Comprehensive Service Level Analysis of Online Taxi Drivers Based on Fuzzy Clustering Combined with Principal Component Analysis

Hong Chen[1,a,*] and Chan Li[1,b]

[1] School of Information Science & Technology, Xiamen University Tan Kah Kee College, Zhangzhou, China，363105

[a]chenhong@xujc.com, [b]lichan@xujc.com

Keywords: Driver service level, Principal component analysis, System clustering, FCM-PCA.

Abstract. Online taxi tourism is one of the important ways of daily tourism. The operator carries out a single evaluation method for the driver's service quality, lacking a comprehensive study of service quality from multiple dimensions of order activity satisfaction, resulting in a high degree of hidden danger to passenger safety and rights. In this paper, an improved principal component analysis (PCA) method, namely Fuzzy C-Mean Clustering (FCM-PCA) based on PCA, is proposed. Experiments show that in the research of target object evaluation, the principal component values and principal component scores of target samples can be used as new indicators for clustering, so as to improve the efficiency of high-dimensional data clustering on the basis of reducing information loss. This study provides a way of thinking for the selection of important service components and a research method for the comprehensive analysis of different drivers' service levels.

1. Introduction

The market competition of the taxi software is evolving into the ecological circle. However, the hidden safety problem of online car booking has also surfaced[1]. This paper closely analyzes the service problems of Didi drivers around the background of passenger travel safety, and provides a more effective comprehensive evaluation and analysis method for drivers compared with the previous single star rating of operators[2].

Comprehensive evaluation methods include statistical analysis, analytic hierarchy process, grey relational analysis and data envelopment analysis (DEA) and so on[3-6]. Bian (2009) adopted the improved method of combining the Fuzzy Comprehensive Method with the Analytic Hierarchy Process (AHP). Practice has proved that this method can be used as technical support for strengthening road safety management.

Recently, Chica-Olmo, Jorge (2017) applies two-step combination of Non-linear Principal Component Analysis (NLPCA) and logical multilevel model (LMLM) to the method of Granada Metropolitan Transport Federation, which was established in 2013[7]. It was not long before Jenelius and Koutsopoulos (2018) proposed a multivariate probabilistic principal component analysis (PPCA) model for link travel time[8]. Xu, K (2018) examines the factors that affect taxi driver response behavior to ride-hailing requests. The results show that empirical research from the driver's point of view is of great significance to the service providers. As Stefanescu (2014) analyzed, travel planners play an important role in public transport operators and passengers using public transport, so does passenger safety. some basic information including garrival time, route, price, distance, interest point, location, connection with other means of transport, transfer times and other information are very important to passengers [9].

The above research shows that the conceptualization and measurement of transport service quality-a fundamental determinant of demand-poses challenges for conducting economic analyses and designing mobility policies. In this paper, an improved principal component analysis (PCA) method, FCM-PCA, is proposed, which is based on PCA and combined with fuzzy C-means (FCM) clustering. The principal component contains most information of original variables and contains more condensed information. Principal Component Analysis (PCA) combines the original correlative features into a few new linear independent features, which maximizes the variance of data in each

dimension of the projected feature subspace[10]. This method can get the final sample evaluation, but our goal is not only to evaluate, but also to classify the samples and get the similarity of the target. Fuzzy C-means (FCM) clustering algorithm is a method of fuzzy grouping of data samples. The membership degree of each object to the grouping center is obtained by optimizing the objective function, which allows the samples to belong to different groups with a certain degree of membership. By clustering, similar properties can be grouped into one group[11-12]. In the study of evaluating target objects, the principal component values and principal component scores of target samples can be clustered as new indicators, which can improve the efficiency of high-dimensional data clustering on the basis of reducing information loss[10]. This method has been validated on the Didi taxi software online platform which provides business services for Chinese enterprises in Guiyang area.

The rest of the paper is organized as follows: The second part introduces the improved principal component analysis algorithm FCM-PCA model. The third part introduces the specific steps of algorithm construction; the fourth part takes droplets as an example to make an empirical analysis of the method. The fifth section draws relevant conclusions and puts forward corresponding suggestions for the empirical results of the case.

2. Optimized Principal Component Analysis: FCM-PCA

To analyze the comprehensive evaluation of the objective function, we first introduce the following assumptions.

