THE KARAP METHOD AS A DATA MINING TOOL

Abstract

In market research the KARAP method is mainly used in connecting each respondent with a particular criterion from p criteria set in the questionnaire. This connection is made through the use of Multivariate Data Analysis and the Euclidean vector space $R^{(p-1)}$ created by p-1 factorial axes.

Overview

An ascending hierarchical classification of the «subject» (rows) of a data table T(n,p) is a procedure that produces a partition sequence of an initial set into non-empty and foreign two-by-two subsets between them, called **classes**, one within the other, interconnecting each time only two classes, which based on a metric, present the smallest distance in each grouping step.

As far as it can be understood, the purpose of ascending hierarchical classification is to group all statistical units of a population into a limited number of homogeneous classes, the so called «**clusters**» as to the behaviour of certain variables, taking into account all variables, so that each one is as much as possible different from the others.

In particular, when using the VACOR method for clustering, clusters are created based on an objective algorithm (Ward's algorithm), apart from the subjective methods that may be developed by each researcher. We say objective algorithm because statistical units' grouping is performed with no a priori hypothesis in the original data table according to x^2 metric.

In classes created by the VACOR method, the multitude of variables that characterise them are identified. This is possible through the aid of MAD programme with table «Variables' contribution in classes' characterization» in combination with table «Variables' contribution in the breakdown of k higher nodes». Consequently, the «subjects» involved in classes' formations are also associated with variables characterising each class.

It is known that using the VACOR method with the known programmes implementing Ascending Hierarchical Classification, it is not possible to connect each «subject» to a particular variable, unless a statistical test of the difference in proportions at 5% confidence level is used, between ratio P_δ and each of the p values corresponding to the sample specified in each «subject» and ratio P_M of each variable to the sum of p variables (Morineau A, 1984).

$$\begin{aligned} H_0 &: P_\delta - P_M = 0 \\ H_1 &: P_\delta - P_M > 0 \end{aligned} \tag{1}$$

Then z value is calculated

$$z = \frac{P_\delta - P_M}{s_p} \tag{2}$$

Where $s_p = \sqrt{\frac{P_M \cdot Q_M}{n}}$ $\quad with \quad$ $Q_M = 1 - P_M$

As discussed by Professor Th. Bechrakis. «Each group's problems are classified based on the test value. Test value is a criterion used for problem selection. characterising each group. The greater the test value for a particular group and a particular problem. the more characteristic this problem is for this particular group» (Bechrakis. Th. p. 74)

Therefore, when value $z>1.645$ results from association (2) the alternative hypothesis H_1 is true, i.e. M variable with P_M ratio strongly characterises the class or subject with a ratio of P_δ. In different z values, where $-1.645<z<1.645$, the variable simply shows medium association (positive or negative) depending on the value it presents), while when value $z<-1.645$ is the result, the absence of variable dependence to the class (or the object) is considered strong.

On the other hand, the association of the subjects with each variable can be performed as long as the data table T(n,p) is analysed through the Correspondence Analysis, the coordinates of p variables and n objects are extracted on factorial axes p-1, and then the distances of each object A from each variable B are calculated,

using the following association, which calculates the distance between two vectors of the vector space R^n (Serge Lang p. 16)

$$\|A - B\| = \sqrt{(A - B) \cdot (A - B)} \qquad (3)$$

Initially, the results of the association of «subjects» with the resulting variables will be compared, firstly following classification using the VACOR method, using the abovementioned test with z distribution, and secondly through the placement of all points (rows and columns) of the data table T(n,p) in the Euclidean vector space $R^{(p-1)}$,based on the **factors** resulting from the implementation of the Correspondence Analysis. This project will also serve as a new process for the classification of «subjects».

The selection of the Euclidean space R^{p-1} was based on the fact that factorial axes create an orthonormal basis for space $R^{(p-1)}$, where p variables are placed based on the coordinates as well as n rows of the data table in their actual places, from which all information of the data table under analysis are provided.

At this point we should present an example to verify the abovementioned suggestions.

Associations between factors

Based on the coincidence table of two qualitative variables A and B the corresponding relative frequencies table is created.

Table 1: Relative frequencies table

A \ B	b_1 b_2 ... ... b_j ..... b_p	
a_1 a_2 . a_i . . a_n	..... f_{ij}	f_i
	f_j	

Correspondence Analysis allows not only the geometric and algebraic ascertainment of deviation from independence state of the two qualitative variables X and Y, but also the exploration of similarities among the corresponding distributions (profiles) of the table's rows and columns, corresponding to the total ratings of the two variables X and Y.

Then, rows profile table is created

Table 2: Row profile table f_J^I

Ratings	$\mathbf{b_1}$……..$\mathbf{b_j}$……$\mathbf{b_p}$	
a_1	.	
.	.	
a_i	.....$f^i_j = f_{ij}/f_i$	1
.	.	
a_n	.	

To make clear why it is preferable to use the profile of a table's row as vectorial expression of the respective statistical unit i. instead of the row with the original data, the answer is as follows: Since two rows are proportional to each other, their profiles will be identical, and when represented on a graph, the graphical representations of the corresponding vectors will coincide, while, to the contrary, rows with the original data will represent two collinear vectors.

This finding is very important since interest in the Correspondence Analysis is focused in the proportions of the «subjects» within «variables» ratings.

Projections of f^J_i points of cloud $N(I)_J$ of the data table's rows on the factorial axes Δ_a (a=1....p-1), which are denoted $F_a(i)$ (where $i=f^i_j$ any row profile), constitute the coordinates of these points upon axes Δ_a. Each coordinate $F_a(i)$ relevant to factorial axis a is called **factor a of profile i** (J-P & F. Benzecri 1980 p. 65)

To define the factorial axes Δ_a at a plane (a=1,2) we use Huyghens' theorem, which states that total inertia I_{Total} of cloud N(I) can be analysed in two parts. The first part relates to inertia $I_{//\Delta}$ along a line Δ_a which crosses the barycentre $G\{=f_J\}$ of the cloud and the other one in inertia $I_{\perp\Delta}$ vertical line Δ_α.

