International Research Journal of Finance and Economics
ISSN 1450-2887 Issue 35 (2010)
© EuroJournals Publishing, Inc. 2010
http://www.eurojournals.com/finance.htm

Greek Census 1981. A Case of Factorial Correspondence Analysis

Paraschos Maniatis
Department of Business Administration at Athens University of Economics and Business
76 Patission St., Athens, GR-104 34
E-mail**:** pman@sch.gr

Abstract

Factorial Correspondence Analysis is not so popular, mainly because of its mathematical complications and the difficulties in interpreting the results. Another reason might be that it is rather remote from the classical body of inferential Statistics, which sometimes ignores the parametric techniques and the significance tests.

The purpose of this study is to present the technique in a simple, comprehensive manner; its mathematical foundations, its application field as well as and the way of interpreting the results. To fix the ideas we apply the analysis to a contingency table, matching population ages and national districts, as resulted from the Greek Census 1981.

Keywords: Factorial Correspondence Analysis, Greek Census, Multidimensional
Jel Classification Codes: H00 - General

Introduction

The term Factorial Correspondence Analysis designates the nature of the technique: Correspondence, because it treats contingency tables and factorial, because its results can be visualised in a factorial plane, defined by two factorial axes. As such, it belongs to the great family of multidimensional descriptive statistics. A close term to the term 'multidimensional', is the term 'multivariate'. However, the two terms should not be confused: the latter refers to table analyses using methods of inferential statistics, while the former is interested in the Geometry of the data, considered as vectors (points) in a multidimensional vector space without worrying too much for problems posed by the parametric distribution of the variables. Visualization of the data is the soul of the multidimensional statistics.

The technique, in its primitive form, known from the beginnings of the 20th century, originated in the works of the great Karl Pearson (1901). There followed advances by other great names of economic and mathematical statistics such as Hotteling (1933), of social statistics, Burt (1950) and many others (Benzıcri, 1982). The theoretical aspects of these advancements were excellent, but useless. They all faced immense computational problems, in particular the problem of diagonalisation of a square matrix of higher dimensions, common to all themes of multidimensional techniques. It was the invasion of the computer in the statistical laboratories, in the late sixties, which gave to the multidimensional techniques strength and development. The first to give correspondence analysis its present form was, no doubt, J.P. Benzıcri (1970, 1973), by introducing the distribution $\chi 2$ as distance between points in the vector space. Further developments of correspondence analysis introduced notions of information theory, proving the equivalence between χ^2 and Information (Entropy), as developed by Shannon (1948) and Khinchin (1957). Extensions of correspondence analysis for more

than two variables are the log-linear techniques, which however, deviate from the spirit of the multidimensional techniques.

Presentation of the Technique

In this section an introduction of the mathematical basis of correspondence analysis is presented. For reasons of simplicity we start with the principal components analysis, which is the model for correspondence analysis.

1. The Principal Components Analysis

Consider a matrix X_{np}, which represents n measurements (cases) on p statistical variables (characters). The general element of the matrix is x_{ij}, i=1, 2….n; j=1, 2,…p. If we subtract from each variable its mean m_j (mean correction) and divide by the standard deviation s_j, we obtain the matrix R_{np} with general element $r_{ij}=(x_{ij}-m_j)/s_j$, the matrix of standardized variables.

The n rows of R can be considered as n points (row vectors) in the $\boldsymbol{R^p}$ Euclidean vector space. If one could see these points, he could also see the p components of each case and hence perceive their distribution and their correlations, measured by (the cosine of) the angles between the vectors. The same convenience would be available for the p vectors in $\boldsymbol{R^n}$ representing the n components of each character), hence the distribution and the correlations between the characters. This being impossible for vector spaces with dimension >3, a good approximation would be to see the projections of the points on a plane, hoping that the loss of information is not important. This is the task of the Principal Components Analysis (PCA), the general principles of which are as follows:

One searches for a vector u of unit norm, $\left|u\right|=1$, such as the sums of the squared projections of the p variables onto u to be maximum. Then, one searches a unit vector v perpendicular to u ($u'v=0$, (') meaning transpose of a vector or matrix) with the same property. In this way one obtains the "trace" of the cluster of points in R^n onto the Euclidean plane R^2. This projection can be reliable or not, depending on the "shape" of the cluster. In the following analysis all vectors are considered as column vectors.

The projection of the vector r_i onto u is

$Proj(r_i,u)=r_i'u/\left|u\right|=r_i'u \Rightarrow [Proj(r_i,u)]^2=(r_i'u)'(r_i'u)=u'(r_ir')u$, since $r_i'u$ is a scalar.

Summing over all i one obtains

$S=\sum[Proj(r_i,u)]^2=\sum u'(r_ir_i')u=u'\sum(r_ir_i')u=u'RR'u=u'Vu$, where $V=\sum(r_ir')=R'R$ is the correlation matrix of Pearson's coefficients of correlation ρ. One now seeks to maximize S under the constraint u'u=1. For this purpose one considers the Lagrangian of S $L=u'Vu-\lambda(u'u-1) \Rightarrow \partial L/\partial u=2Vu-2\lambda u=0 \Rightarrow$

$$Vu=\lambda u \qquad \text{(II.1.1)}$$

The last equation shows that the wanted vector u is an eigenvector of the correlations matrix V, associated to the eigenvalue λ. Pre-multiplication of the equation with u' gives $u'Vu=\lambda$. But since the purpose was to maximize S=u'Vu, it results that

$$\max\{u'Vu\}=\lambda_{max} \qquad \text{(II.1.2)}$$

Therefore, the vector u is the eigenvector matched to the maximum eigenvalue. Further, since V is associated to the positive semi-definite quadratic form u'Vu, it results that all eigenvalues of V are non-negative.

Repeating this exercise for a second vector v, perpendicular to u, one obtains v as the eigenvector of V, which is associated to the second largest eigenvalue of V.

