**Statistics and Application** 统计学与应用, 2021, 10(6), 963-974
Published Online December 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.106101

Hans汉斯

# 一种基于加权均方误差系数的评价因子筛选方法

**张英雪[1]，吴有富[2*]，许　婷[1]**

[1]贵州民族大学数据科学与信息工程学院，贵州　贵阳
[2]贵州交通职业技术学院，贵州　贵阳

收稿日期：2021年11月16日；录用日期：2021年11月30日；发布日期：2021年12月16日

## 摘　要

**评价因子的筛选一直是统计分析中的热门话题，目前对因子筛选方法很多，如综合指数法、全局主成分分析方法等；这些方法在特定的环境中均得到充分的运用，但是当因子间的相关性较强时，这些方法的分析不理想，如在交通助推农村产业的分析中就得不到与实际相符的结果。为了克服此问题，本文提出了一种加权均方误差系数法；并以贵州交通对农村产业的影响为例进行实证分析。实验结果表明，我们的方法是有效。**

## 关键词

**评价因子，综合指数，全局主成分分析，加权均方误差系数**

# An Evaluation Factor Screening Method Based on Weighted Mean Square Error Coefficients

**Yingxue Zhang[1], Youfu Wu[2*], Ting Xu[1]**

[1]School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang Guizhou
[2]Guizhou Vocational and Technical College of Communications, Guiyang Guizhou

Received: Nov. 16th, 2021; accepted: Nov. 30th, 2021; published: Dec. 16th, 2021

## Abstract

**Evaluation factors screening has always been a hot topic in statistical analysis. At present, there**

*通讯作者。

文章引用: 张英雪, 吴有富, 许婷. 一种基于加权均方误差系数的评价因子筛选方法[J]. 统计学与应用, 2021, 10(6): 963-974. DOI: 10.12677/sa.2021.106101

**are many methods for screening factors, such as the integrated index method, the global principal component analysis method, etc.; these methods are fully used in specific environments, but when the correlation between factors is strong, the analysis of these methods is not ideal. For example, in the analysis of the traffic boosting rural industry, the results are not in line with the actuality. To overcome this problem, a weighted mean square error coefficient method is proposed in this paper; and the impact of transportation on rural industries in Guizhou is used as an example for empirical analysis. The experimental results show that our method is effective.**

## Keywords

**Evaluation Factor, Integrated Index, Global Principal Component Analysis, Weighted Mean Square Error Coefficient**



## 1. 引言

评价因子进行筛选一直都是学者们关注的点，然而现在对评价因子筛选的方法很多，例如综合指数、全局主成分分析等方法，无论何种方法，评价指标及其数量的选取、指标权重与评价标准的设定是两个非常重要的方面，并且选择合理的评价指标和评价方法是得出合理结论的前提。张艳芹[1] (2001)通过对均值化后的数据进行标准化系数法加权，进而计算得指标的综合指数，并对综合指数进行非参数检验，从而选取对企业评价的指标。徐雅静[2] (2006)通过变量聚类与全局主成分分析相结合的方法，对我国普通高等教育发展水平进行评价。但是当因子间的相关性较强时，这些方法的分析不理想，如在交通助推农村产业的分析中就得不到与实际相符的结果。为了克服此问题，我们提出了一种加权均方误差系数法。

## 2. 指标筛选原理

本文方法的提出是基于综合指数和全局主成分分析局限性，故在此简述这两种方法。

### 2.1. 综合指数法

该方法是基于加权平均的推广。主要包括两个过程：评价指标的无量纲处理和权重的确定。具体步骤如下：

1) 无量纲处理：通过均值化、极值标准化等方法对数据进行无量纲处理，得到预处理的数据 $Z_{ij}$

2) 权重的计算：通过计算指标的均值($\bar{x}_j$)和标准差($s_j$)，得到标准差系数

$$V_j = \frac{s_j}{\bar{x}_j} \tag{1}$$

将 $V_j$ 归一化处理，得权重：

$$w_j = \frac{V_j}{\sum_{j=1}^{n} V_j} \tag{2}$$

3) 计算综合指数：

$$F_i = \sum_{j=1}^{n} Z_{ij} \times w_j \tag{3}$$

## 2.2. 全局主成分分析法

全局主成分分析方法的具体步骤如下：

1) 建立时序立体数据表 $x = \left(x^1, x^2, \cdots, x^t\right)_{Tn\times p} = \left(x_{ij}\right)_{Tn\times p}$，$Tn$ 表示样本个数，$p$ 表示指标数量。