Among the infinite lines passing through point G, the one that maximises inertia $I_{//\Delta}$ is obtained and thus minimises inertia $I_{\perp\Delta}$.

I.e.

$$_{Total} = I_{//\Delta} + I_{\perp\Delta} \quad (4)$$

Graphically:

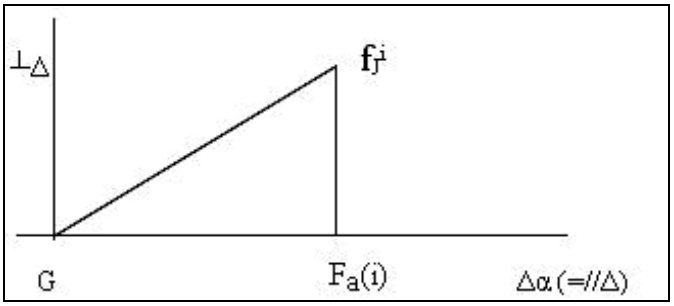

Figure 1: Total inertia breakdown

As we know a characteristic vector $u_a=\varphi_a^J$ associated with factorial axis Δ_a corresponds to every characteristic root λ_a of a square table $S_J^J = f_J^I \circ f_I^J$.

For each vector φ_a^J (j=1....p) its coordinates satisfy the following associations:

$$\sum_{j=1}^{p} f_j \cdot \varphi_a^J = 0 \quad (5)$$

$$\sum_{j=1}^{p} f_j \cdot (\varphi_a^J)^2 = 1 \quad (6)$$

Each factor $F_a(i)$, which is a vector, is calculated using the following association

$$F_a(i) = \sum_{j=1}^{p} f_j^i \cdot \varphi(j) \quad (7)$$

In each factorial axis Δ_a the following associations apply:

$$\sum_{i=1}^{n} f_i \cdot F_a(i) = 0 \quad (8)$$

$$\sum_{i=1}^{n} f_i \cdot (F_a(i))^2 = \lambda_a \quad (9)$$

While for two different factorial axes Δ_r and Δ_s the following applies:

$$\sum_{i=1}^{n} f_i \cdot F_r(i) F_s(i) = 0 \qquad (10)$$

Number $F_a(i)$ in absolute value measures the distance between the centre of gravity $G=\{f_J\}$ of cloud $N(I)_J$ from the projection of profile f_{ij} (which represents row i of the data table) on axis Δ_a.

Generally the following applies

$$d^2\left(f_J^i,\ G\right) = \sum_{a=1}^{p-1} F_a^2(i) \qquad (11)$$

Therefore, the distance separating the centre of gravity $G=\{f_J\}$ from the projection of profile row f_j^i e.g. at factorial plane $\Delta_1 x \Delta_2$ is the hypotenuse of a right triangle with sides $F_1(i)$ and $F_2(i)$. I.e. for factorial level 1x2 the following relationship applies.

$$d^2\left(f_J^i,\ G\right) = F_1^2(i) + F_2^2(i) \qquad (12)$$

The distance $d^2(f_J^i, f_J)$ is also calculated using the following relationship

$$d^2\left(f_J^i,\ G\right) = \sum_{j=1}^{p} \frac{1}{f_i} (f_j^i - f_J)^2 \qquad (13)$$

Relationship 12 indicates that factorial axes, upon which factors $F_a(i)$ are calculated, are rectangular, and therefore the system of p-1 factorial axes constitutes the construction of orthonormal coordinate system in R^{p-1} vector space. Subsequently we will present a numerical example to verify relationship 13.

Numerical example

Using a simple numerical example it is possible to ascertain easily the validity of relationships 12 and 13.

The following coincidence table is given

Table 3: Data Table

Tags	**J_1**	**J_2**	**J_3**	**Sum**
I_1	0	1	0	1
I_2	1	0	1	2
I_3	1	1	0	2
I_4	0	0	1	1
	2	**2**	**2**	**6**

Firstly we find the profiles of rows f_J^I and columns f_I^J. I.e.

$$f_J^I = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{και} \quad f_I^J = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \end{pmatrix}$$

Profiles f_I^J emerged after getting inverse table T(4,3)

For the basic application of Correspondence Analysis finding the symmetric square table S_J^J is required, calculated with the following product.

$$S_J^J = f_J^I \circ f_I^J = \frac{1}{2}\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \cdot \frac{1}{2}\begin{pmatrix} 0 & 2 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \frac{1}{4} \cdot \begin{pmatrix} 2 & 1 & 1 \\ 1 & 3 & 0 \\ 1 & 0 & 3 \end{pmatrix}$$

Then the three characteristic roots of square table S_J^J which are as follows:

$$\lambda_0 = \frac{4}{4} = 1 \;\; \lambda_1 = \frac{3}{4} \;\; \lambda_2 = \frac{1}{4}$$

As is known, in every characteristic root λ_i (apart from the trivial root λ_0) corresponds a characteristic vector $u_a=\varphi_a^J$, which is connected to factorial axis Δ_a.

For each characteristic vector φ_a^J (j=1....p) its coordinates meet, as already mentioned above, relationships 5 and 6

Characteristic vectors' values are presented in relationships 14 and 15

$$\varphi_{11}=0\ ,\ \varphi_{12}=\frac{\sqrt{6}}{2}\ \text{and}\ \varphi_{13}=-\frac{\sqrt{6}}{2} \tag{14}$$

$$\varphi_{21}=-\sqrt{2}\ ,\ \varphi_{22}=\frac{\sqrt{2}}{2}\ \text{and}\ \varphi_{23}=\frac{\sqrt{2}}{2} \tag{15}$$

Using relationship 7, factors F_a^I have the following values.