Assumption 1. (a)the sample data $X \in R^{p\times N}$ has p dimensions and n samples. (b) the data are centralized, i. e. $\sum_{i=1}^{n} x_i = 0$, and the projected new coordinate system is $\{\omega_1, \omega_2, \dots, \omega_p\}$, where ω_i is a standard orthogonal basis, satisfying$\|\omega_i\|_2 = 1, {\omega_i}^{\mathrm{T}}\omega_j = 0(i, j = 1, \dots, p, i \neq j)$.

According to the properties of orthogonal matrix. Abandoning part of the dimension, from p dimension to k dimension, we find a matrix $W_k = \{\omega_1, \omega_2, \dots, \omega_p\}$ is also a standard orthogonal basis[10], where $k < p, \omega_m(m = 1, \dots, k)$.

Assumption 2. The projection of sample $x^{(i)}$ in k-dimensional feature space $z^{(i)} = (z_1^{(i)}, z_2^{(i)}, \dots, z_k^{(i)})^{(T)}$, where $z_m^{(i)} = \omega_m^T x^{(i)}(i = 1, \dots, N)$ is the m -dimensional coordinates in the k-dimensional coordinate system.

Assumption 3. $\mathrm{X} = \{X_1, X_2, \dots, X_p\}$,then the principal component is

$$C_i = \left(u_{1i}X_1 + u_{2i}X_2 + u_{pi}X_p\right)^T. \tag{1}$$

Where $i = 1,2, \dots, p$. Then select a specific number of m principal components to form a new sample data. Generally, the sum of the corresponding eigenvalues of m principal components is more than 85% of the total eigenvalues. In order not to lose too much information, the threshold can be increased to 95%[10].

Assumption 4. Set $\mathrm{C} = \{C_1, C_2, \dots, C_{m+1}\}$ with m principal component and principal component score divide the object into q groupings, and the degree of membership of each object C_i to the t grouping is r_{ti}, then the result of partition can be expressed as a matrix U.

Assumption 5. The target set $\mathrm{C} = \{C_1, C_2, \dots, C_{m+1}\}$, each target is a p -dimensional attribute vector, that is $C_i = \{C_{i1}, C_{i2}, \dots, C_{ip}\}$, divides the target into $c$groupings, and the t-th grouping center is also a p-dimensional vector[12] , i.e.

$$v_t = \{C_{t1}, C_{t2}, \dots, C_{tp}\}. \tag{2}$$

Theorem 1. Reconstructing $x^{(i)}$ based on $z^{(i)}$ feature space, the obtained $x^{(i)'} = \sum_{i=1}^{n} z_j^{(i)}\omega_j = W_k z^{(i)}$ equation to be optimized is

$$\mathrm{W}_{\mathrm{k}}^{*} = \arg_{W_k}\min\|x^{(i)'} - x^{(i)}\|_2^2, s.t.\, \mathrm{W}_{\mathrm{k}}^T W_k = I. \tag{3}$$

The distance between the sample point $x^{(i)'}$ based on projection reconstruction and the original sample point $x^{(i)}$ is minimized. After further simplification, the optimization objectives are as follows:

$$\mathrm{W}_{\mathrm{k}}^{*} = \arg_{W_k}\max \mathrm{tr}\left(\mathrm{W}_{\mathrm{k}}^{T} X X^{T} W_k\right), \mathrm{s.t.}\, \mathrm{W}_{\mathrm{k}}^{T} W_k = I. \tag{4}$$

After solving the problem, it is finally transformed into the eigenvalue problem of XX^T. The eigenvalue decomposition of matrix XX^T is carried out to obtain the eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_p)$. The corresponding eigenvectors $(u_{1i}, u_{2i}, \ldots, u_{pi}), i = 1,2, \ldots, p$.

Theorem 2. Matrix $U = r_{ti}$ is a fuzzy C partition if U satisfies the following conditions:

1) For $\forall\ t, i, r_{ti} \in [0,1]$;

2) For $\forall\ i$, $\sum_{t=1}^{c} r_{ti} = 1$;

3) For $\forall\ t, 0 < \sum_{i=1}^{n} r_{ti} < m + 1$.

