Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation
EJASA:DSS (2011). Vol 2, Issue 1, 15 – 24
e-ISSN 2037-3627, DOI 10.1285/i2037-3627v2n1p15

CUMULATIVE CORRESPONDENCE ANALYSIS TO IMPROVE THE PUBLIC TRAIN TRANSPORT

Pasquale Sarnacchiaro*[(1)], Antonello D'Ambra[(2)]

[(1)] Faculty of Economics, University Unitelma Sapienza, Italy
[(2)] Department of Accounting Management and Quantitative Methods,
Second University of Naples, Italy

Received 03 March 2010; Accepted 08 September 2011
Available online 17 December 2011

Abstract: *To a company, improving customer satisfaction (CS) is a strategic element to increase market share, for example in public transport a growth of users implies a positive effect as private transport reduced.*
Consequently, the use of statistical methods directed to the analysis of CS is very important to get information necessary to support strategic decisions. In this paper, through Taguchi method for design of experiment, we set nine different hypothetic public train transport (called scenarios) and we asked to potential users to evaluate them by an ordinal scale. In order to analyse this data we use an integrated approach based on Cumulative Correspondence Analysis and the Taguchi's index, it allows to find a scenario that maximizes the satisfaction of potential users.

Keywords: *Public transport, cumulative correspondence analysis, Taguchi's statistic, expected customer satisfaction, bootstrap confidence region.*

1. Introduction

Nowadays, in high-tech age it is very difficult to offer users a new type of product. Under the worldwide competition, a new type of product is a very expensive task, it requires time, investments and a thought study of market conditions. To increase the demand for manufactured goods, minor structural changes or no changes are needed. For enterprise, the primary purpose is profit, which can be obtained from the sale of goods or services. It is the user who chooses what

*Corresponding Author. Email: pasquale.sarnacchiaro@unitelma.it goods to buy or services to use. Under these market assumptions, it is not difficult to get the conclusion: the better the manufacturer meets the needs and desires of the consumer, the greater his economic effect will be.

Customer satisfaction (CS) is generally conceived as the sense of satisfaction that a consumer feels when he compares his preliminary expectations with the perceptions of the acquired product. Thus, satisfaction is closely linked with the quality of the product (service). The quality of services is extremely difficult to assess, because services are intangible and heterogeneous. It is not feasible to develop well-defined quality standards that will allow the manufacturer to exercise full control and reject "unsuitable" services before they are sold to the consumer. The direct involvement of consumers in the supply of services leads to take into account quality not as an objective standard characteristic, but as quality element conceived by the consumer.

This definition of quality implies a close relationship between the service and the consumers. In current market conditions, the consumer plays the role of arbiter, so failure will result if CS is not taken into account [7].

Therefore, for a company, the aim to improve CS is strategic to increase market share, moreover, in public transport, if customers/users increase there is a positive effect due to reduced use of private transport.

Consequently, the use of statistical methods directed to the analysis of CS is very important to get information necessary to support strategic decisions. Most statistical studies of CS deal with ordered categorical data.

The aim of this research is to find the most favourable composition for a services of public train transport. To this purpose, through the application of Taguchi method for design of experiment (DOE), we constructed nine hypothetic services of public train transport (called scenario). Then, we submitted these scenarios to potential users that expressed their evaluations by an ordinal scale. In order to analyse the collected data, we used an integrated approach based on Cumulative Correspondence Analysis (CCA) and Taguchi's index [1]. Next, in the graphical representation of CCA, in order to have a deeper analysis of the levels/factors of the scenarios, we introduced the bootstrap confidence region that allowed to identify, from a statistical point of view, the impact of each level on expected CS.

The paper is organized as follows: Section 2 presents the case study, particularly the nine scenarios for the train Naples-Rome obtained by Taguchi’s approach for DOE. Section 3 points out the main aspects of CCA. Section 4 shows the results of CCA applied to the data of the case study and advices an optimal settings for train Naples-Rome; an helpful application of the bootstrap confidence region ends the section.

2. Research plan

The proposed research, which regards the train from Naples to Rome (and vice versa), has the aim to choose the best scenario for this service. To this purpose, we have used the Taguchi Method (TM) for DOE [13], [14] and [15].

