Statistical and Application 统计学与应用, 2015, 4(4), 312-318
Published Online December 2015 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2015.44035

Hans汉斯

The Model on Factors Selection and Prediction of Sand Liquefaction

Dongli Cui, Weiyan Mu

School of Science, Beijing University of Civil Engineering and Architecture, Beijing
Email: 18810972586m0@sina.cn, muweiyan@bucea.edu.cn

Received: Dec. 10th, 2015; accepted: Dec. 27th, 2015; published: Dec. 30th, 2015

Abstract

In order to reduce data dimension, simplify data operation, we adopted the method combining the factor analysis and discriminant analysis, and applied the cumulative variance contribution rate of k in front of more than 85% of the principal components instead of the original related factors of sand liquefaction to analyze, this method didn't reduce sample size, just made the raw data enrichment and comprehensive, did the discriminant analysis based on the factor score data, a set of discriminant results can be obtained. In addition, the extraction methods of principle component analysis were used to get the variable joint degrees, high variable joint degrees indicated the most information can be extracted by factor, then found the corresponding variable, did the discriminant analysis using these variables again, the two discriminant analysis results were compared with the original results and analyzed the misjudgment rate. Results show that the combination of the two methods has strong feasibility in filtering the main factors of sandlique faction and the prediction of sand liquefaction to some extent, and the effect is better.

Keywords

Sand Liquefaction, Factor Analysis, Discriminant Analysis

砂基液化的因素筛选及预测模型

崔栋利，牟唯嫣

北京建筑大学理学院，北京
Email: 18810972586m0@sina.cn, muweiyan@bucea.edu.cn

文章引用: 崔栋利, 牟唯嫣. 砂基液化的因素筛选及预测模型[J]. 统计学与应用, 2015, 4(4): 312-318.
http://dx.doi.org/10.12677/sa.2015.44035

收稿日期：2015年12月10日；录用日期：2015年12月27日；发布日期：2015年12月30日

摘　要

为降低数据维数，简化数据运算，我们采用因子分析和判别分析相结合的方法，运用方差累计贡献率在85%以上的前k个主成分代替原始砂基液化的有关因素，对砂基液化因素进行分析，这种方法并没有缩减样本量，只是对原始数据进行了浓缩和综合，通过对得到的因子得分数据进行判别分析，可得到一组判别结果。另外，利用因子分析的提取方法得到变量的共同度，变量共同度高的表示变量中的大部分信息均能够被因子所提取，选出变量共同度较高的对应的变量，利用这些变量再次进行判别分析，对两次判别分析得到的结果与原结果进行汇总对比，分析误判率。结果表明，这两种方法的结合在一定程度上用于筛选砂基液化的主要因素以及预测砂基液化可行性强，效果较好。

关键词

砂基液化，因子分析，判别分析

1. 引言

砂基液化是砂质地基在地下水压力突然增加时产生流动的现象。疏松的砂性土，特别是粉细砂，经过动载荷作用后将趋于密实，如地震、打桩、爆破及机械振动等。砂基液化能导致地裂缝、错位、滑坡、不均匀沉降等地基失稳现象。砂土、饱和、震动是砂土液化的基本条件。只要采取一定的方法和措施，砂基液化是可以预防和控制的[1]。砂基液化类型的确定是对砂质地基质量和稳定性的一种综合评价，所以利用相关因素以及样本数据对砂基液化进行预测有非常重要的作用，但是影响砂基液化的因素有很多，影响砂土液化的因素包括砂土的成分、砂的密度、砂层的有效覆盖压力及震动的强度和时间等，如果从这些因素来研究砂基是否液化，避免各个因素之间的信息重合，可以更好地分析各个因素对砂基液化的综合影响，从而可以根据这些因素的数据对砂基进行预测，这样可以降低数据维数，简化数据运算，节省计算时间，对于砂基液化问题的研究具有重要作用。

在现实研究过程中，往往需要对所反映事物、现象从多个角度进行观测。因此研究者往往设计出多个观测变量，从多个变量收集大量数据以便进行分析寻找规律。多变量大样本虽然会为我们的科学研究提供丰富的信息，但却增加了数据采集和处理的难度，更重要的是，许多变量之间存在一定的相关关系，导致了信息的重叠现象，从而增加了问题分析的复杂性[2]。所以，我们可以借鉴因子分析浓缩数据的优点，在信息损失较小的前提下，将多因素转换成较少的因素，根据提取的较少因素对未知类别的数据进行预测，将其运用到实际工程项目中，具有重要的作用。