In FCM, the target vectors of m+1 p-dimension attributes are divided into c groupings to form a set of fuzzy partitions:

$$\mathrm{M} = \left\{ \mathrm{U} \in R_{cn} \middle| \underset{\substack{1 \le t \le c \\ 1 \le i \le m+1}}{\forall} r_{ti} \in [0,1], \sum_{t=1}^{c} r_{ti} = 1, 0 < \sum_{i=1}^{n} r_{ti} < m + 1 \right\}. \tag{5}$$

Among them, $R_{c(m+1)}$denotes the space formed by all real $c \times (m + 1)$ matrices[12]. The objective function of the FCM algorithm is

$$J(U, V) = \sum_{t=1}^{c} \sum_{i=1}^{m+1} (r_{ti})^{f} d_{ti}^{2}. \tag{6}$$

Among them, $U \in M, V \in R_{c(m+1)}, f \in [1, \infty]$ are weighted exponents, which determine the similarity between fuzzy classes. d_{ti} is the distance between object C_i and center v_i of the i-th group:

$$d_{ti}^{2} = \|C_i - v_t\|^2. \tag{7}$$

3. Empirical Analysis

According to the service provided by the Didi taxi application, select indicators [12-17]:

Table 1. Hypothesis of related indicators affecting driver's service level.

X	X_2	X_3	X_4	X_5	X_6	X_7
Average star rating	Online time	Number of clicks	Successful singular	Number of completed orders	The actual total kilometres of the order	Vehicle fare income

According to formula (8) of step1 and step2, the correlation matrix is obtained by standardizing the data as follows:

$$R = \begin{pmatrix} 1 & 0.031 & -0.077 & -0.004 & 0.019 & 0.040 & 0.043 \\ & 1 & 0.400 & 0.452 & 0.447 & 0.431 & 0.461 \\ & & 1 & 0.554 & 0.551 & 0.373 & 0.428 \\ & & & 1 & 0.974 & 0.750 & 0.846 \\ & & & & 1 & 0.743 & 0.847 \\ & & & & & 1 & 0.977 \\ & & & & & & 1 \end{pmatrix}.$$

According to step3 the eigenvalues and eigenvectors of the correlation matrix are obtained. The eigenvalues are shown in Table 2.

Table 2. Comparison of Eigenvalues of correlation coefficient matrix.λ_i: Eigenvalues,$i = 1,2, \ldots, 7$.

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
2.046	1.015	0.891	0.791	0.574	0.162	0.085

According to step4 the load matrix is calculated as follows. The contribution rates of variance and cumulative variance are shown in Table 3.

$$R = \begin{pmatrix} & 0.948 & 0.238 & 0.209 & & & \\ -0.293 & & 0.654 & -0.687 & 0.119 & & \\ -0.308 & -0.279 & 0.557 & 0.598 & -0.400 & & \\ -0.459 & & -0.106 & 0.175 & 0.477 & -0.720 & \\ -0.458 & & -0.102 & 0.188 & 0.491 & 0.670 & -0.233 \\ -0.432 & -0.112 & -0.332 & -0.211 & -0.531 & -0.104 & -0.600 \\ -0.460 & & -0.283 & -0.130 & -0.274 & 0.145 & 0.765 \end{pmatrix}$$

Table 3. Variance contribution rate and cumulative variance contribution rate of components. VCR: Variance Contribution Rate. VCCR: Accumulated Variance Contribution Rate. X_i: Component,$i = 1,2, \ldots,7$.

Input	X_1	X_2	X_3	X_4	X_5	X_6	X_7
VCR	0.598	0.147	0.114	0.089	0.047	0.004	0.001
VCCR	0.598	0.745	0.859	0.948	0.995	0.999	1.000

According to the weighted method of step6 and step7, the comprehensive score is estimated, and the weight of the variance contribution rate of each principal component to the total variance contribution rate of the two principal components is taken as the weight to carry out weighted summary to obtain the comprehensive score of each driver.

$$C = (0.598 \times C_1 + 0.147 \times C_2 + 0.114 \times C_3). \quad (13)$$

According to the step8 and step9, the scores are calculated, and the results of systematic clustering of the new principal components and scores are as follows:

Table 4. Evaluation score and clustering results based on principal components (partial data).