DOE can be highly effective when you wish to optimize product and process designs, study the effects of multiple factors on the performance and solve production problems by objectively laying out the investigative experiments. You can also find out which factor should have tighter tolerance and which tolerance should be relaxed. The information from the experiment will tell you how to allocate quality assurance resources based on objective data, and moreover, it will indicate how to combine different factors in their proper settings to get the best results.

By using TM, the DOE can economically satisfy the needs of problem solving and product/process design optimization projects. By learning and applying this technique, engineers, scientists and researchers can significantly reduce the time required for experimental investigations.

TM is a technique to lay out experimental plan in most logical, economical, and statistical way. You can potentially benefit from it when you want to determine the most desirable design of your product/service, the best parameter combination for your process.

The TM, also called orthogonal arrays, allows you to analyze many factors with few runs. TM are balanced, that is, no factor is weighted more or less in an experiment, thus allowing factors to be analyzed independently of each other. Engineering knowledge should guide the selection of factors, levels and responses.

In our case we have selected four control factors (A= Journey time; B= Comfort, C=Cost; D= Frequency) each of them with 3 levels (Table 1). Then the number of different runs in a complete factorial design would be 81.

Table 1. Train Naples-Rome - Control factors and levels.

	Levels		
	1	**2**	**3**
A. Journey time	1:29	1:54	2:09
B. Comfort	70cm	75cm	80cm
C. Cost	€ 19,50	€ 27,60	€ 36,10
D. Frequency	Each four hours	Each two hours	Each hour

In order to analyse this experiment by using TM, the most desirable and suitable design was an L-9 Orthogonal Array (table 2). We chose this design because we supposed that for potential user to image more than nine scenarios was very difficult, even if this choice implied that main effects were confounded with two factors interactions.

Table 2. L_9 orthogonal array and factor assignment – Definition of the scenarios.

	Column (factors)			
Exp. No (Scenario)	**A**	**B**	**C**	**D**
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

Then, we submitted the nine experimental trials to a group of potential users (h=58), particularly we required an evaluation for each scenario using an ordinal scale with five categories (I, II, ...,V), where I was the lowest evaluation and V the highest evaluation.

In order to transform the modalities of an ordinal scale into values expressed in an interval scale, there are many methods proposed in very different area of research, for more details see [16], [2] and [10].

Natural scoring scheme transforms the ordinal categories in natural numbers assigning one to the smallest category and S (total number of categories, five in our case) to the biggest. Natural score is the simplest method and perhaps the prevalent. With this method we assume that the distance between contiguous categories is the same (equally spaced). In our case after quantification with natural score method, for each scenario evaluation we have computed the main statistical indexes (table 3).

Table 3. Train service Naples Rome - Customer Satisfaction evaluation.

	Scenario								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Mean	2,31	2,81	2,67	3,31	1,86	4,22	2,60	4,62	2,83
Median	2,00	3,00	3,00	3,00	2,00	4,00	3,00	5,00	3,00
St. Dev.	3,36	5,66	5,71	7,24	4,66	15,78	0,91	21,83	9,16

3. Cumulative correspondence analysis and Taguchi's statistic

Consider a two-way contingency table, *N*, that cross-classifies *n* statistical units according to *I* row categories and *K* ordered (ascending) column categories. Denote n_{ik} as the (*i; k*)-th joint cell frequency with a relative frequency of $p_{ik} = n_{ik} / n$ which is the (*i; k*)-th element of the matrix *P*. Let $n_{i\bullet}$ and $n_{\bullet k}$ be the *i*-th row and *k*-th column marginal frequencies respectively. Furthermore let D_I be a diagonal matrix where the (*i; i*)-th element is $p_{i\bullet} = n_{i\bullet} / n$. Denote $Z_{ik} = \sum_{j=1}^{k} n_{ij}$ the cumulative frequency of the *i*-th row category up to the *k*-th column category and their consideration provides a way of ensuring that the ordinal structure of the column categories is preserved. Similarly, denote $d_k = \sum_{j=1}^{k} n_{\bullet j} / n = \sum_{j=1}^{k} p_{\bullet j}$ the cumulative relative frequency up to the *k*-th column category.

