APPLICATION OF PRINCIPAL COMPONENT ANALYSIS FOR OPTIMIZATION OF A SYSTEM OPERATION ASSESSMENT CRITERIA SET

Łukasz Muślewski

University of Technology and Life Sciences
Machine Maintenance Department
Prof. S. Kaliskiego Street 7, 85-789 Bydgoszcz, Poland
tel.: +48 52 3408723, fax: +48 52 3286693
e-mail: lukasz.muslewski@utp.edu.pl

Leszek Knopik

University of Technology and Life Sciences
Faculty of Management
Fordońska Street 430, 85-790 Bydgoszcz, Poland
tel.: +48 52 3408228
e-mail: knopikl@utp.edu.pl

Abstract

The analysis of results of experimental tests and a literature survey on the issue reveal that the subject matter connected with determination of the optimal number of a given transportation system operation assessment criteria has a direct influence on the result of the considered assessment. Whole study has been made on the basis of the analysis of data obtained from tests performed in a real municipal transportation system, providing transport tasks in a 400 thousand urban agglomeration. On the basis of the tests, there was determined a set of sixteen criteria for the system operation assessment, depending on preferences of drivers, passengers, workers of the system providing vehicles with serviceability and the last group of results is a final assessment. The research involved tests of significance for correlation coefficients between the assessment criteria. The method of main factors involving analysis of a linear transformation of existing vector of variables X into Y has been used to study the task dimensioning. Proper values, providing basis for determination of a new space correspond to coordinates of the newly created vector Y. On this basis, the dimension of the criteria and their significance space has been concluded. The determined criteria have been accepted for the development of a resultant model for assessment of a transportation system operation quality.

Keywords: *quality, criteria, transport system, analysis of main components, optimization*

1. Introduction

In this paper, problems connected with assessment of transportation systems operation quality have been discussed. It has been defined that: 'Quality of the system – is a set of its features expressed by means of their numerical values, in time t, determining the fulfilment degree of the requirements it has to meet [7].

Thus, in order to build an assessment model of a given system it is necessary to establish the quantity and significance of the set of assessment criteria, which provide basis for determination of features to describe the system in terms of its operation quality.

The assessment process involves using each criterion from set X, for describing them features and on this basis, defining whether and to what extent a given criterion has been fulfilled, in given time t (while making an assessment) [3, 4].

Having in mind adequacy of the developed resultant model, it should be emphasized that identification and specification of the assessment criteria, plays significance an important role in evaluation of the system operation quality. Therefore, in this research there has been used and described a principal component analysis as a tool supporting the process of selection of the most important criteria for operation assessment quality of a given research object.

2. Research object

The research object is a real system of municipal bus transportation belonging to the group of social-technical systems of the type Human – Technical Object – Environment <H – TO – E>, whose main function is safe transporting people within a given quantitative and territorial range, by using transport means operated in the system.

The studied transportation system performs transport tasks on the territory of a town and its suburbs. The transport tasks must be performed punctually, reliably and safely. Therefore, providing the system with proper level of operation quality is of key importance.

The considered system has been identified and its two basic subsystems have been distinguished: logistic including: subsystems of decision and information and the ones which guarantee operation continuity (providing vehicles with serviceability, diagnostic, (fuel supply); and the executive one consisting of elementary subsystems of the type <H – TO> (driver-bus) as well as the environment as a supporting system largely affecting functioning of the whole system and its subsystems.

3. Research Description

On the basis of the analysis of experimental and survey tests results carried out in a real transportation system, a set of criteria was distinguished. Next, they were evaluated in terms of their significance for the system operation quality. In the analyzed set, there were distinguished the following criteria: safety, efficiency, availability, ergonomics, environment friendliness, usability, weather conditions, accessibility, esthetics, information, punctuality, time of the service performance, external factors, damageability, reliability and cost-effectiveness.

Respondents, diversified in terms of sex, age, education were the statistical population who expressed their opinion on the studied subject. They were divided into three groups, according to their preferences – qualitative requirements from the studied transportation system. The first were drivers responsible for performance of transport tasks (group 1). The second group were users of the transport services (group 2), whereas, the third group consisted of workers of the logistics subsystem employed in the considered company (group 3).

