型进行分析, 提高研究的准确性和可靠性。

参考文献

- [1] 钱德凤. 档案治理视角下档案人才队伍建设研究[J]. 档案与建设, 2022(12): 53-56.
- [2] 张文彤, 董伟. SPSS 统计分析高级教程[M]. 第 2 版. 北京: 高等教育出版社, 2013.
- [3] Field, A. (2013) Discovering Statistics Using SPSS. 4th Edition, Sage Publications.
- [4] Tabachnick, B.G. and Fidell, L.S. (2013) Using Multivariate Statistics. 6th Edition, Pearson Education.
- [5] 任雪松, 于秀林. 聚类分析[M]. 第 2 版. 北京: 中国统计出版社, 2009.
- [6] 王协舟, 李音奇. 新时代档案人才队伍现代化建设研究[J]. 兰台内外, 2023(12): 4-7.
- [7] 郭志刚. 社会统计分析方法——SPSS 软件应用[M]. 第 2 版. 北京: 中国人民大学出版社, 2010.