The following statistic has been proposed to assess the association between the row and ordered column variables [11]:

$$T = \sum_{k=1}^{K-1} w_k \left[\sum_{i=1}^{I} n_{i\bullet} \left(\frac{Z_{ik}}{n_{i\bullet}} - d_k \right)^2 \right] \qquad 0 \le T \le \left[n \left(I - 1 \right) \right] \tag{1}$$

where w_k are suitable weights. The link between the Pearson chi-squared statistic and the Taguchi's statistic has been proposed [9]:

$$T = \sum_{k=1}^{K-1} \chi_k^2 \tag{2}$$

Here χ_k^2 is the Pearson chi-squared for $I \times 2$ contingency table obtained by aggregating column categories 1 to k and aggregating the column categories (k+1) to K. For this reason T is also referred to as “cumulative chi-squared statistic". The inertia for the Taguchi's statistic will be T/n.

Correspondence analysis can be performed by partitioning the Pearson chi-squared statistic into the sum of squares of the singular values of the Pearson ratio's. If we define

$$Y = W^{1/2}(U - d1')P'D_I^{-1/2} \tag{3}$$

where W is a diagonal square matrix of dimension *K-1* in which the general term is w_k, U is a lower triangular matrix of dimension $(K-1\times K)$ with all terms equal to 1, d is a column vector of dimension $(K-1)$ contained the cumulative relative frequency d_k (*k=1,2,...,K-1*), **1** is a row vector of dimension (K) with all terms equal to 1. Then, the total inertia can be expressed as $\frac{T}{n} = tr[Y'Y]$.

Realizing the Singular Value Decomposition of the matrix Y, we have $Y = \tilde{A}D_\lambda\tilde{B}'$, where $\tilde{A}$ is a matrix of dimension $[(K-1)\times M]$ composed by the right eigenvectors, $\tilde{B}$ is a matrix of dimension $[I \times M]$ composed by left eigenvectors, D_λ is a diagonal eigenvalues matrix of dimension $[M \times M]$ and $M = \min[I,(K-1)]$ with general term λ_m.

In order to visualize the association between the ordinal column categories and the nominal row categories we use the following row, column and supplementary column coordinates [1]:

$$F = D_I^{-1/2}\tilde{B}D_\lambda \tag{4}$$

$$G = W^{-1/2}\tilde{A}D_\lambda \tag{5}$$

$$G^+ = P'\tilde{B}D_\lambda \tag{6}$$

Respect to the standard coordinates of correspondence analysis [5], we prefer to use these coordinates since they take into account the ordinal nature of column variable [3]. In order to weight the contribution that each modality provides to the response variable (CS), we represent in CCA plot, for each modality of the predictor variables (factor levels), the confidence region obtained by bootstrap procedure [4].

4. Results

Starting from the results of CS obtained on the nine scenarios and presented in the paragraph 2, the favourite scenario, in the sense of the mean/median, was the eighth followed by the sixth (table 3).

However, we observed that the eighth scenario was feasible from a technical point of view, but it was not practicable from an economic point of view (in fact the factor-level *low-cost* and *low journey time* are incompatible).

Moreover, from table 3 we couldn't point out the impact of each factor-level on the CS. So we were not able to identify the factor-level to change to improve CS. In order to answer this question and to determinate the optimal setting for train service we performed CCA.

The data table for CCA is the following: in the rows the different levels for each quality characteristics (factors) and in the columns the ordinal categories (Table 4).

Table 4. Data Matrix for Cumulative correspondence analysis of ordered categories.

		Evaluation				
	Factor	**I**	**II**	**III**	**IV**	**V**
A1	Low Journey time	9	49	42	20	54
A2	Middle Journey time	26	31	42	44	31
A3	High Journey time	31	56	46	34	7
B1	Low comfort	19	53	62	34	6
B2	Middle comfort	31	43	29	20	51
B3	High comfort	16	40	39	44	35
C1	Low cost	13	26	29	35	71
C2	Middle cost	8	52	65	33	16
C3	High cost	45	58	36	30	5
D1	Low frequency	42	67	41	13	11
D2	Middle frequency	11	42	47	47	27
D3	High frequency	13	27	42	38	54

In table 4 the number of statistical units is given by the product of the *number of Scenarios, the number of factors* and *the number of potential users* $\left(9\times 4\times 58\right)$. Starting from Table 4, it is possible to construct four contingency tables (dimensions 12 x 2) obtained by the unification of the ordered modalities of the evaluation variable (I vs II-III-IV-V, I-II vs III-IV-V and so on – Table 5).