Thus, the respondent is a statistical unit – basic unit, and the number of the group of respondents is $N = 150$ (3×50). Significance of the evaluated criteria (16-component vector) in points is the criterion of variability. The scale of grades ranged in {0,1,….10).

4. Principal Component Analysis

In this study, a description of methodological assumptions has been made, and the principal component analysis has been applied for an analysis of sets of grades, determined on the basis of research on a real transportation system.

4.1. Introduction of the Method

A survey usually contains a big number of variables. However, the questions in a questionnaire are related to each other and the survey output is unnecessarily extended... On one hand, we try to describe the obtained data in the most complete way. On the other hand, we are limited by the survey costs. While analyzing the survey, different problems can be

encountered. One of them is the question how to make an optimal reduction of the set of variables without significant loss of information. Popular statistical multidimensional methods do not provide ideal statistical procedures for the best possible choice of a subset of variables. One of such methods is the principal component analysis. However, application of this analysis requires fulfilment of certain assumptions.

This method is carried out on measurable variables (quantitative), though according to literature it is possible to use it for variables of order type. The examined variables should be in a linear dependence with each other and the correlation between features should be measured by Pearson coefficient. If the analyzed variables are not related to each other, then application of the principal component method is not advisable.

On the basis of an analysis of literature [5, 6] it is known that when all coefficients of correlation are smaller than 0.3, application of the principal component method is not effective. The higher correlation coefficients, the more justified this method application is.

At the beginning of the statistical analysis, it is necessary to use test of Bartlett [1]. Bartlett test answers the question whether all correlation coefficients are equal to zero. Application of the method of principal component analysis requires a test with an adequate quantity. Literature [2, 6] suggests that if the correlations are strong, it is enough for the statistical test to have the quantity equal to 50. An assumption about the distribution normality is not necessary for a description of relations between the variables. However, when statistical tests are used to define significance of the components, the assumption about multi-dimensional normality of the studied features distribution is necessary.

4.2. Application of Principal Component Analysis

The analyzed data sets are numerical matrixes with dimensions nxp, where n stands for the number of surveys; p denotes the number of criteria. For a given matrix X(nxp), a matrix of correlation coefficients is determined. Matrix of correlation coefficients is subjected to Bartlett test, which decides about advisability of application of principal component analysis. If R denotes matrix of correlation coefficients, then Bartlett test involves verification of a statistical hypothesis form:

$$H_0: R = I, \tag{1}$$

where I is a unit matrix of dimension p×p.

Hypothesis Ho means that all coefficients of correlation contained in matrix R are equal to zero. Test statistics for this hypothesis has the form:

$$U = -(n-1-\frac{2p+5}{6})\sum_{i=1}^{p}\ln\lambda_i\,, \tag{2}$$

where:

p – number of variables,
n – number of tests
λ_i – i-th value of matrix R.

It is assumed that proper values are arranged in non-ascending order, which means:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p. \tag{3}$$

Statistics U has, with the assumption that hypothesis Ho is true, chi-square distribution with p (p–1) freedom degrees. Results of Bartlett test application for groups 1, 2, 3 are presented in Tab. 1.

The analysis of results of Ho hypothesis testing: R = I shows that for a very low p-value, it is necessary to reject the null hypothesis, for all the analyzed groups of data.

This means that it is justified to use the principal component method.

Tab.1. Bartlett test results

Value of statistics U	p-value
293.29	0.000001
192.39	0.000032
226.56	0.000001
256.69	0.000001

The principal component analysis enables determination of linear transformation of the form:

$$Z = A\,X, \tag{4}$$

where:
A is a matrix of pxp dimension linear transformation,
X is a column matrix of $X^T = [X_1, X_2, \ldots X_p]^T$.
Z is a column matrix containing dependent variables Z1, Z2, …, Zp called components.

The principal component analysis determines the first row vector of matrix A in such a way that component Z1 has the maximal variance with limiting the form:

$$\sum_{i=1}^{p} a_{1i}^2 = 1. \tag{5}$$

Next, the second component is determined so that variance of variable Z2 will be maximal for proper limits.