$$F_a^I=\begin{pmatrix}0 & 1 & 0\\ 1/2 & 0 & 1/2\\ 1/2 & 1/2 & 0\\ 0 & 0 & 1\end{pmatrix}\cdot\begin{pmatrix}0 & -\sqrt{2}\\ \frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2}\\ -\frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2}\end{pmatrix}=\begin{pmatrix}\frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2}\\ -\frac{\sqrt{6}}{4} & -\frac{\sqrt{2}}{4}\\ \frac{\sqrt{6}}{4} & -\frac{\sqrt{2}}{4}\\ -\frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2}\end{pmatrix}$$

We find that relationship 9 is verified

$$\lambda_1=\frac{1}{6}\cdot\left(-\frac{\sqrt{6}}{2}\right)^2+\frac{2}{6}\cdot\left(\frac{\sqrt{6}}{4}\right)^2+\frac{2}{6}\cdot\left(-\frac{\sqrt{6}}{4}\right)^2+\frac{1}{6}\cdot\left(\frac{\sqrt{6}}{2}\right)^2=\frac{3}{4}$$

$$\lambda_2=\frac{1}{6}\cdot\left(\frac{\sqrt{2}}{2}\right)^2+\frac{2}{6}\cdot\left(-\frac{\sqrt{2}}{4}\right)^2+\frac{2}{6}\cdot\left(-\frac{\sqrt{2}}{4}\right)^2+\frac{1}{6}\cdot\left(\frac{\sqrt{2}}{2}\right)^2=\frac{1}{4}$$

Factors of variables G_a^J can be found using the relationship

$$G_a^J=\frac{1}{\sqrt{\lambda_\alpha}}F_a^J\circ f_I^J \tag{16}$$

Which shows

$$G_a^J = \begin{pmatrix} 0 & -\frac{\sqrt{2}}{2} \\ \frac{3\sqrt{2}}{4} & \frac{\sqrt{2}}{4} \\ -\frac{3\sqrt{2}}{4} & \frac{\sqrt{2}}{4} \end{pmatrix}$$

Of the four rows i_1,i_2,i_3 and i_4 the coordinates' values of the two factorial axes Δ_1 and Δ_2 are as follows:

$$F_1(i_1) = \frac{\sqrt{6}}{2}, F_1(i_2) = -\frac{\sqrt{6}}{4}, F_1(i_3) = \frac{\sqrt{6}}{4}, F_1(i_4) = -\frac{\sqrt{6}}{2}$$

$$F_2(i_1) = \frac{\sqrt{2}}{2}, F_2(i_2) = -\frac{\sqrt{2}}{4}, F_2(i_3) = -\frac{\sqrt{2}}{4}, F_2(i_4) = \frac{\sqrt{2}}{2}$$

While the corresponding masses of rows are equal to

$$f_1 = \frac{1}{6}, f_2 = \frac{2}{6}, f_3 = \frac{2}{6}, f_4 = \frac{1}{6}$$

The fact that the two factorial axes Δ_1 and Δ_2 are vertical will be found through the verification of relationship 12, which is a formulation of the Pythagorean Theorem at plane.

E.g. for row i_2 we have

$$F_1(i_2) = -\frac{\sqrt{6}}{4} = -0,612, F_2(i_2) = -\frac{\sqrt{2}}{4} = -0,354$$

Using relationship 13 we have

$$d^2(i_2, f_J) = \frac{1}{\frac{2}{6}}\left(\frac{1}{2} - \frac{2}{6}\right)^2 + \frac{1}{\frac{2}{6}}\left(0 - \frac{2}{6}\right)^2 + \frac{1}{\frac{2}{6}}\left(\frac{1}{2} - \frac{2}{6}\right)^2 = 3 \cdot \frac{1}{36} + 3 \cdot \frac{4}{36} + 3 \cdot \frac{1}{36} = \frac{18}{36} = \frac{1}{2}$$

Thus $$d^2(i_2, f_J) = \frac{1}{2} \rightarrow \quad d(i_2, f_J) = 0,707$$

Using the values of $F_1(i_2)$ and $F_2(i_2)$ of i_2 row in factorial axes Δ_1 and Δ_2 we have

$$F_1^2(i_2) + F_2^2(i_2) = \left(-\frac{\sqrt{6}}{4}\right)^2 + \left(-\frac{\sqrt{2}}{4}\right)^2 = \frac{6}{16} + \frac{2}{16} = \frac{1}{2}$$

Therefore relationship 12 is verified

$$d^2(i_2, f_J) = F_1^2(i_2) + F_2^2(i_2) \qquad (17)$$

Schematically the factorial plane 1x2 is as follows.

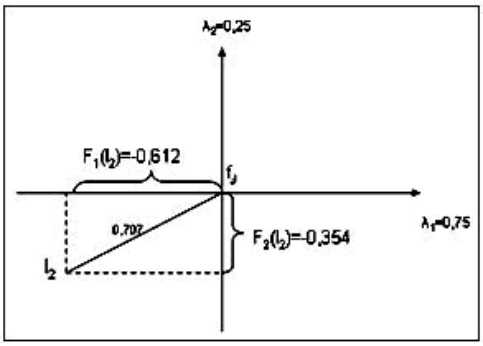

Figure 2: Verification of the relationship $d^2(i_2, f_J) = F_1^2(i_2) + F_2^2(i_2)$

Apparently relationship 17 applies to the multidimensional space R^{p-1}

$$d^2(i_k, f_J) = F_1^2(i_k) + F_2^2(i_k) + ... + F_s^2(i_k) \text{ where k=1….n and s=1…..p-1} \qquad (18)$$

Relationship 18 is Pythagorean Theorem's expression at the multidimensional space R^{p-1} **As far as the distance between a row's profile f_J^i and a variable's profile f_I^j is concerned, the following relationship applies.**

$$d^2\left(f_J^{\,i},\ f_I^{\,j}\right) = \sum_{p=1}^{p-1}\left[F_p(i) - G_p(j)\right]^2 \qquad (19)$$

With the example's data from row i_1 and column j_1 we have

$$d^2(i_1, j_1) = [F_1(i_1) - G_1(j_1)]^2 + [F_2(i_1) - G_2(j_1)]^2 = \left[\frac{\sqrt{6}}{2} - 0\right]^2 + \left[\frac{\sqrt{2}}{2} - (-\frac{\sqrt{2}}{2})\right]^2 = 3,5$$

Thus $$d(i_1, j_1) = \sqrt{3,5} = 1,871$$

Connection among the «rows» and «columns» of a two-dimensional data table

Let in a data table T(n,p) n rows correspond to n respondents, while in p rows of the table the values of p questions correspond to p variables. A detailed reference to the query of identifying a respondent with the variable mostly associated will be made, using a particular example of six qualitative variables (for the measurement of which a 5-point Likert scale is used, where 5 concerned excellent impression), in which 99 persons responded, constituting one of the five classes created through the ascending hierarchical classification application with the VACOR procedure, in 1721 foreign visitors of Thessaloniki.