Table 5. Number of observations by cumulative categories.

I	**II+III+IV+V**	**I+II**	**III+IV+V**	**I+II+III**	**IV+V**	**I+II+III+IV**	**V**
9	165	58	116	100	74	120	54
26	148	57	117	99	75	143	31
31	143	87	87	133	41	167	7
19	155	72	102	134	40	168	6
31	143	74	100	103	71	123	51
16	158	56	118	95	79	139	35
13	161	39	135	68	106	103	71
8	166	60	114	125	49	158	16
45	129	103	71	139	35	169	5
42	132	109	65	150	24	163	11
11	163	53	121	100	74	147	27
13	161	40	134	82	92	120	54

All the chi-squared statistics computed are resulted significant (table 6). The sum of the four chi-squared is equivalent to T-Taguchi’s index [12] and [6].

Table 6. Pearson-Chi-squared and T-Taguchi’s index for contingency tables.

Index	Value	DoF	P-value	Percentage
$\chi^2_{1vs2345}$	93,66	11	0,000	15,06
$\chi^2_{12vs345}$	135,49	11	0,000	21,78
$\chi^2_{123vs45}$	171,51	11	0,000	27,58
$\chi^2_{1234vs5}$	221,31	11	0,000	35,58
T	621,97			

It is well known that classical Pearson’s chi-squared test for contingency table does not have good power against ordinal alternatives [8].

The CCA of ordered categories computed on table 4 showed that the first two eigenvalues explained more than 97,1% of the total inertia (Table 7).

Table 7. Cumulative correspondence analysis - Eigenvalues and decomposition of Taguchi’s index.

λ^2	Percentage	Cumulative percentage
0,2505	84,1	84,1
0,0386	13,0	97,1
0,0080	2,7	99,8
0,0007	0,2	100,0

The sum of the squares of the eigenvalues is equivalent to T-Taguchi’s index divided by the total of the frequency of the table 4. In order to visualize the association between the ordered column categories and the nominal row categories, we computed the row and column coordinates according to the formula (4) and (6) (figure 1).

To represent the column we projected, as supplementary points, the ordered categories of the evaluation variable in dual space. We prefer to use the coordinates expressed in formula (6), instead of the formula (5), because the interpretation is more comprehensible and it is functional to our purpose. In fact, if we assign to each ordered category of the CS the closer level of each factor, through this graphical representation, a typical setting for train Naples-Rome can be defined (Figure 2).

This information, matched with the enterprise economic planning, allows to decision maker to have a very useful tool to set the transport service taking into account the expected level of CS. In fact, if the enterprise for business opportunities can’t offer the scenario that maximise CS (the combination A1-B3-C1-D3), starting from the level of CS achieved and after choosing the acceptable level of CS, the decision maker may define the levels and the factors to modify.

Analysing figure 2 we observed that the “characteristic scenario” corresponding to the lower level CS is not different from the scenario of the next level of CS.

