基于主成分分析的行业标准数据分析

金波（温州市标准化研究院）

摘要：行业标准是我国标准体系的重要组成部分，论文讲解了主成分分析的原理与数学模型，然后利用 SPSS 软件对 2001—2017 年温州各区域企事业单位制定的行业标准进行主成分分析，通过分析结果对温州不同区域的行业标准制定进行评价。

关键词：主成分　分析　行业　标准　温州

Industry Standard Data Analysis Based on Principal Component Analysis

Jin Bo (Wenzhou Institute of Standardization)

Abstract: Industry standard is an important part of the standard system in China. This article roughly explained the principle and mathematical model of principal component analysis, and then using the SPSS software to analyze the industry standard in the various regions of Wenzhou in 2001 to 2017. It also evaluate the industry standards of different regions of Wenzhou through the analysis of the results.

Key words: principal component, analysis, industry, standard, Wenzhou

1　引言

行业标准是我国标准体系不可或缺的一部分，行业标准的建设对于行业健康发展有着举足轻重的作用。2018 年实施的新《中华人民共和国标准化法》（以下简称《标准化法》）虽然取消了强制性行业标准，但仍然保留推荐性行业标准的要求。新《标准化法》中保留“对没有推荐性国家标准、需要在全国某个行业范围内统一的技术要求，可以制定行业标准”这一条款，在国家标准缺失的情况下，行业标准起到补充作用，作为行业规范性文件引导行业良性发展。

1988 年发布的《标准化法》提出行业标准制定至今已有 30 年，但温州市各区域企事业单位参与制定行业标准的重视度和参与度各不相同，文中以 2001—2017 年温州各区域企事业单位参与制定的行业标准作为数据样本，利用主成分分析法进行区域评分评价。

2　主成分分析的原理与模型

2.1　主成分分析原理

主成分分析是通过正交变换，将多个具有一定相关性的原始变量转化为少数几个互不相关的综合变量的降维统计分析方法。降维得到的综合变量基本能够反映原始变量的绝大部分信息，通常表现为原始变量的线性组合。主成分分析最常见的做法就是生成的 n 个互不相关的综合变量（称为 n 个“成分”），按照方差 Var（F_n）越大包含信息越多的原则，从大到小进行排列，再根据特征值、累计方差贡献率等条件进行成分选择分析。

例如，第一个成分的方差 Var（F_1）最大，表示 F_1 包含原始变量的信息最多，如果第一主成分不足以代表原来 n 个变量的信息，再考虑选取第二个主成分 F_2，依次类推可以构造出第三、第四，……，第 n 个主成分。为了有效地反映原来 n 个原始变量的信息，已有的信息尽量不出现在 中，需要是 F_1 与 F_2 不相关，用数学形式表达就是要求 Cov（F_1，F_2）=0，以此类推要求各主成分之间互不相关。

2.2 主成分分析数学模型

设有 n 个样品，每个样品有 p 个指标：X_1，X_2，…，X_p，得到原始数据矩阵：

$$X=\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}=(X_1,X_1,\cdots,X_p)，其中 X_i=\begin{bmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{ni} \end{bmatrix}$$

设 $a_i=(a_{1i}，a_{2i}，\cdots，a_{pi})^T$（$i$=1，2，…，$p$）为 p 个常数向量，用原始数据 X 的 p 个向量 X_1，X_2，…，X_p，作如下线性组合：

$$\begin{cases} F_1=a_1^TX=a_{11}X_1+a_{21}X_2+\cdots+a_{p1}X_p \\ F_2=a_2^TX=a_{12}X_1+a_{22}X_2+\cdots+a_{p2}X_p \\ \vdots \\ F_p=a_p^TX=a_{1p}X_1+a_{2p}X_2+\cdots+a_{pp}X_p \end{cases}$$

简记为：

$F_i=a_i^TX=a_{1i}X_1+a_{2i}X_2+\cdots+a_{pi}X_p$（$i$=1，2，…，$p$）

上述公式需满足：

（1）主成分的系数平方和等于 1，

$a_{1i}^2+a_{2i}^2+\cdots+a_{pi}^2=1$，（$i$=1，2，…，$p$）

（2）主成分之间互不相关，即无重叠的信息，

$\mathrm{Cov}(F_i，F_j)=0$，（$i\neq j$，i，j=1，2，…，p）

（3）主成分的方差依次递减，重要性也依次递减，$\mathrm{Var}(F_1)\geqslant\mathrm{Var}(F_2)\geqslant\cdots\geqslant\mathrm{Var}(F_p)$