Data are included in the research conducted within the framework of ARCHIMIDES III programme entitled «Data Analysis and Knowledge Management Technologies in tourism products' design» with coordinator Professor Dr. Dimitrios Karapistolis.

The questionnaire included six questions relating to foreign visitors' rating of the following: a) the sights of Thessaloniki b) Greek cuisine c) the city's nightlife d) architectural style e) safety and f) locals' friendliness.

The six variables are represented respectively as follows: Δ4, Δ5, Δ6, Δ7, Δ8, Δ9. Given the classification of the 99 persons in one of the groups created by the classification of the 1721 persons using the VACOR method.

Table 4 presents a part of their answers.

Table 4: Part of the data table

A/A	Tags	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
1	**11**	5	3	3	5	5	5
2	**20**	2	2	1	2	3	3
3	**60**	4	2	3	5	3	5
.	.	.	.	.	.	.	.
21	**315**	4	3	4	5	3	5
.	.	.	.	.	.	.	.
67	**1086**	3	2	1	5	3	5
.	.	.	.	.	.	.	.
94	**1623**	4	0	3	5	4	3
.	.	.	.	.	.	.	.
99	**1712**	4	2	1	5	3	2

Given the statistical test of ratios' difference (relationships 1 and 2) at 5% significance level, the connection of each class or respondent with one or more variables is defined.

Having applied the Ascending Hierarchical Classification with the VACOR method to table 4 data, the particular typology was created with the following five homogeneous clusters: 180, 186, 191, 192 and 193, in which when the abovementioned hypothesis testing is used Table 5 is found, where cluster 186 seems to connect to two variables, Δ5 and Δ6, while cluster 192 is more connected to variables Δ5, Δ7 and Δ9, with greater intensity, though, towards Δ5 ($Z_{\Delta 5}$=6.3422).

Table 5: Presents the connection among the 5 clusters and the six variables

Tags	**Δ4**	**Δ5**	**Δ6**	**Δ7**	**Δ8**	**Δ9**
180	-0.4215	-14.422	13.699	0.9611	2.244	-2.414
186	0.9345	1.8842	5.5692	-1.3116	-3.14	-2.392
191	12.4939	-15.39	-15.873	11.057	-2.73	3.528
192	-4.0938	6.3422	-6.8836	1.9318	0.147	2.51
193	-0.2294	0.1392	-1.2727	-1.6656	2.255	0.518

The application of the test in the six variables' values for each respondent results in Table 6, from which it is evident that respondent 20 is more connected to variables Δ5, Δ8 and Δ9 with greater intensity for Δ5 ($Z_{\Delta 5}$=7.1634), while respondent 315 seems to be connected to three variables, Δ5, Δ6 and Δ7 with a greater intensity for Δ6 ($Z_{\Delta 6}$=7.9739).

Table 6: Presents the connection between respondents and variables

Tags	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
11	-0.5345	1.5176	0.6310	0.0817	0.5959	-1.7833
20	-4.8848	7.1634	-4.8654	-4.3151	5.0210	2.4658
60	-1.7209	-2.0604	3.6369	4.0674	-5.8574	2.0785
.	.	.	.	.	.	.
315	-3.4385	2.9401	7.9739	1.9090	-7.1596	-0.0129
.	.	.	.	.	.	.
1623	1.5220	-15.3903	6.7158	8.1673	2.6932	-5.6008
.	.	.	.	.	.	.
1712	4.3130	1.8549	-7.4562	11.7076	-1.2364	-10.0491

With the procedure, of course, there are also many respondents, who are not connected to any variable, such as respondent 11, because the value of the z distribution in included between values $-1.645<z<1.645$, i.e. it is due to random factors, since H_0 applies.

Therefore, from Table 6, we consider, as already mentioned, that the greater z value corresponding to a variable from the row of the 6 values, determines the respondent's connection to that particular variable.

Using as an alternative approach to the problem the connection of respondents with variables, it is appropriate to apply in table 4 data the Correspondence Analysis, and then, according to the respondents' coordinates and the variables of p-1 factorial axes, to determine through the use of Euclidean metrics, the respondents that are mainly connected to particular variables.

In any case, the extraction of results through the use of the Correspondence Analysis presupposes taking into account the explanatory power of the factorial axes.

For this there are three explanatory instances, depending on the explanatory power of the factorial axes; i.e. the use of two or three factorial axes or to use all factorial axes specified by the whole set of variables.

1. With two factorial axes

The analysis of table 4 data through the Correspondence Analysis the following are found:

A) Display of characteristic eigenvalues

Table 7: Histogram of characteristic eigenvalues

Total inertia 0.09244				
Axis	**Inertia**	**%Interpretation**	**Sum**	**|Histogram Eigenvalues**
1	0.0305713	33.07	33.07	|**********
2	0.0285268	30.86	63.93	|*******
3	0.0149623	16.19	80.11	|****
4	0.0097252	10.52	90.63	|***
5	0.0085587	9.37	100.00	|*

From Table 7 it is evident that with the first two factorial axes 63.93% of all information is derived, coming from table 4 data.