2. 模型理论

2.1. 因子分析

因子分析(Factor Analysis)是多元统计分析的一个重要分支，最初是由英国心理学家 C. Spearman 提出的。利用“降维”的思想，在信息损失较小的前提下，将大量的彼此可能存在相关关系的变量，转换成较少的彼此不相关的综合指标。这样既可以减轻收集信息的工作量，又可使各综合指标代表的信息不重叠[3]。

常用的因子分析类型有 R 型因子分析和 Q 型因子分析，分别是针对变量和样本作因子分析。我们的

目的是将多因素转换成较少的因素，所以我们选用 R 型因子分析。设影响砂基液化的因素变量为 $Z_1,Z_2,\cdots,Z_m$，$F_1,F_2,\cdots,F_P$ 为公共因子，因子分析的基本模型表示为：

$$\begin{cases} Z_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1P}F_P + e_1 \\ Z_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2P}F_P + e_2 \\ \cdots \\ Z_m = a_{m1}F_1 + a_{m2}F_2 + \cdots + a_{mP}F_P + e_m \end{cases}$$

将因子分析的数学模型用矩阵形式表示如下：

$$Z = AF + e$$

其中，$Z = (Z_1,Z_2,\cdots,Z_m)^{\mathrm{T}}$，$F = (F_1,F_2,\cdots,F_P)^{\mathrm{T}}$，$e = (e_1,e_2,\cdots,e_m)^{\mathrm{T}}$，

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{bmatrix}$$

A 为因子载荷矩阵，估计 A 的方法有多种，如主成分法、映像因子法、加权最小二乘法、最大似然法等，最常用的是主成分法，a_{ij} 表示在各个因子变量不相关的情况下，第 i 个原始变量和第 j 个因子变量的相关系数，体现了 z_i 在第 j 个公共因子变量上的相对重要性，a_{ij} 的值越大，公共因子 F_j 和原始变量 X_i 的关系就越强。

2.2. 判别分析

判别分析是在分类数目已知的情况下，根据已经确定分类的对象的某些观测指标和所属类别来判断未知对象所属类别的一种统计学方法。判别分析法的思路如下：首先建立判别函数，然后通过已知所属分类的观测值确定判别函数中的待定系数，最后通过得到的判别函数对未知分类的样本进行预测。常用的判别分析法有距离判别法、费希尔判别法、贝叶斯判别法。判别分析在气候分类、农业区划、土地类型划分中有着广泛的应用[4]。

2.3. 因子－判别分析

运用方差累计贡献率在 85%以上的前 k 个主成分代替原始砂基液化的有关因素，对砂基液化因素进行分析，并没有缩减样本量，只是对原始数据进行了浓缩和综合，通过对得到的因子得分数据进行判别分析，可得到一组判别结果[5]。另外，利用主成分分析的提取方法得到变量的共同度，变量共同度高的表示变量中的大部分信息均能够被因子所提取，选出变量共同度较高的对应的变量，利用这些变量再次进行判别分析，对两次判别分析得到的结果与原结果进行汇总对比，分析误判率。

3. 模型实践

在有关地震预报的研究中，有时会遇到砂基液化的问题，影响砂基液化的因素有很多，如砂土的成分、砂的密度、砂层的有效覆盖压力及震动的强度和时间等，从中选择了 7 个有关因素 $X_1 \sim X_7$，分别从已液化和未液化的地层中得到容量分别为 9 与 16 的训练样本，第一组为液化，第二组为未液化，为避免在软件操作中遇到问题，将原始数据中的“液化”与“非液化”用“1”与“2”代替，样本数据如表 1 所示[6]：利用 SPSS 对样本数据进行基本统计，如均值，中位数等统计量，统计结果如表 2 所示。