What is the meaning of the eigenvalues? If U is the matrix of the eigenvectors (which is orthogonal) and Λ is the diagonal matrix of the associated eigenvalues, written in decreasing order, then it holds that $VU=U\Lambda \Rightarrow$

$$U'VU=U'U\Lambda=\Lambda \qquad \text{(II.1.3)}$$

since U is orthogonal, hence, U'U=UU'=I. Denoting the trace of a square matrix by tr one has

$tr(U'VU)=tr(\Lambda)$ (II.1.4)

But $tr(U'V'U)=tr(VU'U)=tr(V)$ and $tr(\Lambda)=\lambda_1+\lambda_2+.....+\lambda_p$, and due to (4) one obtains

$tr(V)= \lambda_1+\lambda_2+.....+\lambda_p$ (II.1.5)

But the trace of a correlation matrix is the sum of the variances of the (standardized) variables (equal to p, for obvious reasons), which is by definition, the total variation in the data. Therefore, the eigenvalues explain the total variance in the data. The above conclusion can be better understood if one considers the variance-covariance matrix instead of the correlations matrix V.

Summarizing the above analysis, the PCA proceeds as follows:

- Diagonalization of the correlation (covariance matrix in the case of non-standardized variables)
- Projection of the cluster of the point in $\boldsymbol{R^n}$ onto axes, which are the eigenvectors of the diagonalized matrix with the largest eigenvalues.

Now, one has a criterion for the quality of the projection, i.e. for the loss of information due to the reduction of the dimension of the data: the larger the percentage $\lambda_1/\Sigma\lambda_i$, $\lambda_2/\Sigma\lambda_i$; i=1, 2,…p the more the variation explained by the two first factorial axes-eigenvectors, the less the loss of information, the better the application of the PCA.

As far as the terminology is concerned, there is strong tradition in the multidimensional statistical analysis to call factorial axes the eigenvectors, factorial plane the plane R^2, inertia the total variation and the fraction $\lambda_i/\Sigma\lambda_i$ relative inertia.

2. The Correspondence Analysis

Correspondence analysis (CORRA) is a method appropriate for the analysis of frequency (contingency) tables of large dimensions.

Let X_{kp} be the contingency table of two qualitative variable Y and Z with k and p levels accordingly. The general element of X is x_{ij}.

Let $n=\Sigma\Sigma x_{ij}$, i=1, 2,…k ; j=1, 2,…p. Dividing all elements of X by n, one obtains the table of relative frequencies P with general element p_{ij}. The marginal sum of row i is denoted by $p_{i.}$ and that of the column j by $p_{.j}$

Therefore, the ratios $p_{ij}/p_{i.}$ are the relative frequencies of the k levels of the variable Z *within* the row i ($\Sigma p_{ij}/p_{i.}$ =1 j=1, 2…k).

Call the ratios $p_{ij}/p_{i.,}$ profile of the row i. The purpose of CORRA is to investigate the similarities (similarity of distribution) between the profiles of the rows and those between of the columns and to identify which elements of X cause strong dissimilarities between rows, between columns, and between rows and columns. The CORRA transforms, for reasons which will be explained later, the initial table X as follows:

Let $R=[r_{ij}]$ $r_{ij}=p_{ij}/(p_i\sqrt{p_{.j}}]$ (II.2.1)

The expectation of the stochastic variable r_j (column j) is

$m_j=\Sigma p_i[p_{ij}/(p_i\sqrt{p_{.j}}]=\Sigma(p_{ij}/\sqrt{p_{.j}})=p_{.j}/\sqrt{p_{.j}})=\sqrt{p_{.j}}$, i=1, 2, ……..k (II.2.2)

and the variance

$v_j=\Sigma p_i[(p_{ij}/(p_{i.}\sqrt{p_{.j}})-\sqrt{p_{.j}}]^2=\Sigma(p_{ij}^2/p_{i.}p_{.j})- p_{.j}$; i=1, 2, ……..k (II.2.3)

The covariance between two columns s and t, r_s and r_t is

$cov(s, t)=\Sigma p_i[(p_{is}/(p_{i.}\sqrt{p_{.s}})-\sqrt{p_{.k}}][(p_{it}/(p_{i.}\sqrt{p_{.t}})-\sqrt{p_{.t}}]$ =>

$cov(s, t)=\Sigma[p_{is}p_{it}/(p_i\sqrt{p_{.s}}\sqrt{p_{.t}})]-\sqrt{p_{.s}}\sqrt{p_{.t}},$; i=1, 2, ……..k (II.2.4)

The Euclidean (squared) distance $d^2(l,q)$ between two rows l,q (two points in $\boldsymbol{R^p}$) is

$d^2(l,q)=\Sigma[(p_{lj}/(p_{l.}\sqrt{p_{.j}})-(p_{qj}/(p_{q.}\sqrt{p_{.j}})]^2$ =>

$d^2(l,q)=\Sigma(1/p_{.j})[(p_{lj}/(p_{l.})-(p_{qj}/(p_{q.})]^2$; j=1, 2…..p (II.2.5)

That is, the distance between two rows is the *weighted distance between the profiles l and q*, weighted by the inverse of the squared mean of each column.

It is now time to explain the transformation $r_{ij}=p_{ij}/(p_{i.}\sqrt{p_{.j}}]$. As indicated by the equation (III.3) the variance of the column j is $v_j=\Sigma(p_{ij}^2/p_{i.}p_{.j})-p_{.j}$; i=1,2,…k. Summing up the variances for j=1, 2,…p one obtains the sum of the total variance (variation) in the data:

$$\text{Sum of variance}=\Sigma[\Sigma(p_{ij}^2/p_{i.}p_{.j})-p_{.j}]=\Sigma\Sigma(p_{ij}^2/p_{i.}p_{.j})-\Sigma p_{.j} \Rightarrow$$

$$\text{Sum of variance}=\Sigma\Sigma(p_{ij}^2/p_{i.}p_{.j})-1;\ i=1, 2,\ldots k\ ;\ j=1,2,\ldots p \qquad \text{(II.2.6)}$$

But it is known that the stochastic variable $n\Sigma\Sigma[(p_{ij}-p_{i.}p_{.j})^2/p_{i.}p_{.j}]=n[\Sigma\Sigma(p_{ij}^2/p_{i.}p_{.j})-1]$ is distributed as χ^2 with (k-1)(p-1) degrees of freedom: $\chi^2= n[\Sigma\Sigma(p_{ij}^2/p_{i.}p_{.j})-1]$, or

$$\chi^2/n=\Sigma\Sigma(p_{ij}^2/p_{i.}p_{.j})-1;\ i=1, 2,\ldots k\ ;\ j=1, 2,\ldots p \qquad \text{(II.2.7)}$$

From (II.2.6) and (II.2.7) it results that

$$\text{Sum of variance}= \chi^2/n \qquad \text{(II.2.8)}$$

Further, it can easily be proved using Shannon's formula for mutual information between two distributions, that the quantity $\Sigma\Sigma(p_{ij}^2/p_{i.}p_{.j})-1$ is the mutual information $I_m(Y,Z)$ between the two qualitative variables Y and Z. Therefore, one can write

$$\text{Sum of variance}= \chi^2/n=\Sigma\Sigma(p_{ij}^2/p_{i.}p_{.j})-1=I_m(Y, Z) \qquad \text{(II.2.9)}$$

In the terminology of correspondence analysis the quantity χ^2/n is called *inertia,* meaning the variation in the data. This variation/information is analyzed by the CORRA according to the principal components technique, as described in the section II.1.