2) 对数据进行标准化。

3) 定义全局数据表的重心：

$$g = \sum_{t=1}^{T}\sum_{i=1}^{n} q_i^t e_i^t \tag{4}$$

$q_i^t$ 表示 $t$ 时刻样本点 $e_i$ 的权重，进而得到全局协方差矩阵 $V$，也即是 $x$ 的相关系数矩阵：

$$V = \sum_{t=1}^{T}\sum_{i=1}^{n} q_i^t \left(e_i^t - g\right)\left(e_i^t - g\right)' \tag{5}$$

4) 求协方差 $V$ 的前 $m$ 个特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$，及对应的特征向量 $\mu_1, \mu_2, \cdots, \mu_m$。

5) 计算主成分及方差贡献率，第 $h$ 主成分为 $F_h = \mu_h' x$；

方差贡献率：

$$a_k = \frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i} \tag{6}$$

累计方差贡献率：

$$a_1 + a_2 + \cdots + a_n = \frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \tag{7}$$

选出的主成分 $F_1, F_2, \cdots, F_m$ 使累计方差贡献率接近 85%。

6) 求因子载荷矩阵 $A = \left(r_{ij}\right)$，$r_{ij}$ 是 $x_i$ 和 $F_i$ 的相关系数，$r_{ij}$ 表示第 $i$ 个变量 $x_i$ 在第 $j$ 个共因子 $F_j$ 上的负荷。

7) 计算指标的主成分系数，由主成分分析结果的因子载荷矩阵中第 $i$ 列数值除以对应第 $i$ 个特征值的开方求得。

8) 求指标权重：

$$w_m = \sum_{i=1}^{p} \frac{a_{mi} * \alpha_i}{p} \tag{8}$$

$a_{mi}$ 表示第 $i$ 个主成分中第 $m$ 个基础指标的系数。

9) 求综合评价函数：

$$F = \sum_{i=1}^{m} \frac{\lambda_i}{q} * f_i \tag{9}$$

$q$ 表示各主成分的特征根之和，$f_i$ 是未经标准化的第 $i$ 个主成分得分。

# 3. 加权均方误差系数法

基于上述综合指数法和全局主成分分析法的局限性和思想，本文提出了加权均方误差系数评价因子

筛选方法。通过计算两序列之间的均方误差(MSE)来反应两组序列之间的波动情况，进而算得序列之间的显著系数，由于每个因子所做的贡献不一致，及所占权重不一样，通过标准差系数法计算权重，从而算得显著度。显著度越大表明与参考序列关联性越强，则表明该评价因子有效。具体步骤如下：

1) 确定参考序列和比较序列

2) 计算参考序列与比较序列之间的均方误差：

$$\frac{\left(x_0(k)-x_i(k)\right)^2}{n}\left(k=1,\cdots,n;i=1,\cdots,m\right) \tag{10}$$

$n$ 表示样本个数，$m$ 表示指标个数。

3) 计算显著系数：

$$\xi_i(k)=\frac{\min\limits_i\min\limits_k\frac{\left(x_0(k)-x_i(k)\right)^2}{n}+\rho\cdot\max\limits_i\max\limits_k\frac{\left(x_0(k)-x_i(k)\right)^2}{n}}{\frac{\left(x_0(k)-x_i(k)\right)^2}{n}+\rho\cdot\max\limits_i\max\limits_k\frac{\left(x_0(k)-x_i(k)\right)^2}{n}} \tag{11}$$

其中 $\rho$ 为分辨系数，一般取 $\rho=0.5$。

4) 通过标准差系数法计算权重：计算标准差系数：

$$V_k=\frac{\sigma_k}{\mu_k}\left(k=1,\cdots,n\right) \tag{12}$$

其中，$\mu_k$ 和 $\sigma_k$ 分别为第 $k$ 个样本的均值和标准差，对标准差系数进行归一化处理，计算权重：

$$w_k=\frac{V_k}{\sum_{k=1}^{n}V_k} \tag{13}$$

5) 计算显著度：

$$r_i=\sum_{i=1}^{n}w_k\cdot\xi_i(k) \tag{14}$$

## 4.实证分析

为了检验我们方法的有效性，我们选取贵州交通对农村产业的影响进行实证分析。由于所选的样本相关性较强，参考徐静雅(2006)的变量聚类与选取典型相关法相结合的思想，在进行贵州交通对农村产业的影响分析时，先对交通变量进行系统聚类，再运用加权均方误差系数法进行典型选取，进而选出交通对农村产业的影响因子。

### 4.1. 构建指标体系和数据来源

交通助推农村产业革命，交通指标从公路里程、投资、车辆、旅客等各个方面选取，农村产业选取第一产业、第二产业、第三产业。因此本文通过贵州省宏观经济库[1]选取贵州省 2001 年~2019 年有关交通和第一产业、第二产业、第三产业的数据，分析哪些交通指标分别对第一产业、第二产业和第三产业有显著促进作用，对第一产业、第二产业和第三产业没有促进作用或促进不明显，应于剔除。对交通指标和第一产业、第二产业和第三产业的符号说明见表 1。由于指标之间存在较强相关性，全都选取用来分析，会存在冗余的现象，介于此要对 32 个交通指标进行指标筛选。