Table 8: Coordinates of respondents and variables for the whole set of factorial axes

Respondents' Coordinates

Tags	FA1	FA2	FA3	FA4	FA5
11	34	12	-3	-15	-34
20	182	-122	-111	46	12
60	-49	51	80	-107	90
73	-3	-111	-135	-103	-238
82	162	73	-101	-60	-33
88	-81	-35	-182	-103	-108
99	-7	104	92	-48	24

Variables' Coordinates

Tags	GA1	GA2	GA3	GA4	GA5
Δ4	-153	-29	136	114	-78
Δ5	441	-165	97	-34	-63
Δ6	91	451	70	-34	37
Δ7	-152	-65	-1	-177	-29
Δ8	27	29	-229	57	-64
Δ9	-10	-97	-8	38	169

Then we observe factorial plane 1x2

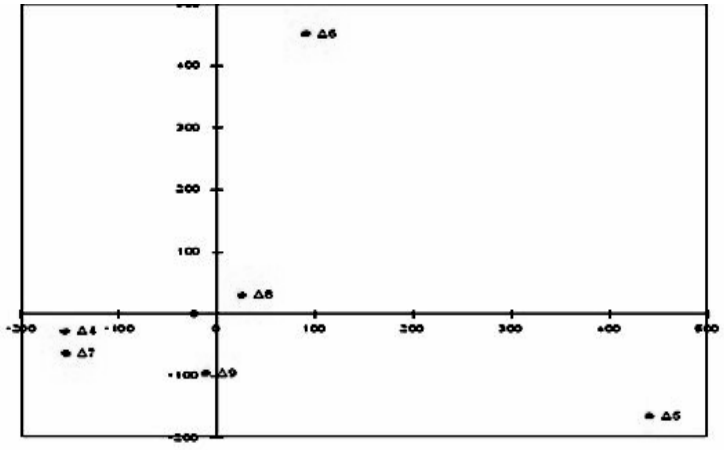

Figure 3: Variables' factorial plane 1x2

Based on the first two factorial axes, plane 1x2 is divided into four subspaces. In the 1st subspace points with both coordinates positive are placed. Following the same procedure, depending on the signs of the points' coordinates they are placed in the 2nd, 3rd, or 4th subspace.

Therefore variables Δ6 and Δ8 are located in the 1st subspace, no variable is located in the 2nd subspace, and variables Δ4, Δ7 and Δ9 are located in the 3rd subspace, while variable Δ5 is located in the 4th subspace.

The profiles of the 99 respondents are also placed on factorial plane 1x2, depending on their coordinates' signs, where one group of «subjects» in the 2nd subspace is «orphaned» from a variable. (Points' selection was based on COR and CTR criteria).

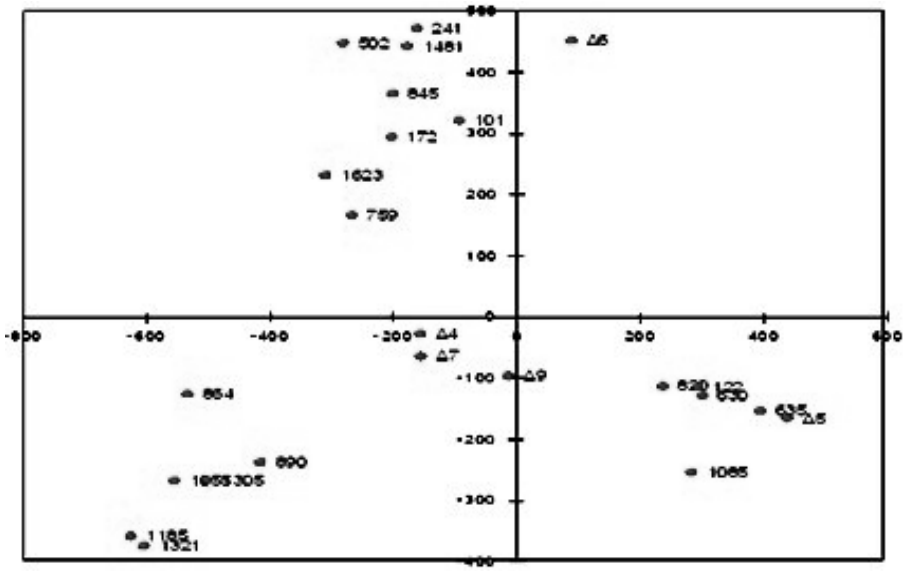

Figure 4: Factorial plane 1x2 of «objects» and variables

Since the first two factorial axes create an orthogonal coordinate system, to measure the distance between a variable and a «subject» Euclidian metric between two points is used, through the use of their coordinates on the two axes. Using the MAD software, the smallest distance of each «subject» among the six variables is calculated, resulting, thus to the following finding.

Table 9 shows that 7 respondents {122,238,420,630,635,820 and 1085} are connected to variable Δ5, since they have the smallest distance from the other five variables, while 8 respondents are connected to variable Δ4.

Table 9: Connection between «subjects» and variables based on two factors

Tags	A4	A5	A6	A7	A8	A9
Number	8	7	5	14	45	20
1	88	122	101	358	11	20
2	172	238	241	389	60	73
3	554	420	502	629	82	116
4	644	630	845	645	99	209
5	693	635	1481	864	176	213
6	759	820		890	246	378
7	1548	1085		1006	298	596
8	1623			1055	301	641
9				1185	312	643
10				1279	315	678
11				1305	355	703
12				1321	368	704
13				1433	399	705
14				1712	452	879
15					536	882
16					553	1086
17					555	1105
18					567	1328
19					702	1503
20					745	1696
21					914	
22					922	
23					990	
24					1000	
25					1019	
26					1020	
27					1114	
28					1127	
29					1140	
30					1156	
31					1172	
32					1200	
33					1269	
34					1307	
35					1383	
36					1482	
37					1530	
38					1540	
39					1546	
40					1570	
41					1580	
42					1619	
43					1626	
44					1633	
45					1694	

2) With three factorial axes

If the first three axes are used and given that they constitute a three-orthogonal coordinates system in the three-dimensional space, then eight subspaces are created, in which they are placed, depending on the points coordinates' signs.