Table 1. Data of sand liquefaction
表 1. 砂基液化数据

序号	组别	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	1	6.6	39	1.0	6.0	6	0.12	20
2	1	6.6	39	1.0	6.0	12	0.12	20
3	1	6.1	47	1.0	6.0	6	0.08	12
4	1	6.1	47	1.0	6.0	12	0.08	12
5	1	8.4	32	2.0	7.5	19	0.35	75
6	1	7.2	6	1.0	7.0	28	0.30	30
7	1	7.5	52	1.0	6.0	12	0.16	40
8	1	7.5	52	3.5	7.5	10	0.16	40
9	1	7.5	52	1.0	6.0	6	0.16	40
10	2	8.4	32	1.0	5.0	4	0.35	75
11	2	8.4	32	2.0	9.0	10	0.35	75
12	2	8.4	32	2.5	4.0	10	0.35	75
13	2	6.3	11	4.5	7.5	3	0.20	15
14	2	7.0	8	4.5	4.5	9	0.25	30
15	2	7.0	8	6.0	7.5	4	0.25	30
16	2	7.0	8	1.5	6.0	1	0.25	30
17	2	7.2	6	3.5	4.0	12	0.30	30
18	2	7.2	6	1.0	3.0	3	0.30	30
19	2	7.2	6	1.0	6.0	5	0.30	30
20	2	5.5	6	2.5	3.0	7	0.18	18
21	2	7.5	52	1.0	5.0	5	0.16	40
22	2	8.3	97	0.0	6.0	5	0.15	180
23	2	8.3	97	2.5	6.0	5	0.15	180
24	2	8.3	89	0.0	6.0	10	0.16	180
25	2	8.3	56	1.5	6.0	13	0.25	180

Table 2. Statistics description on factors of sand liquefaction
表 2. 砂基液化因素基本统计

		1	2	3	4	5	6	7
N	有效	25	25	25	25	25	25	25
	缺失	0	0	0	0	0	0	0
均值		7.3520	36.4800	1.9000	5.8600	8.6800	0.2192	59.4800
中位数		7.2000	32.0000	1.0000	6.0000	7.0000	0.2000	30.0000
众数		7.20[a]	6.00	1.00	6.00	5.00[a]	0.16	30.00
方差		0.723	811.260	2.208	2.032	33.143	0.008	3249.593
偏度		−0.325	0.755	1.257	−0.208	1.723	0.137	1.543
峰度		−0.683	−0.036	1.203	0.347	4.329	−1.272	1.004
最小值		5.50	6.00	0.00	3.00	1.00	0.08	12.00
最大值		8.40	97.00	6.00	9.00	28.00	0.35	180.00

[a]存在多种方式。已显示最小值。

3.1. Bartlett 球形度检验

Bartlett 球形度检验的原假设为相关系数矩阵为单位阵，如果 Sig 值小于 0.05，则表示变量之间存在相关关系，由此可否定相关矩阵为单位阵的原假设[7]，即此可认为各变量之间存在显著的相关性，Bartlett 形度检验结果表 3 表明，Bartlett 值 = 115.053，$P = 0.000$，因此适合做因子分析。

3.2. 砂基液化因素的筛选及预测

把7项砂基液化因素作为变量，利用SPSS软件进行因子分析，选择大于等于的1的特征值，并对他们的方差贡献率和累计贡献率进行汇总，汇总结果见表4，由于 $\lambda_4 = 0.989$，非常接近于1，且由大到小排列

的前三个特征值的累计贡献率未达到85%，但加上 λ_4 之后的累计贡献率达到90.799%，所以保留 λ_4。因子分析得到的因子得分数据见表5，把得到的因子得分数据看做四个变量，利用四个变量进行判别分析，对样本进行再次分类。另外，利用主成分分析的提取方法得到变量的共同度，结果见表6，变量共同度高的表

Table 3. Bartlett sphericity test
表 3. Bartlett 球形度检验

Bartlett的球形度检验	上次读取的卡方	115.053
	自由度	21
	显著性	0.000

Table 4. The first four characteristic value in the sequence and variance contributive rate
表 4. 前 4 个特征值及方差贡献率

顺序编号	λ_1	λ_2	λ_3	λ_4
特征值	2.549	1.691	1.125	0.989
方差贡献率（%）	36.419	24.151	16.077	14.132
累计贡献率（%）	36.419	60.570	76.648	90.799