Further, some elementary calculations show that the distance between two rows is the χ^2/n between two distributions, one theoretical T_i and one observed O_i.

For i=1, 2,…n the quantity $\Sigma[O_i-T_i]^2/T_i$ is distributed as χ^2 with n-1 degrees of freedom: $\chi^2= \Sigma[O_i-T_i]^2/T_i= \Sigma[O_i-np_i]^2/np_i= n\Sigma[(O_i/n)-p_i]^2/p_i= n\Sigma(1/p_i)[(O_i/n)-p_i]^2 \Rightarrow$

$\chi^2/n=\Sigma(1/p_i)[(O_i/n)-p_i]^2$. In general, for two distributions X and Y it holds

$$\chi^2/n=\Sigma(1/p_i)[(X_i/n)-(Y_i/n)]^2,\ i=1,2\ldots n \qquad \text{(II.2.10)}$$

which is the form (II.2.5).

From a statistical point of view, two points close to each other show similarity of their distributions (similar profiles). In an informational context, the meaning of the closeness of two points is that if one aggregates the two neighboring points, that is, if the two rows are aggregated in one, the loss of information will be small. The last remark characterizes the technique of correspondence analysis: it is a technique of visualizing a large contingency table with minimum loss of information.

After obtaining the projection of the k points in $\boldsymbol{R^p}$ onto the factorial level, the procedure is repeated in exactly the same manner to the transposed matrix X' and the p points in $\boldsymbol{R^k}$ are projected onto the same factorial level. In this way one obtains the projection of the profiles of the variables Y and Z onto $\boldsymbol{R^2}$. It can be proved that the analysis in $\boldsymbol{R^P}$ and the analysis in $\boldsymbol{R^k}$ result to the same non-zero eigenvalues, which are at most p-1, with $p \leq k$.

3. Rules of Interpretation of the Points in the Plane

One of the most delicate problems in the correspondence analysis- and in all multidimensional techniques- is the interpretation of the points-projections in the factorial plane. The following rules of interpretation result from the way of construction of the CORRA:

- Two points of the same qualitative variable close to each other designate similarity of distribution. If the points are remote from each other they have quite different distributions. The terminology of CORRA describes such points as "opposed".
- One point close to the origin (centroid) of the axes follows the average distribution. A remote point exhibits distribution quite different from the average distribution.
- Two points of different variables close to each other indicate that the frequency in the intersection of the row and the column of the contingency table X is an outlier in the row and the column in which it belongs.

As mentioned above, the overall quality of the application is measured by the ratios $\lambda_1/\Sigma\lambda_i$ and $\lambda_2/\Sigma\lambda_i$. However, it is possible for two points to be close to each other *in the plane* while they are remote in space. For this reason one has always to check the squared direction cosines of the vector. The sum of all squared direction cosines equals 1. This means that if the sum of the squared direction cosines of the vector with the two factorial axes is close to one, the point is close to the factorial plane, i.e. the place of the point on the factorial is reliable. The last consideration permits one to define the sum of the squared direction cosines of the vector with the two factorial axes as "quality" of the point's projection.

Further, the interpretation of the proximity between two points of different variables is dangerous, is still a subject of disputation, and requires a cautious treatment. Nevertheless, a theoretical consideration offers a base of justification of the interpretation of the proximity between points of different variables. One can prove that between the component i of an eigenvector ψ in $\boldsymbol{R}^{\boldsymbol{k}}$ and the component j of an eigenvector φ in $\boldsymbol{R}^{\boldsymbol{p}}$, which correspond to the same eigenvalue q, the following mutual relations exist

$$\psi_{iq}=(1/\sqrt{\lambda_q})\Sigma(p_{ij}/p_{i.})\varphi_{jq} \; ; \; j=1, 2,\ldots p \qquad \text{(III.1a)}$$

$$\varphi_{jq}=(1/\sqrt{\lambda_q})\Sigma(p_{ij}/p_{.j})\psi_{jq} \; ; \; i=1, 2,\ldots k \qquad \text{(III.1b)}$$

The above formulae, also called "transition formulae", show that each component of an eigenvector ψ, associated to the eigenvalue q, is the weighted average, (centroid, barycentre in French CORRA terminology), at a constant factor $1/\sqrt{\lambda_q}$, of the profile of the row i, weighted by the components of φ, associated to the same eigenvalue - and inversely. This gives a justification basis for conceiving each point of the variable Y as the centroid of its surrounding point of Z - and inversely.

Application of Correspondence Analysis to the Greek Census 1981

For an illustration of correspondence analysis we apply the method to the Greek census of 1981. In particular, we investigate the contingency table Greek districts - Age of population.

Data, definition of variables, levels of variables and abbreviations

The districts are classified in 11 classes as in the following table 1

Table 1: Greek districts classification

District No.	District name	Abbreviation
1	Athens	ATH
2	Rest of Athens District (Sterea Ellada)	RATHD
3	Peloponnese	PEL
4	Ionian Islands	ION
5	Epirus	EPI
6	Thessalia	THL
7	Thessalonica	THN
8	Rest of Thessalonica District (Macedonia)	RTHND
9	Thrace	THRA
10	Aegean Islands	AEG
11	Crete	CRE

The ages are divided in 18 classes as in table 2

Table 2: Age classes

Age order	Age class	Abbreviation
1	[-05)	-05
2	[05-10)	05-10
3	[10-15)	10-15
4	[15-20)	15-20
5	[20-25)	20-25
6	[25-30)	25-30
7	[30-35)	30-35
8	[35-40)	35-40
9	[40-45)	40-45
10	[45-50)	45-50
11	[50-55)	50-55
12	[55-60)	55-60
13	[60-65)	60-65
14	[65-70)	65-70
15	[70-75)	70-75
16	[75-80)	75-80
17	[80-85)	80-85
18	[85-)	85-