3 行业标准数据分析

3.1 行业标准指标确定

本文分析的数据以温州企事业单位参与制定的行业标准为对象，时间跨度从 2001 年至 2017 年，数据主要来源为温州市标准化创新服务平台，从约 17.7 万份行业标准数据中，通过条件筛选与信息比对的方式确定行业标准的起草单位、行政区域等内容，总共选出的行业标准数量为 737 份（见表 1）。

表 1 行业标准统计数据

区域	标准参与数（X_1）	主持制定（X_2）	主导制定（X_3）	有采用先进标准（X_4）	实施少于 5 年（X_5）
乐清市	218	12	59	6	121
瑞安市	175	32	41	9	128
鹿城区	118	40	43	9	51
龙湾区	161	34	38	13	68
永嘉县	145	20	43	7	81
苍南县	33	1	8	2	24
瓯海区	92	37	16	3	30
平阳县	56	8	22	0	39
泰顺县	1	0	0	0	0
市辖区	25	11	7	2	25
洞头区	2	0	2	0	2

（1）标准参与数（X_1）：在 737 份筛选出的行业标准中，部分行业标准的起草单位不止一家温州企事业单位参与，所以重复计数，最终确定标准参与总样本量为 1026 项。

（2）主持制定（X_2）、主导制定（X_3）：企事业单位制定行业标准的参与程度，本文以第 1 起草单位为主持制定，第 2、3 为主导制定，第 4 及以后排名为参与制定来区分。

（3）有采用先进标准（X_4）：采用 ISO、IEC、ITU 等国际先进标准，或国际先进的区域标准，对标准的技术水平有一定的影响。

（4）实施少于 5 年（X_5）：相关文献研究，标

准的标龄对经济的增长有抑制作用，同时，在通常情况下 5 年正好是一个标准的复审周期，以实施日期起的 5 年为分界点。

在指标的确定过程中，由于数据的共线性等因素，删除了参与起草制定（排名第 4 及以后）、无采用先进标准、实施大于等于 5 年等指标的统计数据。市辖区涵盖温州市本级、经济技术开发区等企事业起草的行业标准，“市辖区”非行政区概念。

3.2 行业标准主成分分析

表 2 相关系数矩阵

/		标准参与数	主持制定	主导制定	有采用先进标准	实施少于 5 年
相关系数	标准参与数	1.000	0.622	0.959	0.806	0.938
	主持制定	0.622	1.000	0.572	0.772	0.465
	主导制定	0.959	0.572	1.000	0.769	0.899
	有采用先进标准	0.806	0.772	0.769	1.000	0.692
	实施小于5年	0.938	0.465	0.899	0.692	1.000

表 3 KMO 和巴特利特检验

KMO 取样适切性量数		0.795
巴特利特的球形度检验	上次读取的卡方	51.276
	自由度	10
	显著性	0.000

表 4 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	4.025	80.497	80.497	4.025	80.497	80.497
2	0.688	13.769	94.266	—	—	—
3	0.171	3.423	97.689	—	—	—
4	0.091	1.811	99.500	—	—	—
5	0.025	0.500	100.000	—	—	—

表 5 成分得分系数矩阵

/	成分 1
标准参与数	0.242
主持制定	0.186
主导制定	0.235
有采用先进标准	0.223
实施小于5年	0.225

运用SPSS统计分析软件对行业标准统计数据(表1)进行主成分分析，从表2的相关系数矩阵来看，各指标间存在较高的相关性。

由于相关性较高，利用表3进行数据检查，查看数据是否符合主成分分析要求，从表3检验结果来看，KMO数值接近0.8，显著性小于0.05，表1的数据符合主成分分析法的要求，能够运用主成分分析。

主成分选择一般以特征值大于1，从表4中可知，累计方差贡献率只有80.497%，按累计方差贡献率大于85%为选择原则，需要考虑选取成分2作为第二主成分，但第二主成分的特征值远小于1，经过测试与方析，提取成分1作为第一主成分基本能够表现5个原始变量信息，故只选择F_1。