Therefore, if we symbolise points located in the 1st, 2nd, 3rd, and 4th subspace with the symbol ↑next to the point's identity, while for points located in the 5th, 6th, 7th, and 8th subspace we set the symbol↓, we have for the first time in world literature, an illustration of three-dimensional space at plane, (without using the perspective method), therefore these diagrams hereinafter will be called **Karapistolis diagrams**.

The implementation of this particular procedure results firstly in Table 10 through the placement of variables in the 8 subspaces, and secondly in Table 11 through the placement of the 99 respondents to the corresponding subspaces, as well as the corresponding Karapistolis diagram.

Table 10: The eight subspaces with the variables located in them

Subspace	1rst	2nd	3rd	4th	5th	6th	7th	8th
Number	1	0	1	1	1	0	2	0
	Δ6		Δ4	Δ5	Δ8		Δ7	
							Δ9	

Table 11: The eight subspaces with the respondents located in

Subspace	1rst	2nd	3rd	4th	5th	6th	7th	8th
Number	23	6	10	9	11	10	15	15
	298	60	358	116	11	101	73	20
	312	99	645	209	82	172	88	122
	315	301	879	630	246	241	213	176
	399	502	1055	635	355	554	389	238
	567	536	1105	643	452	759	596	368
	702	1546	1185	703	553	845	629	378
	745		1321	704	555	1383	644	420
	914		1328	705	1140	1481	693	641
	922		1433	882	1172	1548	864	678
	1019		1712		1307	1623	890	820
	1020				1694		1006	990
	1114						1086	1000
	1127						1279	1085
	1156						1305	1482
	1200						1503	1696
	1269							
	1530							
	1540							
	1570							
	1580							
	1619							
	1626							
	1633							

Below is the diagram

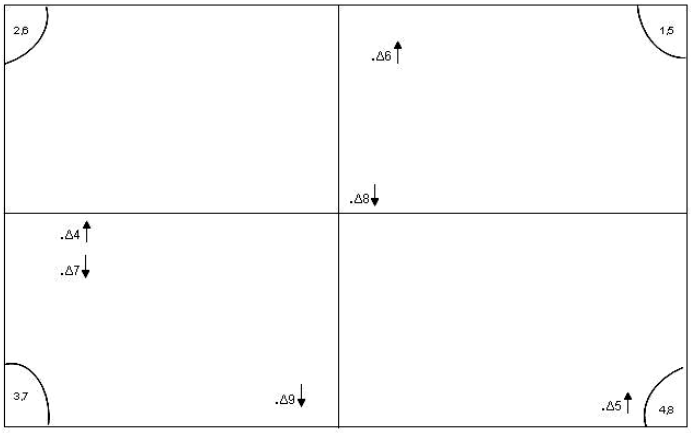

Figure 5: Variables' factorial space 1x2x3

It is evident from Karapistolis diagram that variables Δ4 and Δ7, as well as variables Δ6 and Δ8 are located in different subspaces in the three dimensional space, with all that this might mean for their interpretation - at a rate of 80.11%, versus their interpretation provided by factorial plane 1x2 at a rate of 63.93%.

According to table 11, based on the signs of the points' three coordinates, their placement at the eight subspaces, variable Δ5 and the 9 respondents belonging to the 4th subspace had to be interconnected through the use of the three factors, however,

when calculating minimum distances among respondents and variables, Table 12 occurs, which informs us that only 4 respondents (122,630,635 and 1085) are connected to variable Δ5. The interesting is that among the nine respondents of the 4th subspace, together with variable Δ5 (Table 11), only two - 630 and 635 – are actually connected to variable Δ5, since they are at smaller distance, while the other 7 are connected to other variables, such as respondent 116 who is connected to variable Δ9 located in the 7th subspace, due to the smaller distance as compared to that from variable Δ5.

Table 12: Respondents' connection to the variables according to three factors

Tags	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9	
Number	15	4	6	16	11	47	
1	60	122	101	389	82	11	705
2	99	630	241	554	88	20	745
3	301	635	502	629	172	73	820
4	358	1085	845	644	452	116	879
5	536		1269	693	553	176	882
6	645		1481	759	555	209	914
7	702			864	990	213	922
8	1019			890	1000	238	1086
9	1020			1006	1140	246	1114
10	1105			1055	1503	298	1127
11	1156			1279	1694	312	1172
12	1185			1305		315	1200
13	1321			1383		355	1307
14	1433			1548		368	1328
15	1546			1623		378	1482
16				1712		399	1530
17						420	1540
18						567	1570
19						596	1580
20						641	1619
21						643	1626
22						678	1633
23						703	1696
24						704	

This different connection of respondents with variable Δ5 is due to a percentage of 16.19% of the information provided by the 3rd factorial axis.

3. With the whole set of factorial axes

Using all five factorial axes an orthonormal base is created at R^5, where all variables and all respondents are placed in their real positions, from where all information of the data table is provided. The connection of the respondents based on the minimum distance among variables and "subjects" using all coordinates is presented in table 13, as opposed to their connections using the z distribution.

It is evident from table 13 that 7 respondents {122,238,420,630, 635, 820,1085} are connected to variable Δ5,(as in the two factors case), while when it comes to respondents connected to variable Δ4, it can be seen that increasing the percentage of information from 63.11% to 100%, the number and the respondents connected to

the variable are also differentiated, an observation which is also true for some other variables.

Therefore, 8 respondents are connected to variable Δ4 with the first two actors, 15 respondents are connected to three factors and 16 are connected to all actors. It is also worth noting that none of the 8 respondents that seems to be connected to variable Δ4, according to the two factors, does not seem to be connected to the 15 respondents when the three factors are used, while when the whole set of factors is used, only 10 of the 15 continue to be connected to variable Δ4.

This differentiation in terms of the number of respondents and the respondents that are connected to each variable, depending on the number of factors taken into account, means that the only correct procedure to identify the connection between respondents and variables in a data table, is to use all factors extracted from the Correspondence Analysis, since in this case an n-dimensional orthonormal space is created, which illustrates the actual picture of relationships among the points depicting the respondents and the points depicting the variables, offering 100% of information, independently of the participation of each «subject» in factorial axes formation.