One of the main reasons for using the principal component analysis is verification of hypothesis of the form:

$$H_0 : \lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_p. \tag{6}$$

In relations to the alternative hypothesis:
H_1 : not all $\lambda_{k+1}, \lambda_{k+2}, \ldots, \lambda_p$ .are equal.

Testing statistics for Ho hypothesis has the form:

$$\chi^2 = -(n-k)\left[\sum_{j=1}^{p} \ln \lambda_j - q \ln \frac{1}{q} \sum_{i=k+1}^{p} \lambda_i\right], \tag{7}$$

where:
χ^2 has distribution
χ^2 z df = q(q+1)/2 – 1, q = n – k freedom degrees, for the assumption that the hypothesis is true.

Statistical hypothesis described by dependence (6) was verified successively for k = 0,1,2…p–2. Results of the verification are presented in Tab. 2 which contains proper values λ_i, i = 1, 2, …, p for „set 1". The analysis of data from Tab. 2 shows that the 10 highest proper values vary considerably.

Tab. 2. Results of testing for "set 1"

No.	Proper values	p – value	Test result
1	3.65	0.000	X
2	2.17	0.000	X
3	1.89	0.000	X
4	1.63	0.000	X
5	1.22	0.000	X
6	0.98	0.000	X
7	0.83	0.001	X
8	0.74	0.001	X

No.	Proper values	p – value	Test result
9	0.72	0.002	X
10	0.54	0.027	X
11	0.41	0.090	
12	0.35	0.118	
13	0.28	0.183	
14	0.23	0.238	
15	0.15	0.556	
16	0.11		

Table 3, contains proper values λ_i, i = 1, 2, …, p for "set 2". On the basis of the analysis of data from Tab. 3, it can be concluded that only the 4 highest proper values differ significantly. This means that "set 2" can be a set with too many features.

Tab. 3. Results of tests for "set 2"

No.	Eigenevalues	p – value	Test result
1	2.65	0.000	X
2	2.30	0.000	X
3	1.80	0.011	X
4	1.69	0.046	X
5	1.26	0.283	
6	1.11	0.461	
7	0.93	0.678	
8	0.85	0.808	

No.	Eigenevalues	p – value	Test result
9	0.69	0.955	
10	0.51	0.990	
11	0.45	0.974	
12	0.44	0.907	
13	0.42	0.850	
14	0.29	0.980	
15	0.28	0.789	
16	0.23		

Table 4 contains proper values λ_i, i = 1, 2, …, p for "set 3". The analysis of data from Tab. 4 shows that the 9 highest proper values differ significantly from each other.

Tab. 4. Results of tests for "set 3"

No.	Eigenevalues	p – value	Test result
1	3.00	0.000	X
2	2.48	0.000	X
3	1.93	0.001	X
4	1.33	0.019	X
5	1.17	0.031	X
6	0.94	0.044	X
7	0.88	0.028	X
8	0.87	0.022	X

No.	Eigenevalues	p – value	Test result
9	0.74	0.040	X
10	0.63	0.075	
11	0.57	0.123	
12	0.46	0.382	
13	0.31	0.895	
14	0.23	0.980	
15	0.20	0.934	
16	0.18		

Table 5 contains proper values λ_i, i = 1, 2, …, p for the analyzed sets of received results combined. The analysis of data contained in Tab. 5 shows that all the proper values differ significantly statistically, for significance level $p < 0.033$.

Tab. 5. Results of testing for a summary set of data (total)

No.	Eigenevalues	p – value	Test result
1	4.161	0.000	X
2	2.783	0.000	X
3	1.563	0.000	X
4	1.181	0.000	X
5	0.957	0.000	X
6	0.892	0.000	X
7	0.778	0.000	X
8	0.651	0.000	X

No.	Eigenevalues	p – value	Test result
9	0.603	0.001	X
10	0.49	0.015	X
11	0.415	0.032	X
12	0.4	0.020	X
13	0.359	0.025	X
14	0.343	0.033	X
15	0.212	0.039	X
16	0.111		

Proper values contained in Tab. 2-5 are in non-ascending order. A chart of successive proper values, in dependence on the proper value number, in an ordered sequence, described by dependence (3) is often used in graphical presentations of the principal component analysis. Such a chart is called an avalanche. In order to place all the sequences in one chart, they need to be normalized by dividing each proper value by the prime (maximal).