The data were obtained from the National Statistical Service of Greece. The frequencies express thousand of habitants. The data is shown in the following table 3

Table 3: The data

	ATH	RATHD	PEL	ION	EPI	THL	THN	RTHND	THRA	AEG	CRE	Total Rows
-05	237	91	80	13	26	57	70	99	29	33	43	778
05-10	217	91	76	14	26	56	65	99	27	33	42	746
15-20	218	95	85	14	28	59	70	111	29	33	43	785
15-20	223	83	76	11	25	48	70	96	26	27	35	720
20-25	244	73	62	10	20	44	75	87	33	30	32	710
25-30	238	69	61	10	20	43	64	76	23	27	31	662
30-35	236	66	58	11	21	45	61	74	21	29	33	655
35-40	189	60	49	8	17	42	52	70	18	22	27	554
40-45	208	73	62	11	21	51	63	92	25	24	30	660
45-50	204	75	67	12	22	49	62	95	24	25	31	666
50-55	205	75	70	13	22	48	62	92	23	27	32	669
55-60	155	55	54	12	17	35	40	56	16	24	25	489
60-65	129	48	48	10	14	31	31	43	12	20	24	410
65-70	120	48	52	10	16	31	32	56	14	25	24	428
70-75	92	42	48	10	13	25	25	46	11	22	22	356
75-80	61	29	34	17	8	18	16	31	7	15	16	252
80-85	35	17	21	4	5	9	9	17	4	8	8	137
85-	17	9	10	2	3	5	5	10	3	5	5	74
Total Col	3028	1099	**1013**	**192**	**324**	**696**	**872**	**1250**	**345**	**429**	**503**	**9751**

Source: National Statistical Service of Greece. Frequencies x 1000

To make the above table commeasurable with the statistical terminology we have the variables:
Y: District, qualitative variables with 11 levels (characters in the CORRA terminology)
Z: Age, qualitative variables with 18 levels (characters)
X: the contingency table of Y and Z. We apply the CORRA on the contingency table.

Statistical treatment of the data

For the treatment of the data- calculations, tables and graphs we have used the STATISTICA package. The results are shown in the following files:

Corra 1-Data (Original Data)

	1 ATH	2 RATHD	3 PEL	4 ION	5 EPI	6 THL	7 THN	8 RTHND	9 THRA	10 AEG	11 CRE
-05	237	91	80	13	26	57	70	99	29	33	43
05-10	217	91	76	14	26	56	65	99	27	33	42
10-15	218	95	85	14	28	59	70	111	29	33	43
15-20	223	83	76	11	25	48	70	96	26	27	35
20-25	244	73	62	10	20	44	75	87	33	30	32
25-30	238	69	61	10	20	43	64	76	23	27	31
30-35	236	66	58	11	21	45	61	74	21	29	33
35-40	189	60	49	8	17	42	52	70	18	22	27
40-45	208	73	62	11	21	51	63	92	25	24	30
45-50	204	75	67	12	22	49	62	95	24	25	31
50-55	205	75	70	13	22	48	62	92	23	27	32
55-60	155	55	54	12	17	35	40	56	16	24	25
60-65	129	48	48	10	14	31	31	43	12	20	24
65-70	120	48	52	10	16	31	32	56	14	25	24
70-75	92	42	48	10	13	25	25	46	11	22	22
75-80	61	29	34	17	8	18	16	31	7	15	16
80-85	35	17	21	4	5	9	9	17	4	8	8
85-	17	9	10	2	3	5	5	10	3	5	5

Corra 2-Row Percentages

Percentages of Row Totals (corra1-data.sta)											
	1 ATH	**2 RATHD**	**3 PEL**	**4 ION**	**5 EPI**	**6 THL**	**7 THN**	**8 RTHND**	**9 THRA**	**10 AEG**	**11 CRE**
-05	30.46272	11.69666	10.28278	1.670951	3.341902	7.326478	8.997429	12.72494	3.727506	4.241645	5.526992
05-10	29.08847	12.19839	10.18767	1.876676	3.485255	7.506702	8.713137	13.27078	3.619303	4.423592	5.630027
10-15	27.7707	12.10191	10.82803	1.783439	3.566879	7.515924	8.917197	14.14013	3.694268	4.203822	5.477707
15-20	30.97222	11.52778	10.55556	1.527778	3.472222	6.666667	9.722222	13.33333	3.611111	3.75	4.861111
20-25	34.3662	10.28169	8.732394	1.408451	2.816901	6.197183	10.56338	12.25352	4.647887	4.225352	4.507042
25-30	35.95166	10.42296	9.214502	1.510574	3.021148	6.495468	9.667674	11.48036	3.47432	4.07855	4.682779
30-35	36.03053	10.07634	8.854962	1.679389	3.206107	6.870229	9.312977	11.29771	3.206107	4.427481	5.038168
35-40	34.11552	10.83032	8.844765	1.444043	3.068592	7.581227	9.386282	12.63538	3.249097	3.971119	4.873646
40-45	31.51515	11.06061	9.393939	1.666667	3.181818	7.727273	9.545455	13.93939	3.787879	3.636364	4.545455
45-50	30.63063	11.26126	10.06006	1.801802	3.303303	7.357357	9.309309	14.26426	3.603604	3.753754	4.654655
50-55	30.64275	11.21076	10.46338	1.943199	3.28849	7.174888	9.267564	13.75187	3.437967	4.035874	4.783259
55-60	31.69734	11.24744	11.04294	2.453988	3.476483	7.157464	8.179959	11.45194	3.271984	4.907975	5.112474
60-65	31.46341	11.70732	11.70732	2.439024	3.414634	7.560976	7.560976	10.4878	2.926829	4.878049	5.853659
65-70	28.03738	11.21495	12.14953	2.336449	3.738318	7.242991	7.476636	13.08411	3.271028	5.841121	5.607477
70-75	25.8427	11.79775	13.48315	2.808989	3.651685	7.022472	7.022472	12.92135	3.089888	6.179775	6.179775
75-80	24.20635	11.50794	13.49206	6.746032	3.174603	7.142857	6.349206	12.30159	2.777778	5.952381	6.349206
80-85	25.54745	12.40876	15.32847	2.919708	3.649635	6.569343	6.569343	12.40876	2.919708	5.839416	5.839416
85-	22.97297	12.16216	13.51351	2.702703	4.054054	6.756757	6.756757	13.51351	4.054054	6.756757	6.756757