[1]数据来源：贵州省宏观经济数据库(https://guizhou.gov.cn/)。

**Table 1.** Symbol description
**表 1.** 符号说明

| 指标 | 符号 |
|---|---|
| 第一产业(亿元) | $y_1$ |
| 第二产业(亿元) | $y_2$ |
| 第三产业(亿元) | $y_3$ |
| 一级公路(公里) | $x_1$ |
| 二级公路(公里) | $x_2$ |
| 三级公路(公里) | $x_3$ |
| 四级公路(公里) | $x_4$ |
| 等外公路(公里) | $x_5$ |
| 公路路网密度(公里/百平方公里) | $x_6$ |
| 高速公路车辆通行费收入(万元) | $x_7$ |
| 民用汽车拥有量(万辆) | $x_8$ |
| 载客汽车(万辆) | $x_9$ |
| 载货汽车(万辆) | $x_{10}$ |
| 私人汽车拥有量(万辆) | $x_{11}$ |
| 公路货物运输量(万吨) | $x_{12}$ |
| 公路货物周转量(亿吨公里) | $x_{13}$ |
| 公路旅客运输量(万人) | $x_{14}$ |
| 公路旅客周转量(亿人公里) | $x_{15}$ |
| 公路总里程(公里) | $x_{16}$ |
| 等级公路里程(公里) | $x_{17}$ |
| 高速公路里程(公里) | $x_{18}$ |
| 路网及农村公路建设投资(万元) | $x_{19}$ |
| 汽车站场投资(万元) | $x_{20}$ |
| 重点公路投资(万元) | $x_{21}$ |
| 公路建设投资(万元) | $x_{22}$ |
| 交通固定资产投资(亿元) | $x_{23}$ |
| 交通运输、仓储和邮电通信业(亿元) | $x_{24}$ |
| 国道(公里) | $x_{25}$ |
| 省道(公里) | $x_{26}$ |
| 县道(公里) | $x_{27}$ |
| 乡道(公里) | $x_{28}$ |
| 建制村通油路率(%) | $x_{29}$ |
| 乡镇通公路率% | $x_{30}$ |
| 建制村通公路率% | $x_{31}$ |
| 乡镇通油路率% | $x_{32}$ |

## 4.2. 相关分析

在进行指标筛选前先分别计算交通指标与第一产业、第二产业和第三产业的相关系数，通过计算指标之间的 Pearson Correlation 分析指标之间存在正向还是反向的相关关系。相关系数的计算公式为：

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} \tag{15}$$

从相关系数表 2 可以看出三级公路($x_3$)、等外公路($x_5$)和乡镇通公路率($x_{30}$)分别和第一产业、第二产业和第三产业的相关系数都较低，甚至三级公路($x_3$)对第一产业存在负相关，说明即使偏远郊区的公路里程逐渐增加，但是对于第一产业、第二产业和第三产业的促进作用不明显，应于剔除。在 2017 年贵州已经实现了建制村通油路率达 100%，黔货能出山，人也能走出去，进而越来越多的年轻人出去打工、进厂，成为留守老人和留守儿童的人越来越多，从而导致农村劳动力低下，渐渐的地荒了，所以，第一产业与三级公路($x_3$)、等外公路($x_5$)和乡镇通公路率($x_{30}$)的相关系数都较低甚至为负，符合现状。第二产业和第三产业分别主要指工业和服务业，即使乡村的道路通行，对于工业和服务业的是存在促进作用，但是促进作用不显著。高速公路车辆通行费收入($x_7$)、民用汽车拥有量($x_8$)、载客汽车($x_9$)、高速公路里程($x_{18}$)和第一产业的相关程度很大，相关系数都高达 0.99。民用车辆和高速公路里程的增加，人们想去哪里都方便。贵州各个有特色的地方都通路了，民用车辆的增加和高速公路增加相结合促进旅游业和新农家乐的发展，进而第三产业得到了显著的促进。四级公路($x_4$)、高速公路车辆通行费收入($x_7$)、载货汽车($x_{10}$)、公路货物运输量($x_{12}$)、公路货物周转量($x_{13}$)、等级公路里程($x_{17}$)、高速公路里程($x_{18}$)和第二产业的相关系数也是高达 0.99，主要都是和货物相关的数据，和第二产业息息相关。民用汽车拥有量($x_8$)、私人汽车拥有量($x_{11}$)、公路货物周转量($x_{13}$)、等级公路里程($x_{17}$)和第三产业的相关系数也高达 0.99，第三产业主要是服务业，现在人民的生活水平提升，不再满足于物质需求，更依赖于满足精神需要，通过民用车辆拥有量的增加，可以看出人民经济提升后，更依赖于享受服务。根据第一产业、第二产业和第三产业与交通的相关系数得出以上结果都是符合贵州省的现状。