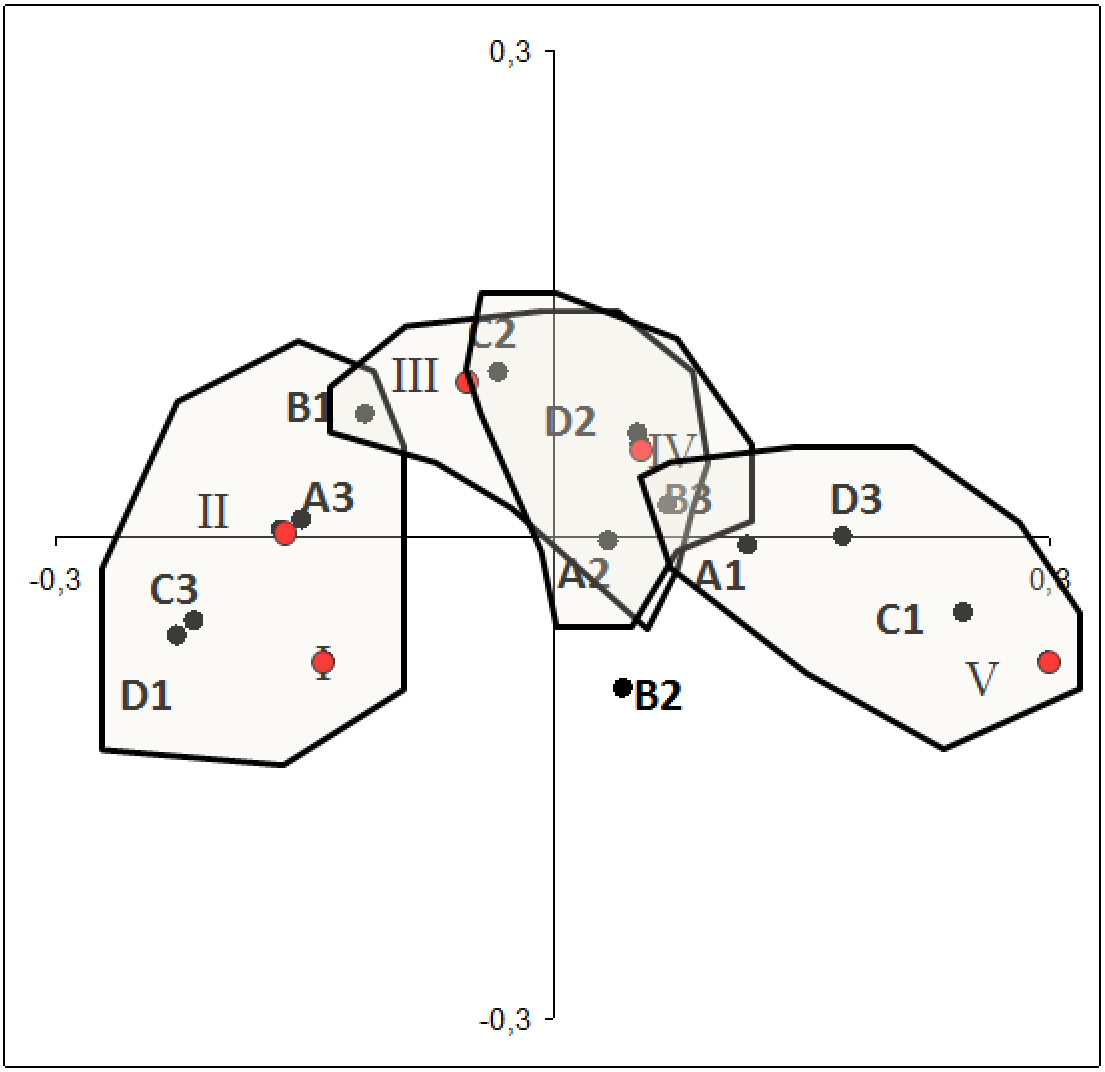


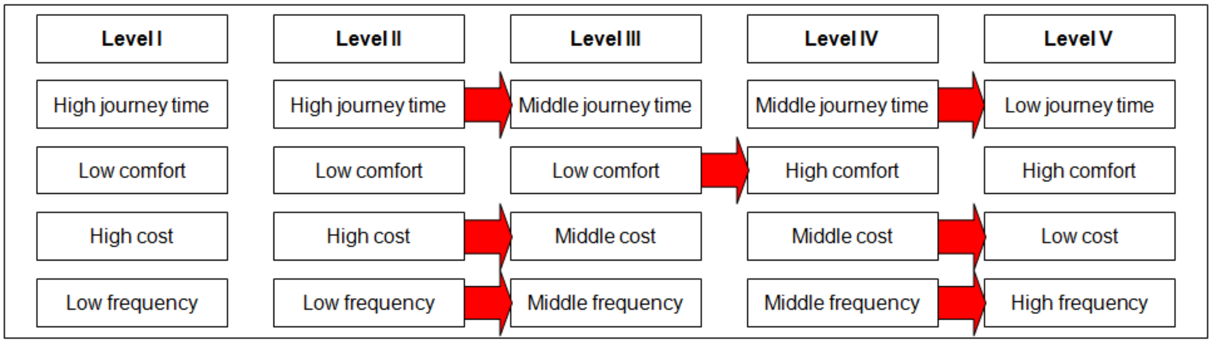
Figure 1. Cumulative correspondence analysis - Graphical representation on the first two axis.