Corra 3-Col Percentages

Percentages of Column Totals (corra1-data.sta)											
	1 ATH	**2 RATHD**	**3 PEL**	**4 ION**	**5 EPI**	**6 THL**	**7 THN**	**8 RTHND**	**9 THRA**	**10 AEG**	**11 CRE**
-05	7.826948	8.280255	7.897335	6.770833	8.024691	8.189655	8.027523	7.92	8.405797	7.692308	8.548708
05-10	7.166446	8.280255	7.502468	7.291667	8.024691	8.045977	7.454128	7.92	7.826087	7.692308	8.349901
10-15	7.199472	8.644222	8.390918	7.291667	8.641975	8.477011	8.027523	8.88	8.405797	7.692308	8.548708
15-20	7.364597	7.55232	7.502468	5.729167	7.716049	6.896552	8.027523	7.68	7.536232	6.293706	6.95825
20-25	8.058124	6.642402	6.120434	5.208333	6.17284	6.321839	8.600917	6.96	9.565217	6.993007	6.361829
25-30	7.859974	6.278435	6.021718	5.208333	6.17284	6.178161	7.33945	6.08	6.666667	6.293706	6.163022
30-35	7.793923	6.00546	5.725568	5.729167	6.481481	6.465517	6.995413	5.92	6.086957	6.759907	6.560636
35-40	6.241744	5.459509	4.837117	4.166667	5.246914	6.034483	5.963303	5.6	5.217391	5.128205	5.367793
40-45	6.869221	6.642402	6.120434	5.729167	6.481481	7.327586	7.224771	7.36	7.246377	5.594406	5.964215
45-50	6.73712	6.824386	6.614018	6.25	6.790123	7.04023	7.110092	7.6	6.956522	5.827506	6.163022
50-55	6.770145	6.824386	6.910168	6.770833	6.790123	6.896552	7.110092	7.36	6.666667	6.293706	6.361829
55-60	5.11889	5.00455	5.330701	6.25	5.246914	5.028736	4.587156	4.48	4.637681	5.594406	4.970179
60-65	4.260238	4.367607	4.738401	5.208333	4.320988	4.454023	3.555046	3.44	3.478261	4.662005	4.771372
65-70	3.963012	4.367607	5.133268	5.208333	4.938272	4.454023	3.669725	4.48	4.057971	5.827506	4.771372
70-75	3.038309	3.821656	4.738401	5.208333	4.012346	3.591954	2.866972	3.68	3.188406	5.128205	4.373757
75-80	2.014531	2.638763	3.356367	8.854167	2.469136	2.586207	1.834862	2.48	2.028986	3.496503	3.180915
80-85	1.155878	1.546861	2.07305	2.083333	1.54321	1.293103	1.03211	1.36	1.15942	1.864802	1.590457
85-	0.561427	0.818926	0.987167	1.041667	0.925926	0.718391	0.573394	0.8	0.869565	1.165501	0.994036
Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Corra 4-Total percentages

Percentages of Total (corra1-data.sta)												
	1 ATH	**2 RATHD**	**3 PEL**	**4 ION**	**5 EPI**	**6 THL**	**7 THN**	**8 RTHND**	**9 THRA**	**10 AEG**	**11 CRE**	**TOTAL**
-05	2.4305199	0.9332376	0.8204287	0.1333197	0.2666393	0.5845554	0.7178751	1.0152805	0.2974054	0.3384268	0.4409804	7.9786689
05-10	2.2254128	0.9332376	0.7794072	0.143575	0.2666393	0.5743001	0.6665983	1.0152805	0.2768947	0.3384268	0.4307251	7.6504974
10-15	2.2356681	0.9742591	0.8717055	0.143575	0.28715	0.6050661	0.7178751	1.1383448	0.2974054	0.3384268	0.4409804	0504564
15-20	2.2869449	0.8511947	0.7794072	0.1128089	0.256384	0.4922572	0.7178751	0.9845144	0.2666393	0.2768947	0.3589375	7.3838581
20-25	2.5023075	0.7486412	0.6358322	0.1025536	0.2051072	0.4512358	0.7691519	0.8922162	0.3384268	0.3076608	0.3281715	7.2813045
25-30	2.4407753	0.7076197	0.6255769	0.1025536	0.2051072	0.4409804	0.6563429	0.7794072	0.2358732	0.2768947	0.3179161	6.7890473
30-35	2.4202646	0.6768537	0.5948108	0.1128089	0.2153625	0.4614911	0.6255769	0.7588965	0.2153625	0.2974054	0.3384268	6.7172598
35-40	1.9382627	0.6153215	0.5025126	0.0820429	0.1743411	0.4307251	0.5332786	0.7178751	0.1845965	0.2256179	0.2768947	5.6814686
40-45	2.1331146	0.7486412	0.6358322	0.1128089	0.2153625	0.5230233	0.6460876	0.943493	0.256384	0.2461286	0.3076608	6.7685366
45-50	2.0920931	0.7691519	0.687109	0.1230643	0.2256179	0.5025126	0.6358322	0.9742591	0.2461286	0.256384	0.3179161	6.8300687
50-55	2.1023485	0.7691519	0.7178751	0.1333197	0.2256179	0.4922572	0.6358322	0.943493	0.2358732	0.2768947	0.3281715	6.8608348
55-60	1.5895806	0.5640447	0.5537894	0.1230643	0.1743411	0.3589375	0.4102143	0.5743001	0.1640857	0.2461286	0.256384	5.0148703
60-65	1.3229412	0.4922572	0.4922572	0.1025536	0.143575	0.3179161	0.3179161	0.4409804	0.1230643	0.2051072	0.2461286	4.204697
65-70	1.230643	0.4922572	0.5332786	0.1025536	0.1640857	0.3179161	0.3281715	0.5743001	0.143575	0.256384	0.2461286	4.3892934
70-75	0.943493	0.4307251	0.4922572	0.1025536	0.1333197	0.256384	0.256384	0.4717465	0.1128089	0.2256179	0.2256179	3.6509076
75-80	0.6255769	0.2974054	0.3486822	0.1743411	0.0820429	0.1845965	0.1640857	0.3179161	0.0717875	0.1538304	0.1640857	2.5843503
80-85	0.3589375	0.1743411	0.2153625	0.0410214	0.0512768	0.0922982	0.0922982	0.1743411	0.0410214	0.0820429	0.0820429	1.4049841
85-	0.1743411	0.0922982	0.1025536	0.0205107	0.0307661	0.0512768	0.0512768	0.1025536	0.0307661	0.0512768	0.0512768	0.7588965
Total	**31.053225**	**11.270639**	**10.388678**	**1.9690288**	**3.3227361**	**7.1377295**	**8.9426725**	**12.819198**	**3.5380987**	**4.3995488**	**5.1584453**	**100**