**Table 2.** Correlation coefficients for the three industries and transport indicators
**表 2.** 三大产业与交通指标的相关系数

| 交通指标 \ 产业 | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|
| $x_1$ | 0.947130518 | 0.915162851 | 0.911640951 |
| $x_2$ | 0.985299667 | 0.971916544 | 0.980917771 |
| $x_3$ | −0.065178691 | 0.018328808 | 0.002368529 |
| $x_4$ | 0.981369039 | 0.991649386 | 0.989711754 |
| $x_5$ | 0.188878478 | 0.273995233 | 0.257460123 |
| $x_6$ | 0.828755255 | 0.873757304 | 0.864858153 |
| $x_7$ | 0.994505 | 0.991616931 | 0.989427909 |
| $x_8$ | 0.99230012 | 0.988009761 | 0.992292284 |
| $x_9$ | 0.990840426 | 0.983380518 | 0.988686163 |
| $x_{10}$ | 0.97496967 | 0.991324197 | 0.987426858 |
| $x_{11}$ | 0.989106085 | 0.987529863 | 0.991849698 |
| $x_{12}$ | 0.978495631 | 0.992683115 | 0.974668151 |

**Continued**

| | | | |
|---|---|---|---|
| $x_{13}$ | 0.989360594 | 0.996391581 | 0.992009132 |
| $x_{14}$ | 0.492065359 | 0.510354215 | 0.490308595 |
| $x_{15}$ | 0.86635105 | 0.907784302 | 0.894217357 |
| $x_{16}$ | 0.8287966 | 0.873794397 | 0.864920278 |
| $x_{17}$ | 0.98230157 | 0.99216596 | 0.990579667 |
| $x_{18}$ | 0.995739538 | 0.995599224 | 0.986380728 |
| $x_{19}$ | 0.855706843 | 0.835394277 | 0.787233029 |
| $x_{20}$ | 0.907555674 | 0.888572124 | 0.881739101 |
| $x_{21}$ | 0.92972416 | 0.9540565 | 0.918518719 |
| $x_{22}$ | 0.958275913 | 0.966075135 | 0.923498696 |
| $x_{23}$ | 0.964510428 | 0.969416594 | 0.927614416 |
| $x_{24}$ | 0.904902878 | 0.932738618 | 0.889112795 |
| $x_{25}$ | 0.921825665 | 0.889780858 | 0.889879811 |
| $x_{26}$ | 0.924210559 | 0.887709042 | 0.889738995 |
| $x_{27}$ | 0.909133179 | 0.878460128 | 0.886866221 |
| $x_{28}$ | 0.832386214 | 0.783337581 | 0.793820333 |
| $x_{29}$ | 0.974155895 | 0.96732494 | 0.94947448 |
| $x_{30}$ | 0.23099079 | 0.261543258 | 0.255984859 |
| $x_{31}$ | 0.691582575 | 0.735150936 | 0.72173451 |
| $x_{32}$ | 0.647323003 | 0.707040806 | 0.696402804 |

### 4.3. 聚类分析

通过徐静雅(2006)的变量聚类 + 选取典型相关法的思想，先对交通指标进行变量聚类。聚类是将研究对象进行分类，使得类与类之间距离最大，点与点之间的距离最小。本文采用系统聚类，其原理：将每一个点都看成单独的一类，通过离差平方和(Ward's method)计算类与类之间的距离，选择距离最近的合成新类，循环反复直到所有的点都在同一类为止，最后结果会给出谱系图。

系统聚类需要解决三个问题：确定点与点之间的距离；确定类与类之间的距离；聚类数目的确定。

1) 计算点与点之间的距离，主要方法有：绝对值距离、欧氏距离、切比雪夫距离、马氏距离、余弦距离。本文采用欧氏距离：

$$d_{ij}=\sqrt{\sum_{k=1}^{n}\left(x_{ik}-x_{jk}\right)^2},i=1,2,\cdots,m;j=i+1,i+2,\cdots,m \tag{16}$$

2) 计算类与类之间的距离，主要方法有：类平均法、可变法、重心法、最长距离法、最短距离法、离差平方和。这里采用离差平方和法：假设原样本为 $q$ 类，则第 $i$ 类的离差平方和定义为：

$$S_i=\sum_{j=1}^{N_i}\left(x_{ij}-\overline{x}_i\right)'\left(x_{ij}-\overline{x}_i\right) \tag{17}$$