Charts of avalanches for the four analyzed sets are presented in Fig. 1.

The analysis of the above chart confirms that the considered charts get stabilized along with the growth of the proper value number. However, the stabilization degree is different has been shown before in statistical tests.

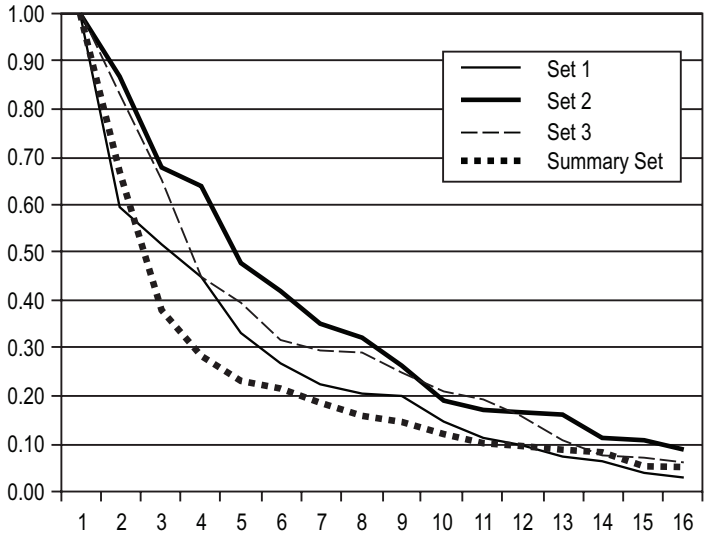

Fig. 1 Charts of avalanches for the analyzed data sets

5. Conclusions

Analyzing the results of surveys carried out with the use of principal component method, it can be said that there are definite reasons for limiting dimensionality, that is, reducing the number of accepted criteria, only for 'set 2' which contains only 4 statistically different proper values. Since this set is made up of the studied transportation system users who are for the most part the ones to set qualitative requirements – expectations connected with its functioning, it needs to be accounted for in final decisions on reduction of the considered criterion vector dimensionality.

On the basis of the analysis of results of data obtained for groups 1 and 3, it can be said that, respectively, 10 an 9 of proper values and the correlation of the other ones vary considerably from each other which makes it possible to conclude that it does not provide basis for reduction of any of them, in the analyzed set of criteria.

Whereas, from the analysis of the assessment results obtained from the summary set (N=150) it results that all proper values are statistically different. On this basis, it can be said that the considered set of criteria should not be reduced to their subsystems.

Distinct differentiation in the speed of the avalanche curves stabilization, for the analyzed groups, confirms advisability of further study on the differences between the groups and their statistical dimensionality.

References

[1] Bartlett, M. S., *A note on the multiplying factors for various chi square approximations*, Journal of the Royal Statistical Society, 16, series B, 296 – 298, 1954.
[2] Comrey, A. L., *A First Course in Factor Analysis*, Academic Press, New York 1973.
[3] Muślewkski, Ł., *Control Method for Transport System Operational Quality,* Journal of KONES Powertrain and Transport, Vol. 16, No. 3, Zakopane 2009.
[4] Muślewski, Ł., *Evaluation Method of Transport Systems Operation Quality,* Polish Journal of Environmental Studies, Vol. 18, No. 2A, Hard, Olsztyn 2009.
[5] Tabachnik, B. G., Fidell, L., *Computer – Assisted Research Design and Analysis*, Allyn & Bacon, Boston 2001.
[6] Tabachnik, B. G., Fidell, L., *Using Multivariate Statistics,* Haper & Row, New York 1996.
[7] Woropay, M, Muślewski, Ł., *Jakość w ujęciu systemowym*, ITeE, Radom 2005.