C_1	C_2	C_3	score	cluster
223	29	21	140.011	5
145	85	39	103.651	5
218	55	47	143.807	1
123	57	35	85.923	3
270	188	267	219.534	5
238	73	91	163.429	3

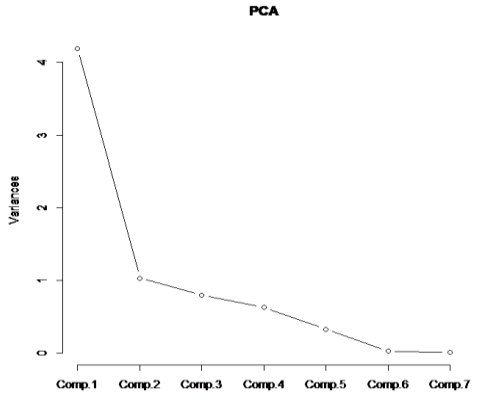

Fig. 1. Gravel map. Abscissa: principal components, ordinates: corresponding method contribution rate.

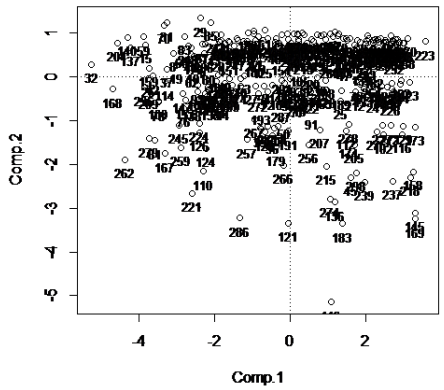

Fig. 2. Distribution of score aggregation of driver's comprehensive service level, abscissa: first principal component, ordinate: second principal component.

According to the principle that the cumulative variance contribution rate is more than 95%, three principal components are selected. The cumulative variance contribution rate is 99.5%, m=3. It can also be seen from the gravel map that m=3 is more suitable.

It can be seen from the principal component load matrix that the load values of principal component C_1 on X_4 (number of successful orders), X_5 (number of completed orders), X_6 (actual total kilometers

of orders) and X_7 (fare income) are very large, which can be regarded as the principal component of driver's order service. C_2 has a significantly larger load on X_1 (average star) than other indicators, which can be regarded as the main component of passengers' satisfaction with service. C_3 has the second largest load value on X_2 (online time) and X_3 (number of single clicks), and can be regarded as the main component of driver's online activity. Combining the explanation of each principal component with the scores and comprehensive scores of each driver on the two principal components, the comprehensive service level of each driver can be evaluated.

With the first principal component as the abscissa and the second principal component as the ordinate, the service composition maps of each driver are drawn, as shown below.

From Component Figure 3, we can see that the service score aggregation location of most drivers is: $C_1 \in (-3,3), C_2 \in (-1,0) \cup (0,1)$. Moreover, the number of positive values aggregated by C_1 is larger than that aggregated by negative values. Therefore, it can be judged that the score of C_1 (driver's order service) tends to be positive, that is to say, the general difference of order service quality is positive.

Algorithmic efficiency refers to the execution time of the algorithm. Table 5 shows that the improved algorithm can significantly reduce the execution efficiency of the algorithm. The efficiency comparison of the improved algorithm is as follows:

Table 5. Comparing the execution time of four algorithms. Unit: second.

Data volume	PCA	FCM	Total time	FCM-PCA
1mb	50	55	105	95
2mb	50	60	110	100
5mb	65	80	145	135
8mb	80	90	170	140
1gb	90	90	180	145

4. Conclusions and Suggestions

In summary, FCM-PCA is used to evaluate and cluster the services provided by 300 drivers of Didi taxi software. It has good practical pertinence. The experiment selects 7 initial indexes and extracts 3 principal components. Finally, the extracted principal components and scores are clustered as the basis for evaluating the service level of drivers. The conclusion shows that this method can effectively understand the comprehensive service situation of drivers and the psychology of passengers, and will have stronger pertinence, which will be of great significance to the improvement and formulation of policies. In addition, according to the efficiency of the algorithm in Table 5, the improved FCM-PCA algorithm not only expands the clustering function on the basis of principal component analysis, but also improves the speed of the algorithm.