其中 $\overline{x}_i$ 为第 $i$ 类变量均值，$N_i$ 为第 $i$ 类变量数量，假设将 $G_p$ 和 $G_q$ 合并成一个新类 $G_r$，则定 $G_p$ 和 $G_q$ 的平方距离为：

$$D_{pq}^2 = S_r - \left(S_p + S_q\right) \tag{18}$$

其中 $S_p$ 和 $S_q$ 分别为 $G_p$ 和 $G_q$ 类的离差平方和， $S_r$ 为新类 $G_r$ 的离差平方和。

3) 聚类数目的确定。聚类数目的确定一直都是研究难点，聚类数目确定得合适将事半功倍。判断聚类数目的方法：调整兰德系数法[3]、贝叶斯准则[4]、K 平均算法、K 中心聚类算法(K-mediods)、基于 Calinsky Criterion 准则、基于 AP 算法[5] (Affinity propagation Clustering Algorithm)等等。本文运用 AP 算法来确定聚类数目。AP 算法具体步骤：

1) 再开始 AP 算法前，将吸引度矩阵 $R$ 和归属度矩阵初始化为 0 矩阵；

2) 更新吸引度矩阵：

$$r_{t+1}\left(i,k\right) = \begin{cases} S\left(i,k\right) - \max_{j\neq k}\left\{a_t\left(i,j\right) + r_t\left(i,j\right)\right\}, i \neq k \\ S\left(i,k\right) - \max_{j\neq k}\left\{S\left(i,j\right)\right\}, i = k \end{cases} \tag{19}$$

3) 更新归属度矩阵：

$$a_{t+1}\left(i,k\right) = \begin{cases} \min\left\{0, r_{t+1}\left(k,k\right) + \sum_{j\neq i,k}\max\left\{r_{t+1}\left(j,k\right),0\right\}\right\}, i \neq k \\ \sum_{j\neq k}\max\left\{r_{t+1}\left(j,k\right),0\right\}, i = k \end{cases} \tag{20}$$

4) 根据衰减系数 $\lambda$ 对两个公式进行衰减

$$\begin{aligned} r_{t+1}\left(i,k\right) &= \lambda * r_t\left(i,k\right) + \left(1-\lambda\right) * r_{t+1}\left(i,k\right) \\ a_{t+1}\left(i,k\right) &= \lambda * a_t\left(i,k\right) + \left(1-\lambda\right) * a_{t+1}\left(i,k\right) \end{aligned} \tag{21}$$

一直重复步骤 2~步骤 4，直到矩阵稳定或者达到最大迭代次数，算法结束，最终取 $a+r$ 最大的 $k$ 作为聚类中心。

由于指标交通指标存在不同量纲，然而不同量纲无法进行准确的数据分析，所以要消除指标量纲带来的影响。在聚类分析之前对数据进行无量纲化处理，将数据 Z-score 标准化到[−1, 1] [6]，使得不同单位和量纲的指标处于同一数量级，避免了不同量纲对指标筛选造成的影响。再运用 AP 算法客观的确定聚类数目，为了保证聚类数目的准确性，亦采用了 k-平均聚类算法对聚类数目进行再次确定，从而和 AP 算法进行综合分析，分析可知交通指标最佳聚类数目为 7 类。对交通指标，使用 SPSS 软件系统聚类中的 Ward's method 对其进行 R 型聚类，点与点之间的距离采用欧氏距离，得到聚类结果见表 3 和聚类谱系图见图 1。

**Table 3.** Clustering results
**表 3.** 聚类结果

| 聚类 | 指标 |
|---|---|
| 第一类 | $x_1, x_{19}, x_{20}$ |
| 第二类 | $x_2, x_4, x_{10}, x_{12}, x_{17}, x_{29}$ |
| 第三类 | $x_3, x_{30}, x_{31}$ |
| 第四类 | $x_5, x_6, x_{15}, x_{16}, x_{32}$ |
| 第五类 | $x_7, x_8, x_9, x_{11}, x_{13}, x_{18}$ |
| 第六类 | $x_{14}, x_{25}, x_{26}, x_{27}, x_{28}$ |
| 第七类 | $x_{21}, x_{22}, x_{23}, x_{24}$ |