Corra 5-Col Cos2 (Columns Analysis: Coordinates, Squared Cosines, Quality, etc)

Column Coordinates and Contributions to Inertia										
	1 COLUMN_N	2 COORDIN.	3 COORDIN.	4 MASS	5 QUALITY	6 RELATIVE	7 INERTIA	8 COSINE²	9 INERTIA	10 COSINE²
ATH	1	0.085843893	-0.04798482	0.310532	0.9915876	0.22300829	0.2508234	0.7555206	**0.3062971**	**0.23606699**
RATHD	2	-0.035310092	0.03615269	0.112706	0.7818597	0.02710515	0.0154024	0.3817124	0.0631042	0.40014726
PEL	3	-0.124935019	0.012633387	0.103887	0.8821191	0.13672926	0.1777343	0.8731905	0.0071028	0.00892854
ION	4	-0.378352767	-0.16685274	0.01969	0.8830171	0.28073536	0.3089502	0.7392487	0.2348266	0.14376841
EPI	5	-0.046279057	0.034694928	0.033227	0.63099	0.01297104	0.0078002	0.403954	0.0171339	0.22703607
THL	6	-0.0154993	0.02444211	0.071377	0.2238612	0.01966451	0.0018794	0.0642012	0.0182669	0.15966005
THN	7	0.10196363	0.014949369	0.089427	0.8799531	0.07946517	0.1019062	0.8614358	0.0085613	0.01851732
RTHND	8	-0.006407525	0.071308063	0.128192	0.7625478	0.06344617	0.0005769	0.0061077	0.2792329	0.75644005
THRA	9	0.060857039	0.040735368	0.035381	0.4066473	0.0343556	0.0143626	0.2808251	0.0251502	0.12582227
AEG	10	-0.127959597	-0.04625411	0.043995	0.7223696	0.0830175	0.0789581	0.6388899	0.0403215	0.0834797
CRE	11	-0.085782458	0.000337674	0.051584	0.7075314	0.03950196	0.0416062	0.7075204	0.000003	0.000011

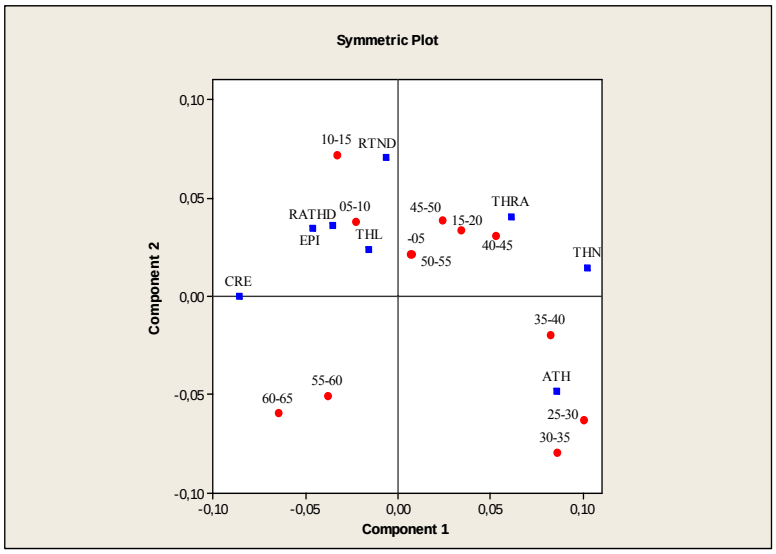
Corra 6-Row Cos2 (Rows Analysis: Coordinates, Squared Cosines, Quality, etc)

Row Coordinates and Contributions to Inertia (corra1-data.sta)										
	1 ROW_NUMB	2 COORDIN.	3 COORDIN.	4 MASS	5 QUALITY	6 RELATIVE	7 INERTIA	8 COSINE²	9 INERTIA	10 COSINE²
-05	1	0.00657139	0.022074319	0.079787	0.4355034	0.007155373	0.0003776	0.03545309	0.01665457	0.40005032
05-10	2	-0.02297945	0.038594494	0.076505	0.6662468	0.017058045	0.004428	0.17437387	0.04881675	0.49187294
10-15	3	-0.03317418	0.072674445	0.080505	0.9760199	0.038758535	0.009711	0.16830439	0.18214296	0.807715475
15-20	4	0.03347129	0.033906741	0.073839	0.6512231	0.018950479	0.0090672	0.32140305	0.03636502	0.329820047
20-25	5	0.11421214	-0.026020387	0.072813	0.7536055	0.097613006	0.1041061	0.71642022	0.0211186	0.03718526
25-30	6	0.09998238	-0.062659125	0.06789	0.9749022	0.071385707	0.0743872	0.69998108	0.11418418	0.274921075
30-35	7	0.0852452	-0.078952244	0.067173	0.9391207	0.07109725	0.0535026	0.50549986	0.17936989	0.4336208
35-40	8	0.08196565	-0.018892028	0.056815	0.8409962	0.035192638	0.0418376	0.79857253	0.00868653	0.042423623
40-45	9	0.05264604	0.03141925	0.067685	0.661153	0.028332237	0.0205622	0.487514	0.02862298	0.173639029
45-50	10	0.02343455	0.0391756	0.068301	0.5782754	0.018122169	0.0041113	0.15239467	0.04490398	0.425880763
50-55	11	0.00713006	0.022039577	0.068608	0.3649938	0.007426229	0.0003823	0.03458097	0.01427618	0.330412845
55-60	12	-0.03798848	-0.050056886	0.050149	0.8378862	0.017401367	0.0079324	0.30621177	0.05382894	0.531674445
60-65	13	-0.06465877	-0.059157978	0.042047	0.6303869	0.037718504	0.0192678	0.34314441	0.06303616	0.287242498
65-70	14	-0.10754274	0.0083206	0.043893	0.7802586	0.048189461	0.0556416	0.77561569	0.00130176	0.004642946
70-75	15	-0.18037513	0.011372143	0.036509	0.9058622	0.096929663	0.1301957	0.90227575	0.00202262	0.003586498
75-80	16	-0.35519692	-0.124973362	0.025844	0.8975299	0.300586176	0.3573817	0.79866116	0.17290786	0.098868763
80-85	17	-0.21102574	0.010403386	0.01405	0.8440793	0.054708535	0.068578	0.8420328	0.0006514	0.00204648
85-	18	-0.21522095	0.058458031	0.007589	0.8327048	0.033374624	0.0385296	0.77549155	0.01110962	0.057213215