通过表5“成分得分系数阵”，可以将第一主成分F_1表示为原始变量的线性组合：

$$F_1=0.242X_1+0.186X_2+0.235X_3+0.223X_4+0.225X_5$$

由于本案例仅提取一个主成分，后期不再合成综合得分模型，用生成的第一主成分得分分析，成分得分图如图1所示。

3.3 行业标准分析评价

对表1的行业标准统计数据进行排名，得到表6排名表。

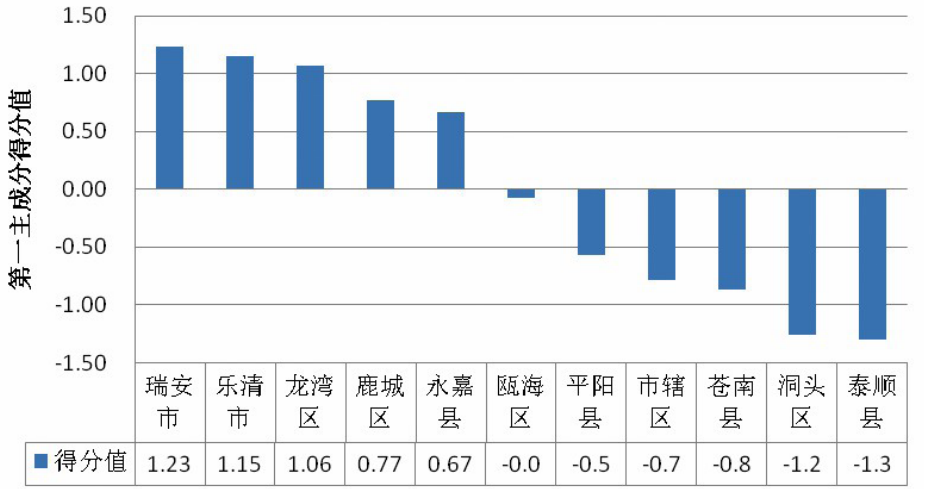

图1 成分得分图

表6 各区域排名表

区域	标准参与数（X_1）	主持制定（X_2）	主导制定（X_3）	有采用先进标准（X_4）	实施少于5年（X_5）
乐清市	1	6	1	5	2
瑞安市	2	4	4	2	1
鹿城区	5	1	2	2	5
龙湾区	3	3	5	1	4
永嘉县	4	5	3	4	3
苍南县	8	9	8	7	9
瓯海区	6	2	7	6	7
平阳县	7	8	6	9	6
泰顺县	11	10	11	10	11
市辖区	9	7	9	8	8
洞头区	10	10	10	10	10

排名表非常清晰地表示出各个区域在不同变量之间的优劣状态，在标准的参与数量上体现了企事业单位参与标准起草的积极性，进行比较后发现乐清市（218）、瑞安市（175）、龙湾区（161）、这些温州的工业强市（区）分列第1、2、3名；“一流企业做标准”是大家的共识，谁掌握标准的制定权，谁也就掌握了标准的话语权，主持制定正是从标准话语权的重要性出发进行比较，前3名分别是鹿城区（40）、瓯海区（37）、龙湾区（34）；采用ISO、IEC、ITU等国际上先进标准，在一定程度上能够提升标准的质量与水平，促进行业技术进步与发展，前3名分别是龙湾区（13）、鹿城区（9）、瑞安市（9）；标准的实施日期少于5年，表明标准相对较新，正处于促进经济、技术发展的阶段，前3名分别是瑞安市(128)、乐清市(121)、永嘉县(81)。

从上述排名分析可以看出，以不同的变量进行统计数据比较，温州各区域都有着自身的一些优势，区域彼此之间不分伯仲。单一变量统计排名不能综合区分区域的强弱，所以本文案例希望利用主成分分析，简化变量，利用1~2个新变量（F_1，F_2）解释原始变量（X_1，X_2，…，X_5），本文第一主成分已经涵盖了原始5个变量的信息，所以只提取第一主成分F_1，并计算了相关的成分得分，成分得分的排名可理解为区域排名综合得分，从图1成分得分图的排名看，瑞安市（1.23）、乐清市（1.15）、龙湾区（1.06）分列前3位。就以乐清来讲，虽然其在“标准参与数”“主导制定”位列第1，“实施日期少于5年”位列第2，但同样重要的“主持制定”（第6位）、“采用先进标准”（第5位）排名比较居中，并没有绝对的优势，经过主成分分析后，根据成分得分排名后，乐清市仅位列第2位。

4 结语

由于时间跨度从2001年到2017年，个别行业标准题录信息有所缺失，对数据的完整性有一定的影响，例如，在比对出的737份行业标准中，文成县在整个时间跨度中均未有行业标准出现，所以也就无法列入分析。在分析中的各指标，均由行业标准题录信息归纳出来，在一定程度上存在相关性，使得主成分分析的个别条件未能达到“最优”。虽然在数据上存在一定的不足，但从整体而言，本文从客观的数据出发，利用主成分降维分析的思路，寻求客观评价方法，对温州各区域的行业标准制定情况进行了综合直观的评价。

参考文献

[1] 张文霖．主成分分析在满意度权重确定中的应用[J]. 市场研究，2006（6）：18-22.

[2] 胡波，郭丽．实用统计分析方法与技术[M]. 北京：化学工业出版社，2012.

[3] 林海明，杜子芳．主成分分析综合评价应该注意的问题[J]. 统计研究，2013，30（8）：25-31.

[4] 韩小孩，张耀辉，孙福军，等．基于主成分分析的指标权重确定方法[J]. 四川兵工学报，2012，33（10）：124-126.