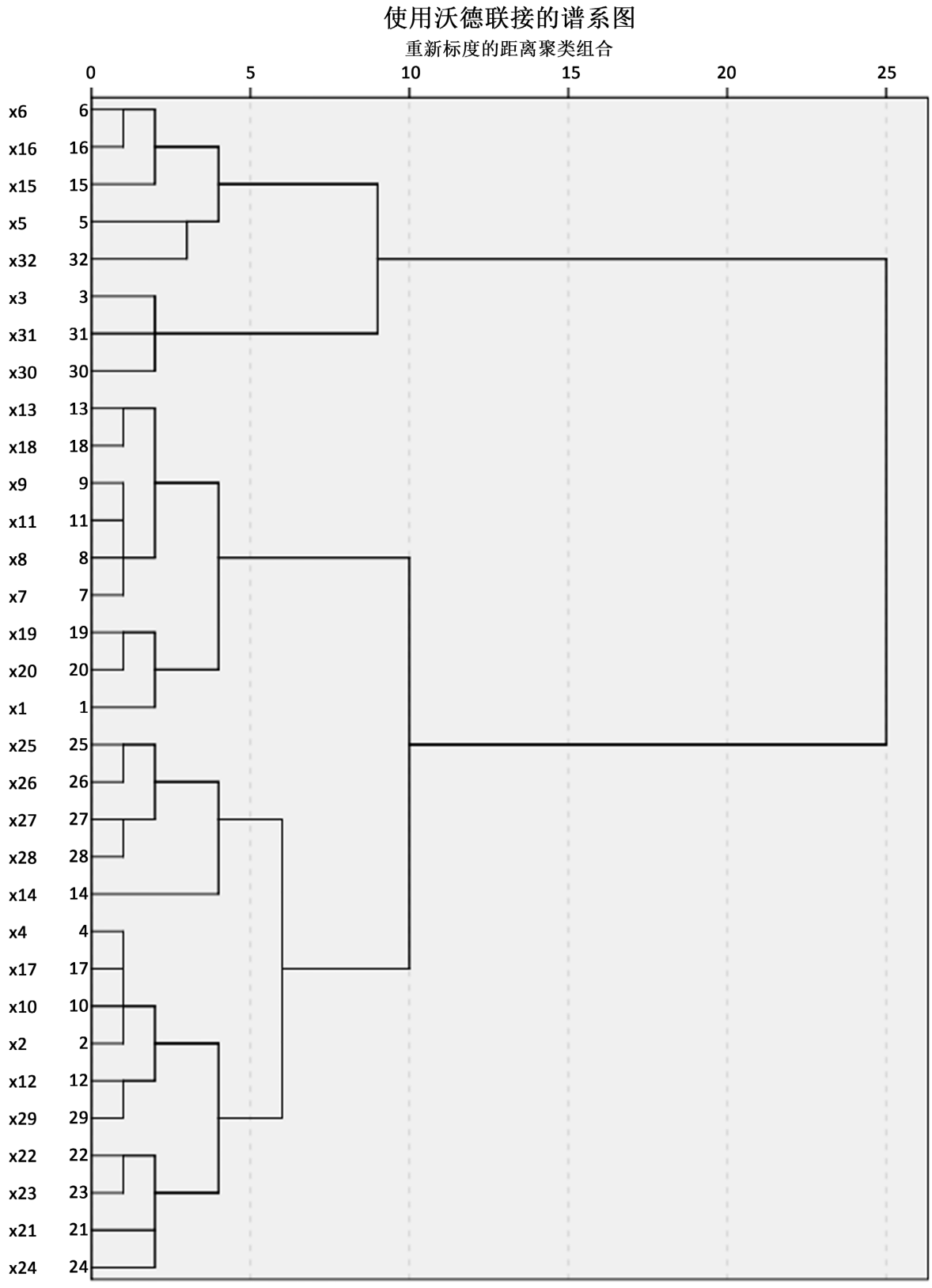


**Figure 1.** Clustering spectrum chart
**图 1.** 聚类谱系图

### 4.4. 交通指标选取

构建的交通指标体系并不都对第一产业、第二产业和第三产业有显著的助推作用，有些指标对农村产业没有明显作用，可能存在冗余或者抑制作用，需将其剔除。聚类分析已经将交通指标聚为 7 类，本文分别以第一产业、第二产业和第三产业为参考序列，运用加权均方误差系数法分别计算农村产业与交通指标的显著度，再在每一类中分别选取对第一产业、第二产业和第三产业显著度最大的指标。