The connection of subjects with variables based on Euclidian metric and maximum z distribution value are presented in the following table.

Table 13: Connection of subjects and variables with the two different procedures (Euclidian metric and z distribution)

Connecting to the Euclidean metric								Connecting to the metric Z							
VARIABLE	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9		VARIABLE	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
NUMBER	16	7	9	20	26	21		NUMBER	6	34	19	11	21	3
1	99	122	101	60	11	116		1	536	11	99	60	73	213
2	301	238	172	358	20	209		2	645	20	241	358	82	1433
3	536	420	241	389	73	213		3	1055	209	301	629	88	1696
4	645	630	502	567	82	298		4	1105	238	315	879	368	
5	702	635	845	629	88	312		5	1185	246	399	1086	378	
6	759	820	1269	644	176	315		6	1548	298	502	1279	389	
7	1019	1085	1481	693	246	378		7		312	845	1305	452	
8	1020		1570	864	355	596		8		355	1114	1321	554	
9	1055		1633	879	368	641		9		420	1156	1328	644	
10	1105			890	399	643		10		553	1172	1623	693	
11	1127			914	452	678		11		555	1269	1712	759	
12	1185			1006	553	703		12		567	1481		864	
13	1200			1086	554	704		13		596	1530		890	
14	1546			1279	555	705		14		630	1540		990	
15	1548			1305	990	745		15		635	1546		1000	
16	1626			1321	1000	882		16		641	1570		1006	
17				1328	1140	922		17		643	1580		1140	
18				1383	1172	1114		18		678	1619		1383	
19				1623	1307	1156		19		702	1633		1482	
20				1712	1482	1433		20		703			1503	
21					1503	1696		21		704			1694	
22					1530			22		705				
23					1540			23		745				
24					1580			24		820				
25					1619			25		882				
26					1694			26		914				
								27		922				
								28		1019				
								29		1020				
								30		1085				
								31		1127				
								32		1200				
								33		1307				
								34		1626				

The ultimate purpose of this project is to prove the most effective solution for finding the connection among variables and subjects, having used the z distribution and the Euclidean metric. This purpose can be implemented as long as data deriving from table 13 are trained using machine learning classifiers, the SVM learning machine, in particular.

Overview of data training using machine learning classifiers

Machine learning is a field of artificial intelligence which concerns algorithms and methods allowing computers to «learn». The aim of machine learning is to create models using a dataset, through the use of a computer system.

Various techniques of machine learning have been developed, which are used depending on the nature of the problem and fall within one of the following two types:

1. Supervised learning

2. Unsupervised learning

In supervised learning the system is requested to «learn» a concept or function from a data set, which constitutes the description of a model.

In unsupervised learning the system should find out on its own correlations or groups in a data set, creating prototypes, without knowing whether they exist, how many and which they are.

In the present project supervised learning will be used, where the system should «learn» inductively a function called **target function** and constitutes an expression of the model describing the data.

The target function is used to predict the value of a variable, which is called **output variable,** based on the values of a set of variables, which are called **input variables** or **characteristics.**

In supervised learning two types of problems (learning tasks) are identified, classification problems and regression problems.

Classification relates to the creation of prediction models for discrete ranges (classes/categories).

Regression relates to the creation of prediction models for numerical values.

Support Vector Machines (SVM)

Support Vector Machines (SVMs) are characterised as learning machines and they are based on Statistical Learning Theory and on Perceptron-type neural networks. They were proposed by Vladimir Vapnik.

In the case of classification, SVMs try to find a hypersurface, which separates in the space of examples the negative from the positive examples. SVMs are characterised by the following stages:

1. Training: In this phase parameters' calculations of the learning model are performed using the appropriate learning data set.

2. Test: The calculated parameters model (support vectors) is tested in terms of its ability in achieving successful result estimation in a non-trained data set.

3. Performance estimation: The appropriate performance indicators of the model are calculated, mainly the error rate, aiming at the investigation of the model's generalisability.

Support Vector Machines (SVMs) belong to the Supervised Machine Learning algorithms with remarkable success in classification problems. Similar to most machine learning algorithms, they represent objects to be classified as feature vectors.

In our case, the respondents are the objects to be classified and features provide information, such as whether the respondent to be classified is connected to variable A or B.

To use SVMs in classification problems with more than two classes two categories of approaches have been proposed:

- Direct: Finding the differentiating hypersurfaces in a step (Vapnik, 1998; Crammer and Singer, 2000)

- Indirect: Combination of the results of a set of binary SVMs: one-versus-one, one-versus-all (Vapnik, 1998)

Indirect approaches are simpler and easier to implement, but none of the approaches returns probabilities.

Implementation of the SVM learning machine

To implement the proposed comparison six qualitative variables Δ4 to Δ9 are used again, the values of which concern the answers given by the 99 foreign visitors of Thessaloniki.

Table 14 presents the connection of objects with the corresponding variables, firstly according to the minimum distance using Euclidean metric, and secondly with the maximum value of the z distribution.

Table 14: Connection of the 99 objects and the 6 variables according to Euclidean metric (MET1) and maximum value of the z distribution (MET2)

	Δ4	.....	Δ9	MET1
11	5		5	5
20	2		3	5
60	4		5	4
73	3		2	5
82	3		4	5
88	2		2	5
99	4		4	1
101	5		5	3
116	3		5	6
122	3		5	2
172	4		5	3

	Δ4	.....	Δ9	MET2
11	5		5	2
20	2		3	2
60	4		5	4
73	3		2	5
82	3		4	5
88	2		2	5
99	4		4	3
101	5		5	3
116	3		5	2
122	3		5	2
172	4		5	3

Note 1: The values of variables MET1 and MET2 from 1 to 6 correspond to variables Δ4 to Δ9. The coincidence between MET1 and MET2 values reaches 49.5%

Data training with the Support Vector Machine SVM

Table 13 data training using the Support Vector Machine (SVM) it is found that the process of objects and variables connection through Euclidean metric is superior to the corresponding one using the z distribution. This finding arises since learning performance rate of table 13 data, relating to the connection of objects through Euclidean metric is higher (78.89% Table 15) than the one resulting from the z distribution (71.11% Table 16).