In addition, according to the ranking of the final score of the principal component, we have come to the conclusion that the top four drivers in the comprehensive ranking, whether high or low, are all in the top 10 of C_1 (driver order service), which shows that the principal component has the largest proportion to the comprehensive score. At the same time, from the variance contribution rate in Table 2, we can still observe that the variance contribution rate of C_1 is 0.598, which is far greater than other principal components and plays a vital role in the driver's service. Therefore, Didi taxi app should pay more attention to this aspect in terms of education and training. For drivers, C_1 is also an indicator of great flexibility. If operators want to rapidly increase the number of customers and improve passenger safety factor, they should pay attention to the capabilities mentioned in this regard.

Acknowledgement

This research was financially supported by the 2017 Project of the 13th Five-Year Plan for Education Science in Fujian Province (Grant No. FJJKCG17-131) , Natural Science Foundation of Fujian, China(Project No. 2018J01101) and Project of School-level research incubation of Xiamen University Tan Kah Kee College (Project No. 2018L02).

References

[1] H. Y. Bian, and Y. M. Wang, The safety evaluation on road passenger transportation enterprises based on modified AHP method, *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*. vol. 4, pp. 413-419, 2009.

[2] K. Xu, L.P. Sun, J.C. Liu, and H.S. Wang, An empirical investigation of taxi driver response behavior to ride-hailing requests: A spatio-temporal perspective, *Plos One*, vol. 13, pp, 6, 2018.

[3] J. Chica-Olmo, GS. Gachs-Sanchez, and C. Lizarraga, Route effect on the perception of public transport services quality, *Transport Policy*, vol. 67, pp. 40-48, 2017.

[4] D-O. Juan, D-O. Rocio, and E. Laura, Index numbers for monitoring transit service quality, *Transportation Research Part A-policy and Practice*,vol. 84, pp. 13-30, 2016.

[5] D-O. Luigi; I. Angel, and C. Patricia, The quality of service desired by public transport users, *Transport Policy*, vol.18, pp. 217-227, 2011.

[6] J. Ji, and X.L. Gao, Analysis of people's satisfaction with public transportation in Beijing, *Habitat International*, vol.34, pp. 464-470, 2010.

[7] F. Pier Alda, and M. Giancarlo, Citizens evaluate public services: a critical overview of statistical methods for analysing user satisfaction, *Journal of Economic Policy Reform*, vol.17, pp. 236-252, 2014.

[8] F. Gaetano, C. Chiara, and C. Luciano, Short-term traffic predictions on large urban traffic networks: applications of network-based machine learning models and dynamic traffic assignment models, *International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pp. 93-101, 2015.

[9] J. Asensio, and A. Matas, Commuters' valuation of travel time variability, *Transportation Research Part E-logistics and Transportation Review*, vol. 44, pp. 1074-1085, 2008.

[10] L. Marielle, J. J. Meulman, and P.J. F. Groenen, Nonlinear principal components analysis: Introduction and application, *Psychological Methods*, vol.12, pp. 336-358, 2007.

[11] Z. F. Wang, K. H. Xu, D. P. Kong and L. Z. Li, Cooperative task assignment method based on fuzzy clustering ——auction mechanism, *Fire Control and Command Control* , vol. 44, pp. 102-111, 2019.

[12] F. Pier Alda, P. Laura, and F. V. Carlo, A two-step approach to analyze satisfaction data, *Social Indicators Research*, vol.104, pp. 545-554, 2011.

[13] J. Erik, and H. N. Koutsopoulos, Travel time estimation for urban road networks using low frequency probe vehicle data, *Transportation Research Part B-methodological*, vol.53 , pp. 64-81, 2013.

[14] W. L. Min, and W. Laura, Real-time road traffic prediction with spatio-temporal correlations, *Transportation Research Part C-emerging Technologies*, vol. 19, pp. 606-616, 2013.

[15] S. Axel, and Z. Henryk, Travel time prediction using floating car data applied to logistics planning, *Ieee Transactions on Intelligent Transportation Systems*, vol. 12 , pp. 243-253, 2011.

[16] V. L. J. W. C, and V. H C. P. I. J, Short term traffic and travel time prediction models, *Transportation Research Circular*, pp. 22-41, 2012.

[17] Q.Ye, S. W. Y, and W. S. C, Short-term traffic speed forecasting based on data recorded at irregular intervals, *Ieee Transactions on Intelligent Transportation Systems*, vol. 13, pp. 1727-1737, 2012.