**Table 4.** Significance of the three major industry and transport indicators
**表 4.** 三大产业与交通指标的显著度

| 交通指标 | 符号 | 第一产业 | 第二产业 | 第三产业 |
|---|---|---|---|---|
| | | 显著度 | 显著度 | 显著度 |
| 一级公路 | $x_1$ | 0.914471546 | 0.902840112 | 0.907093229 |
| 路网及农村公路建设投资 | $x_{19}$ | 0.917078665 | 0.909259484 | 0.895036347 |
| 汽车站场投资 | $x_{20}$ | 0.948275625 | 0.94157145 | 0.939886274 |
| 二级公路 | $x_2$ | 0.71409911 | 0.701826009 | 0.821465955 |
| 四级公路 | $x_4$ | 0.804805821 | 0.840873383 | 0.910830875 |
| 载货汽车 | $x_{10}$ | 0.737412168 | 0.769999074 | 0.864519047 |
| 公路货物运输量 | $x_{12}$ | 0.894812264 | 0.943764923 | 0.936864733 |
| 等级公路里程 | $x_{17}$ | 0.764454674 | 0.784307128 | 0.879947134 |
| 建制村通油路率 | $x_{29}$ | 0.868916272 | 0.856557724 | 0.889558323 |
| 三级公路 | $x_3$ | 0.741610082 | 0.72702952 | 0.801198132 |
| 乡镇通公路率 | $x_{30}$ | 0.742754892 | 0.722898605 | 0.798517352 |
| 建制村通公路率 | $x_{31}$ | 0.791745742 | 0.778923856 | 0.841483448 |
| 等外公路 | $x_5$ | 0.763233799 | 0.778735239 | 0.832287324 |
| 公路路网密度 | $x_6$ | 0.896712816 | 0.90338573 | 0.927033815 |
| 公路旅客周转量 | $x_{15}$ | 0.908724328 | 0.911690293 | 0.931952905 |
| 公路总里程 | $x_{16}$ | 0.896724723 | 0.903397848 | 0.927042485 |
| 乡镇通油路率 | $x_{32}$ | 0.842511477 | 0.854171839 | 0.893272524 |
| 高速公路车辆通行费收入 | $x_7$ | 0.859747979 | 0.875929133 | 0.823747296 |
| 民用汽车拥有量 | $x_8$ | 0.933401923 | 0.935989219 | 0.916177335 |
| 载客汽车 | $x_9$ | 0.874559804 | 0.875059894 | 0.81336821 |
| 私人汽车拥有量 | $x_{11}$ | 0.868688174 | 0.887025427 | 0.832798163 |
| 公路货物周转量 | $x_{13}$ | 0.929070267 | 0.961271537 | 0.927635572 |
| 高速公路里程 | $x_{18}$ | 0.92175621 | 0.940426233 | 0.883091207 |
| 公路旅客运输量 | $x_{14}$ | 0.704830392 | 0.677384371 | 0.782878316 |
| 国道 | $x_{25}$ | 0.904620322 | 0.85780787 | 0.904106811 |
| 省道 | $x_{26}$ | 0.880696918 | 0.833062988 | 0.890004069 |
| 县道 | $x_{27}$ | 0.830092591 | 0.788076631 | 0.863406794 |
| 乡道 | $x_{28}$ | 0.808834317 | 0.759649603 | 0.83848401 |

**Continued**

| | | | | |
|---|---|---|---|---|
| 重点公路投资 | $x_{21}$ | 0.879222495 | 0.910618873 | 0.927816813 |
| 公路建设投资 | $x_{22}$ | 0.897358171 | 0.915929317 | 0.930464777 |
| 交通固定资产投资 | $x_{23}$ | 0.896332059 | 0.912990591 | 0.927958932 |
| 交通运输、仓储和邮电通信业 | $x_{24}$ | 0.913211311 | 0.941519348 | 0.953723336 |

通过加权均方误差系数法得出交通指标与第一产业、第二产业和第三产业的显著度见表 4，分析可知第一类中一级公路($x_1$)、路网及农村公路建设投资($x_{19}$)和汽车站场投资($x_{20}$)三个指标，汽车站场投资($x_{20}$)对于第一产业、第二产业和第三产业的显著度都是最大的，并且显著度都在 0.93 以上。对于第二类中二级公路($x_2$)、四级公路($x_4$)、载货汽车($x_{10}$)、公路货物运输量($x_{12}$)、等级公路里程($x_{17}$)、建制村通油路率($x_{29}$)六个指标，公路货物运输量($x_{12}$)对于第一产业、第二产业和第三产业的显著度都是最大，并且显著度都在 0.89 以上。对于其余几类，在每类中分别是建制村通公路率($x_{31}$)、公路旅客周转量($x_{15}$)、国道($x_{25}$)、交通运输、仓储和邮电通信业($x_{24}$)对第一产业、第二产业和第三产业的显著度最大。不同的是，在第五类中，高速公路车辆通行费收入($x_7$)、民用汽车拥有量($x_8$)、载客汽车($x_9$)、私人汽车拥有量($x_{11}$)、公路货物周转量($x_{13}$)、高速公路里程($x_{18}$)六个指标，对于第一产业，民用汽车拥有量($x_8$)的显著度最大，而对于第二产业和第三产业，公路货物周转量($x_{13}$)的显著度最大。因此通过在每类中分别选取与第一产业、第二产业和第三产业显著度最大的交通指标，结果为：