Table 5. The data of factor score
表 5. 因子得分数据

序号	FACT__1	FACT__2	FACT__3	FACT__4
1	−0.62765	−1.14399	−0.0535	−0.13097
2	−0.7387	−1.11677	0.80016	−0.04464
3	−0.88688	−1.67808	−0.06836	−0.06744
4	−0.99793	−1.65086	0.7853	0.01889
5	0.53343	1.22544	1.598	0.94486
6	−0.78408	0.68299	3.30067	0.43588
7	0.00388	−0.68408	0.73773	−0.04143
8	−0.0885	−0.41892	−0.22528	1.59054
9	0.11493	−0.71131	−0.11594	−0.12776
10	0.8474	1.20579	−0.40744	−1.00425
11	0.71216	1.05979	0.43013	1.65931
12	0.64331	1.55566	−0.10351	−0.99045
13	−1.11598	0.04109	−1.34995	1.62514
14	−0.78629	0.93307	−0.72979	−0.03506
15	−0.75436	0.90018	−1.69057	2.07096
16	−0.4562	0.29309	−0.80609	−0.28642
17	−0.65383	1.25908	0.00791	−0.66539
18	−0.35381	0.90241	−0.55659	−2.17433
19	−0.36651	0.66185	−0.04678	−0.45768
20	−1.54028	−0.22727	−0.47942	−1.50987
21	0.12533	−0.63263	−0.3333	−0.70478
22	1.99857	−1.11639	−0.36208	−0.32985
23	1.85701	−0.71734	−1.15316	0.48696
24	1.83789	−0.93395	0.37219	−0.2933
25	1.4771	0.31113	0.44967	0.03109

Table 6. Variable degree of common
表 6. 变量共同度

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
初始值	1.000	1.000	1.000	1.000	1.000	1.000	1.000
提取	0.942	0.949	0.39	0.751	0.508	0.965	0.861

提取方法：主成份分析

Table 7. Grouping contrast of sample prediction
表 7. 样本预测分组对比

序号	原组别	预测 1	预测 2
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	2
6	1	1	2
7	1	1	1
8	1	1	1
9	1	1	1
10	2	2	2
11	2	2	2
12	2	2	2
13	2	2	2
14	2	2	2
15	2	2	2
16	2	2	2
17	2	2	2
18	2	2	2
19	2	2	2
20	2	2	2
21	2	2	2
22	2	2	2
23	2	2	2
24	2	2	2
25	2	2	2

示变量中的大部分信息均能够被因子所提取，选出变量共同度较高的对应的变量，利用这些变量再次进行判别分析，对样本进行第二次分类，对两次判别分析得到的结果与原结果进行汇总对比，分析误判率。从表6中我们可以看出X_1，X_2，X_6，X_7变量共同度较高，所以对这四个变量进行判别分析，最后对原来分组和两次判别分析的分组结果进行汇总和对比。结果见表7，其中，预测1和预测2是分别对因子得分数据和X_1，X_2，X_6，X_7进行判别分析得到的。

3.3. 结论

由表 7 可以得出，对因子得分数据进行判别分析得到的预测与原来的分组完全一致，一致性为 100%，

对 X_1，X_2，X_6，X_7 进行判别分析得到的预测与原来的分组几乎一致，只有第五组，第六组发生误判，误判率为 8%，相对较低。在实际生活中，我们经常会遇到高维数据，运用本文的模型对相关指标进行筛选和对数据的分组进行预测，可以达到降低数据维数，简化数据运算，节省计算时间的重要作用。结果表明，这两种方法的结合在一定程度上用于筛选砂基液化的主要因素以及预测砂基液化可行性强，效果较好。

参考文献 (References)

[1] 李学文. 中国袖珍百科全书[M]. 北京: 长城出版社, 2001: 5301-5309.

[2] 陈胜可. 统计分析从入门到精通[M]. 北京: 清华大学出版社, 2013: 349-360.

[3] 王鹏泽, 刘鹏飞, 等. 因子、聚类及判别分析在烟叶风格特色评价中的应用[J]. 中国烟草科学, 2015, 36(2): 20-25.

[4] 邵良杉, 徐波. 基于因子分析与 Fisher 判别分析法的隧洞围岩分类研究[J]. 公路交通科技, 2015, 32(7): 98-100.

[5] 王玉杰, 王千. 主要土壤肥力因素指标的筛选模型[J]. 生物数学学报, 2000, 15(2): 163-168.

[6] 梅长林, 范金城. 数据分析方法[M]. 北京: 高等出版社, 2006: 142-164.

[7] 何晓群. 多元统计分析[M]. 北京: 中国人民大学出版社, 2012: 143-154.