Furthermore, through the use of Euclidean metric, after 20 data learning repetitions percentages above average are much higher (7 out of 20 repetitions above 80% with a maximum value of 100% and a minimum value of 61.11%) as compared to the corresponding percentages resulting from the z distribution, presenting a maximum value of 88.89% only and a minimum value of 44.44%.

Table 15: SVM training based on the Euclidean metric

MET_EU (20%)	
0,8333	
0,7778	
0,7778	
0,7778	
0,8333	
0,8889	
0,7222	
0,7778	78,89%
0,9444	
0,7778	
1,0000	
0,7778	
0,6111	
0,7222	
0,7222	
0,8889	
0,7222	
0,8333	
0,6667	
0,7222	

Table 16: SVM training based on the z distribution

MET_Z (20%)	
0,7778	
0,6667	
0,7778	
0,6667	
0,7778	
0,6667	
0,4444	
0,5556	71,11%
0,6111	
0,8333	
0,7778	
0,6667	
0,6667	
0,7778	
0,7222	
0,7778	
0,7778	
0,6667	
0,8889	
0,7222	

The new procedure for the classification of data table rows. The KARAP method

The proposed classification procedure answers the concern existing in each Ascending Hierarchical Classification through the VACOR method, i.e. that it is not possible to accurately identify objects connected with classes variables.

For this reason, the proposed new classification procedure of n objects of a data table T(n,p), achieves to the extent that the researcher wishes, homogeneity of

objects in terms of their exclusive connection to each variable or if desired in terms of a combination of variables

The proposed classification's algorithm, named **KARAP**, which is implemented through the MAD program is as follows:

1. Logical table 0-1 is created, derived from data table T(n,p),either using variables' ratings, or Likert scales.Each object's numbering corresponds from 1 to n.

2. Logical Table 0-1 is analysed using the Correspondence Analysis to export the variables and objects' coordinates on the factorial axes.

3. Using the F_a and G_a coordinates it is possible to find the connection of each object with each variable based on Euclidean metric.

If desired by the researcher, s/he can follow the steps below to classify objects in terms of a combination of variables, due to their great number.

4. Table Burt is created, corresponding to logical table 0-1

5. Ascending Hierarchical Classification is applied using the VACOR method on the Burt table's data

6. Based on hierarchy's typology, which results from the segmentation of the dendrogram into k clusters, objects are classified depending on the variables with which they are connected, according to step 3.

The implementation of the proposed classification according to steps 1 to 3, the following results were found using the 84 data of table 4 (all visitors not answering to a criterion were intentionally not included).

Table 17 clearly shows the perceptions of all 84 respondents of this particular class regarding Thessaloniki in terms of the six criteria used in this study.

For example they did not like or they were indifferent towards Greek cuisine (Δ_{51}, Δ_{52} and Δ_{53} percentage 13/84=15.48%), while they endorsed locals friendliness

(Δ_{94} and Δ_{95} 21/84=25%), while they were also absolutely negative towards the city's nightlife, at a percentage of 17.86%.

Table 17: Classification of the 84 respondents based on the karap method

	The Sights of Thessaloniki					Greek Cuisine					The city's nightlife				
Tags	**Δ41**	**Δ42**	**Δ43**	**Δ44**	**Δ45**	**Δ51**	**Δ52**	**Δ53**	**Δ54**	**Δ55**	**Δ61**	**Δ62**	**Δ63**	**Δ64**	**Δ65**
Number	**0**	**2**	**4**	**0**	**5**	**3**	**2**	**8**	**0**	**0**	**6**	**4**	**5**	**0**	**0**
		20	420		554	101	879	11			73	122	567		
2		1140	630		643	629	1696	246			88	635	702		
3			820		705	1503		298			358	990	914		
4			1085		1114			315			1279	1000	1020		
5					1626			355			1433		1546		
6								399			1712				
7								922							
8								1482							
9															
10															
11															
12															
13															
14															
15															
	Architectural style					**Safety**					**Locals' Friendliness**				
Tags	**Δ71**	**Δ72**	**Δ73**	**Δ74**	**Δ75**	**Δ81**	**Δ82**	**Δ83**	**Δ84**	**Δ85**	**Δ91**	**Δ92**	**Δ93**	**Δ94**	**Δ95**
Number	**2**	**0**	**0**	**10**	**6**	**0**	**1**	**1**	**1**	**3**	**0**	**0**	**0**	**6**	**15**
1	452			82	60		645	1156	238	553				99	116
2	1105			301	312					555				536	176
3				368	678					693				1019	209
4				1269	745									1127	213
5				1530	1307									1172	378
6				1540	1383									1200	596
7				1570											641
8				1580											644
9				1619											703
10				1633											704
11															882
12															1006
13															1086
14															1548
15															1694

Conclusion

1. With the Karapistolis diagrams it is possible to graphically illustrate three dimensional space at plane, so that the researcher does not have to perform a classification to avoid confusion of adjacent points, belonging in different R^3 subspaces.

2. With the SVM learning machine it is possible to objectively evaluate each classification type, no matter which metric was used to create it, aiding the researcher in comparing results between two different classification methods, as well as in identifying homogeneity of each classification's classes.

3. It is deemed necessary, every time the connection of objects to specific variables is to be investigated, firstly to use the whole set of actors resulting from the Correspondence Analysis and secondly to use the proposed classification procedure, named karap, since the researcher can identify the uniqueness of the

connection of objects within the classes in each rating of variables, thus interpreting more easily the behaviour of the whole set of objects.

4. The karap method provides to the researcher compact classes, in terms of objects' behaviour uniqueness for each cluster of variables, since it only includes objects, whose profile is connected to specific variables of each cluster.