Differently, if the enterprise needs to pass from a level II to a level III of CS, the enterprise have to improve simultaneously the factors “cost”, “frequency” and “journey time”.

In the same way, to pass from level III to level IV, the client asks to improve significantly the factor “comfort”.

The last scenario matching to the maximum level of CS (V) is not very interesting because it corresponds to the best level for each control factor.

Subsequently, we represent in CCA plot, for each modality of the predictor variable the confidence region obtained by bootstrap procedure. Particularly, starting from the original dataset (Table 4) we represent this data matrix through the product of two complete disjonctive table. The first one of dimensions (*12*x*2088*) and the second of dimensions (*2088*x5), then we compute 1000 bootstrap samples on the first data matrix sampling the columns. Bootstrap method consists in the extraction with replacement of repeated random samples (bootstrap sample). From each of the bootstrap samples, we compute a “news” original dataset and afterwards we calculate the first two vectors for the row coordinates according to formula (6).



Level I	Level II	Level III	Level IV	Level V
High journey time	High journey time	Middle journey time	Middle journey time	Low journey time
Low comfort	Low comfort	Low comfort	High comfort	High comfort
High cost	High cost	Middle cost	Middle cost	Low cost
Low frequency	Low frequency	Middle frequency	Middle frequency	High frequency

Figure 2. The characteristic scenario for each level of customer satisfaction.

Finally, in order to represent the bootstrap confidence region (95%), we calculate mean and variance for each of the row coordinates.