第一产业选取的交通指标为：民用汽车拥有量($x_8$)、公路货物运输量($x_{12}$)、公路旅客周转量($x_{15}$)、汽车站场投资($x_{20}$)、交通运输、仓储和邮电通信业($x_{24}$)、国道($x_{25}$)和建制村通公路率($x_{31}$)。

第二产业选取的交通指标为：公路货物运输量($x_{12}$)、公路货物周转量($x_{13}$)、公路旅客周转量($x_{15}$)、汽车站场投资($x_{20}$)、交通运输、仓储和邮电通信业($x_{24}$)、国道($x_{25}$)和建制村通公路率($x_{31}$)。

第三产业选取的交通指标为：公路货物运输量($x_{12}$)、公路货物周转量($x_{13}$)、公路旅客周转量($x_{15}$)、汽车站场投资($x_{20}$)、交通运输、仓储和邮电通信业($x_{24}$)、国道($x_{25}$)和建制村通公路率($x_{31}$)。

汽车站场投资($x_{20}$)、国道($x_{25}$)的增加和建制村通公路率($x_{31}$)的提高为交通推动产业发展奠定了基础，使得通行无阻，促进了贵州旅游业的发展，旅游业带动一方经济的发展，环环相扣。贵州省是一个民族特色很浓厚的省份，然而也是比较贫困的省份。在交通不发达时期，人民想去民族特色浓厚的地方体验民族文化比较艰难。交通的改善，山里的路通了，让具有浓厚民族文化的鼓楼，村寨，吊脚楼成为了旅游盛地。民用汽车拥有量($x_8$)和公路旅客周转量($x_{15}$)的增加是人民出行的基础，针对旅游有两种说法：自驾游和穷游，自驾游的增多进而就间接代表民用汽车拥有量增多，穷游的即通过选择公用出行方式，进而使得公路旅客周转量增多。公路货物运输量($x_{12}$)和公路货物周转量($x_{13}$)主要为第二产业和第三产业提供保障，公路货物运输量和公路货物周转量的稳固增加，进而代表着第二产业和第三产业稳固发展。进而贵州逐步发展“交通 + 生态旅游”，旅游业同时带动第一产业、第二产业和第三产业稳固发展。

所以，通过交通对农村产业的影响结果分析，得到民用汽车拥有量($x_8$)、公路货物运输量($x_{12}$)、公路货物周转量($x_{13}$)、公路旅客周转量($x_{15}$)、汽车站场投资($x_{20}$)、交通运输、仓储和邮电通信业($x_{24}$)、国道($x_{25}$)和建制村通公路率($x_{31}$)对农村产业的影响较强，进而可知加权均方误差系数筛选出来了交通评价因子符合实际且有效。

## 5. 结语与建议对策

本文提出的加权均方误差系数法对评价因子进行筛选，并将此方法运用到贵州交通对农村产业的影响进行实证分析，结果筛选出对农村产业显著性较高的 7 个交通指标，通过上述的分析，表明提出的方

法筛选出来的交通评价因子是有效的。也将该方法与灰色关联法进行比较，发现提出的方法提高了数据结果的精度，进而说明了加权均方误差优。并且通过选出的交通评价因子，可以给贵州交通有关部门进行建议，对于有些还未过上小康生活的地方，可以通过改善交通来改变生活，并带领大家致富。道路的通行，带动大量人口流动，进而发展旅游业和有特色的农业，进而有效的助推农村产业，最终实现农业强、农村美、农民富。

## 参考文献

[1] 张艳芹. 非参数检验方法在指标选取中的应用[J]. 上海统计, 2001(5): 23-26.

[2] 徐雅静, 汪远征. 变量聚类——全局主成分分析在我国普通高等教育发展水平评价中的应用[J]. 数理统计与管理, 2006(5): 566-573.

[3] Hubert, L. and Arabie, P. (1985) Comparing Partitions. *Journal of Classification*, **2**, 193-218. https://doi.org/10.1007/BF01908075

[4] Schwartz, G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464. https://doi.org/10.1214/aos/1176344136

[5] Dueck, D. and Frey, B.J. (2007) Non-Metric Affinity Propagation for Unsupervised Image Categorization. 2007 *IEEE* 11*th International Conference on Computer Vision*, Rio de Janeiro, 14-21 October 2007. https://doi.org/10.1109/ICCV.2007.4408853

[6] 韩胜娟. SPSS 聚类分析中数据无量纲化方法比较[J]. 科技广场, 2008(3): 229-231.