Corra 7- Eigenvalues

Eigenvalues and Inertia for all Dimensions (corra1-data.sta)					
	1 SINGULAR	2 EIGEN__V	3 PERC. OF	4 CUMULATV	5 CHI_SQUA
1	0.095516567	0.009123415	67.17369302	67.17369302	88.96241596
2	0.048315422	0.00233438	17.18752611	84.36121914	22.7625396
3	0.032667115	0.00106714	7.857119896	92.21833903	10.40568617
4	0.022429187	0.000503068	3.703982045	95.92232108	4.905420211
5	0.016107963	0.000259466	1.910394469	97.83271555	2.530057524
6	0.013818631	0.000190955	1.405956317	99.23867187	1.861997831
7	0.00792363	0.000063	0.462264078	99.70093594	0.61220587
8	0.004899199	0.000024	0.176722556	99.8776585	0.234044979
9	0.003403602	0.000012	0.08529417	99.96295267	0.112960522
10	0.002243146	0.000005	0.03704733	100	0.049064148

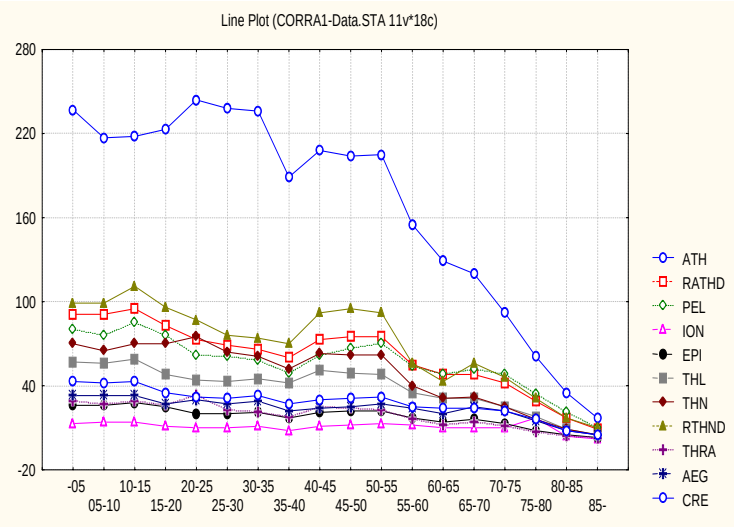
G-Corra1-Factorial Plane (Graph of the Factorial Plane)



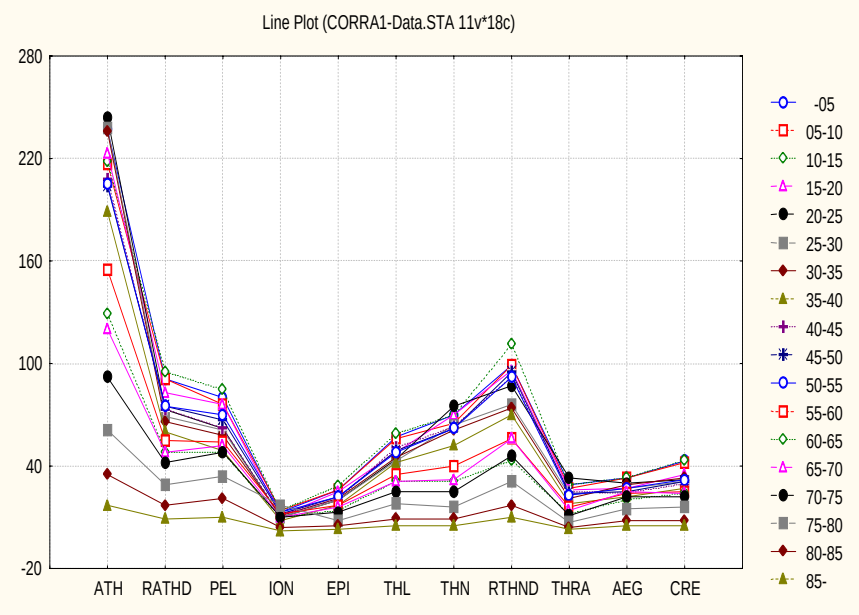
Analysis of Contingency Table

Axis	Inertia	Proportion	Cumulative	Histogram
1	0,0091	0,6717	0,6717	**********
2	0,0023	0,1719	0,8436	*******
3	0,0011	0,0786	0,9222	***
4	0,0005	0,0370	0,9592	*
5	0,0003	0,0191	0,9783	
6	0,0002	0,0141	0,9924	
7	0,0001	`0,0046	0,9970	
8	0,0000	0,0018	0,9988	
9	0,0000	0,0009	0,9996	
10	0,0000	0,0004	1,0000	
Total	**0,0136**			

G-CORRA2-COL PERCENTAGES (graph of the columns percentages)



G-CORRA3-ROW PERCENTAGES (graph of the rows percentages)



Findings of the analysis

IV.3.1 Value of χ^2, degrees of freedom, p-value and total inertia (all shown in file CORRA1-DATA)

χ^2 =132.44

d.o.f=(18-1)(11-1)=170

Total inertia=I= χ^2/n=132.44/9751=0.01358

The found value of $\chi 2$ =132.44 with 170 d.o.f corresponds to a p-value 0.9849, which at level of significance 5% does not reject the null hypothesis that the two qualitative variables are independent, that is, the ages are distributed more or less proportionally in the districts-and vise-versa. Nevertheless, there exist similarities and dissimilarities in the row and in the column distributions, which will be detected by the correspondence analysis.