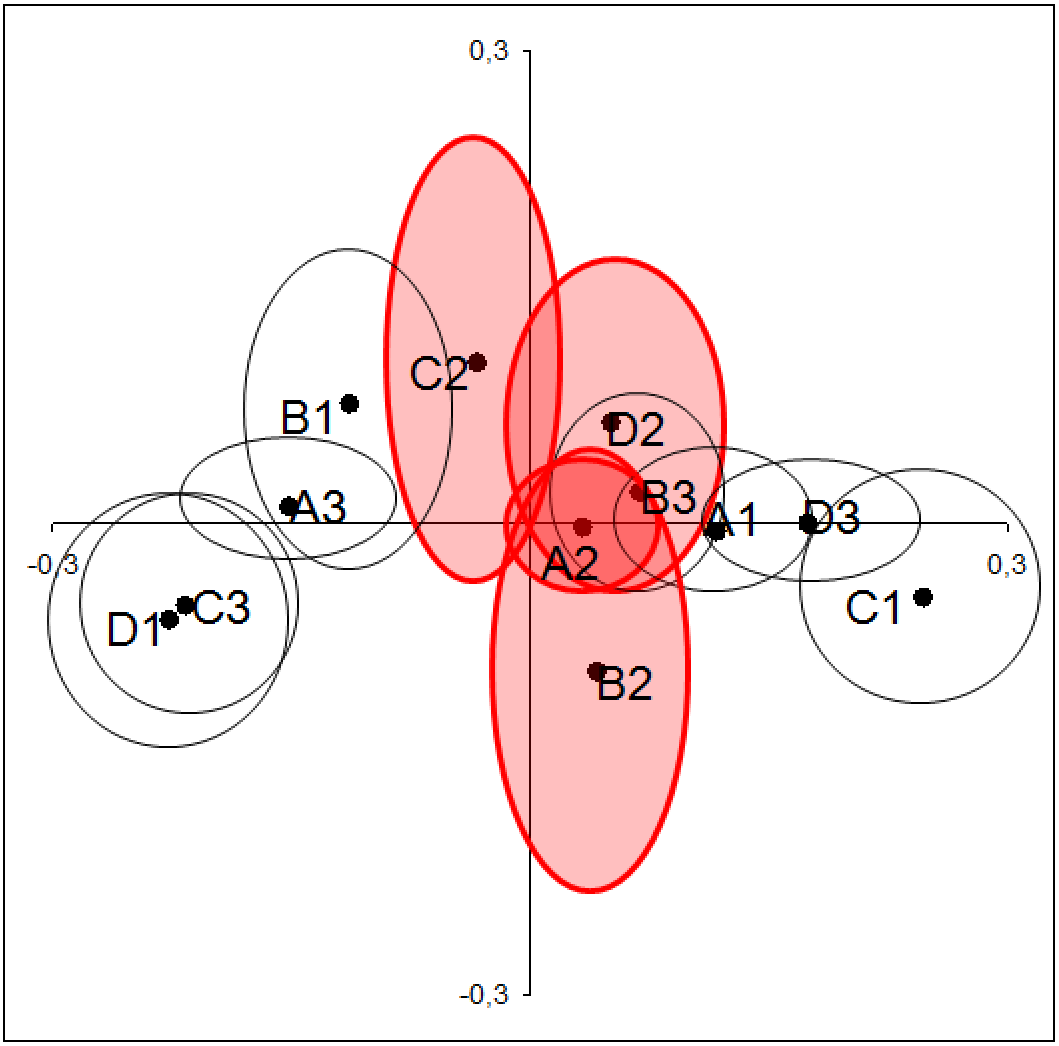


Figure 3. Bootstrap Confidence region for factor levels.

In this way, we verified the impact of the factor level of control variables on CS (figure 3), particularly if the confidence region includes the origin of axes, we state that the modality is not statistically significant. Therefore, in our case, we can say that the influent modalities on CS are the lowest and highest levels for each control factor.
In this paper we have considered the problem to find the most favourable composition for a services of public train transport. To this purpose, we used an integrated approach based on the application of TM for DOE and CCA. Next, in the graphical representation of CCA, in order to have a deeper analysis of the levels/factors of the scenarios, we introduced the bootstrap confidence region that it allowed to identify, from a statistical point of view, the impact of each level on expected CS.

References

[1]. Beh, E. J. , D'Ambra, L., Simonetti B. (2010). Correspondence Analysis of Cumulative Frequencies Using a Decomposition of Taguchi's Statistic, *Comunication in Statistics*, 40 (9), 1620-1632.

[2]. Bradley, R.A. Katti, S.K. Coons, I.J., (1962). Optimal scaling for ordered categories. *Psychometrika*, 27, 355-374.

[3]. D’Ambra L., Koksoy O., Simonetti B. (2009). Cumulative correspondence analysis of ordered categorical data from industrial experiments. *Journal of applied statistics*. 36, 1315-1328.

[4]. Efron B., Tibshirani R.J. (1993). *An introduction to the bootstrap*, Chapman & Hall

[5]. Greenacre, M. J. (1984). *Theory and application of correspondence analysis*. London: Academic Press.

[6]. Hirotsu, C. (1986). Cumulative chi-squared as a tool of goodness of fit, *Biometrika*, 73, 165-173.

[7]. Krivobokova O.V. (2009). Evaluating Customer Satisfaction as an Aspect of Quality Management. World Academy of Science, *Engineering and Technology*, 53.

[8]. Nair, V.N. (1986). Testing in industrial experiments with ordered categorical data. *Technometrics*, 28(4), 283-291.

[9]. Nair, V. N. (1987). Chi-squared type tests for ordered alternatives in contingency tables. *Journal of the American Statistical Association*, 82, 283-291.

[10]. Parsa, R. A. Smith, W.B., (1993). Scoring Under Order Constraints in Contingency Tables, *Communications in Statistics*, 22 (12), 3537-3552.

[11]. Taguchi, G. (1966). *Statistical Analysis* (in Japanese). Tokyo: Maruzen,

[12]. Taguchi, G. (1974). A new statistical analysis for clinical data, the accumulating analysis, in contrast with the chi-square test. *Saishin Igaku*, 29, 806-813.

[13]. Taguchi, G. (1991a). *Taguchi methods: Case studies from the U.S. and Europe*. Michigan: American Supplier Institute.

[14]. Taguchi, G. (1991b). *Taguchi methods: Research and development*. Michigan: American Supplier Institute.

[15]. Taguchi, G. (1991c). *Taguchi methods: Signal-to-noise ratio for quality evaluation*. Michigan: American Supplier Institute.

[16]. Zanella, A. Cantaluppi, G., (2004). Simultaneous transformation into interval scales for a set of categorical variables. *Statistica*, 2, 401-426.