Quality of the overall application (shown in file CORRA7-EIGENVALUES)

The first eigenvalue, assigned to the first (horizontal) factorial axis covers 67.71% of the total inertia and the second eigenvalue, assigned to the second (vertical) factorial axis, covers 17.19%, i.e. the first two factorial axes represent ("explain") cumulatively 67.71+19.19=84.37% of the total inertia. This means very reliable overall projection of the points onto the factorial plane, the distances between the points are quite reliable.

Quality of the projection of the individual points (shown in files CORRA5-COL COS2 and CORRA6-ROW COS2 under the title "Quality")

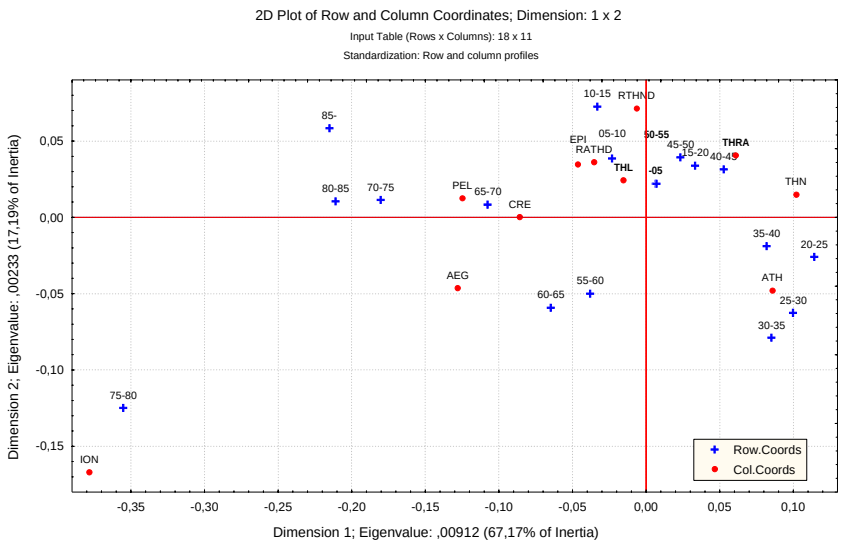
The age classes are well placed, except the ages **–05** (0.4355) and **50-55** (0.3649); poor sum of squared direction cosines.

The districts are also well placed, except the districts **THL** (0.2238) and **THRA** (0.4066)

Discussion of the findings

For easy reference we present the factorial plane in the following graph 1, which represents the factorial plane.

Figure 1: The factorial plane



In the above graph we read:

a/ The older ages cover the left area of the plane while the younger are concentrated on the right part. That is opposition exists between the distributions of the older ages (considered together) in relation to the younger ages (considered together)

b/ the districts clearly form four distinct clusters:

{ION} alone

{PEL, CRE, AEG}

{EPI, RATHD, RTHND}

{ATH, THN}

In each cluster the ages are distributed quite uniformly.

c/ ION is close to the age 75-80. A check in the profiles can verify that this age covers an extraordinary high place in the distribution of the ages in ION- and vice-versa: ION covers a high place in the distribution of ages.

Bb/ ATH and THL attract the more active ages 25-40, while PEL, CRE and AGE concentrate the older (and less active) ages

d/ The outliers -05, 50-55, THL and THRA follow quite particular distributions and they should be checked individually in the corresponding frequency tables.

Conclusion

The purpose of this study was to present the correspondence analysis rather than its concrete application of the given contingency table. The latter would demand much more considerations pertaining to a more detailed analysis of the profiles and, most importantly, the socio-economic reasons which create the given distribution of the ages on the Greek districts. For example, why the Ionian Islands, Peloponnese, the Aegean Islands show higher percentages of old ages, while Athens and Thessalonica attract the younger, active ages. This is a task going far beyond the scope of the study. The application of the correspondence analysis gave a panoramic and reliable view of the inter-relations in the data. The method has raised objections, relating to the interpretation of the points and the factorial axes. This is a problem common to all multidimensional techniques. Further, a common reaction to the results of correspondence analysis is "Yes, we knew these results in advance". The answer resembles "Yes, but you have forgotten them- and now you have refreshed and enriched your memory"

Bibliography

[1] Basilevsky, A. (2005) *Applied Matrix Algebra in the Statistical Sciences*, New York: Dover Publications
[2] Benzıcri, J.P. (1970) Distance distributionnelle et mıtrique de χ^2 en analyse
[3] Benzıcri, J.P. (1982) Histoire et prıhistoire de l' analyse des donnıes, Paris: Dunod
[4] Benzıcri, J.P. et coll. (1973) *L' analyse des donnıes, t.I: La taxinomie, t.II:L' analyse* des correspondences, Paris: Dunod
[5] Burt, C. (1950) The factorial Analysis of Qualitative Data. British *Journal of factorielle des correspondences*, Paris: Dunod
[6] Greenacre, M. (1984) Theory and Applications of Correspondence Analysis, London: Academic Press
[7] Hotteling, H. (1933) Analysis of a Complex of Statistical variables into Principal Components. *Journal of Educational Psychology*, 24, pp. 417-441
[8] Johnson, R.A., Wichern, D.W. (1988) *Applied Multivariate Statistical Analysis*, New Jersey: PRENTICE HALL
[9] Khinchin, A. I. (1957) *Mathematical Foundations of Information Theory*, New York: Dover Publications
[10] Lebart, L., Morineau, A., Warwick, K.W. (1984) Multivariate Descriptive Statistical Analysis, Correspondence Analysis and Related Techniques for Large Matrices, New York: Wiley matematiques pour l' analyse des donnıes, Paris: Dunod
[11] Nakache, J. P., Chevalier, A., Morice, V. (1981) *Exercises commentıs de*
[12] Pearson, K. (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, 2(11), pp.559-572
[13] Perlis, S. (1991) *Theory of matrices*, New York: Dover Publications, *Statistical Psycology*, 3(3) pp.166-185
[14] Shannon, C. E. (1948) The mathematical theory of communication. *Bell Syst, Techn. Journ.*, 27 pp. 379-423 ; 623-656
[15] Ventsel, H. (1973) *Thıorie des Probabilitıs*, Moscow: MIR Publications

Software for Correspondence Analysis

Microsoft, SPSS package
Microsoft, STATISTICA package
Microsoft, MINITAB package