__draft__

AFC97: A New Software Implementation for Correspondence Analysis

NIKOS KOUTSOUPIAS
Department of Marketing
Technological Institute of Thessaloniki
Odysseos 32, 56625 Thessaloniki
GREECE
nickk@uom.gr

Abstract: - The extensive utilization of the statistical data analysis method of Correspondence (Factor) Analysis in a wide range of scientific fields and the need for further improvement of the method's results and their interpretation led us in a new, independent software implementation of that method.

This paper proposes an alternative view of the method's format of results in an effort to aid the data analyst come up with clear, concise and well perceived findings related to the phenomenon s/he analyzes. The method of Correspondence Analysis is re-implemented in graphical (MS-Windows) environment with emphasis on the design of the interpretation tables that are now produced in MS Excel format and the capability for selective visualization of important information emerging from the data analyzed.

Key-Words: Statistical Software, Data Analysis Software, Correspondence Analysis, Factor Analysis

1. Introduction

During the last few decades Data Analysis is gaining increasing acceptance mainly due to the extensive dissemination of computers and the weaknesses of traditional statistical methods when applied on large data collections.

Furthermore, personal computers offer great assistance in the development of new visualization methods that, in a great extend, support the fundamental principles of the Statistical Data Analysis methods.

The application of data analysis methods is impossible without the use of computer software. Along with known commercial statistical software packages such as Statgraphics [14], S-Plus [12], SAS/STAT [11], SPSS [13] that have included data analysis functions, since the early 80's, many independent researchers developed specialized software packages for the analysis of data on their respective fields. Significant efforts where Blumenthal's MULTISTAT software [3], the MacMul package by Thioulouse [15], Felsenstein's [4] PHYLIP statistical software for the analysis of genetical data, the multivariate statistical package MVSP by Kovach [9], Karakos' [6] PRAXITELIS data analysis software as well as Karapistolis' [7] SPA package for the statistical analysis of marketing data.

Commercial statistical packages present the statistical data analysis results in a brief, condensed manner and software developed by individual researchers is mainly not user friendly. Furthermore, most of the software mentioned thus far, is incapable of setting selection or importance criteria.

2. The AFC97 Software by Example

The primary motive that led us to rebuild and re-implement the method from the beginning was the notable absence of a user-friendly data analysis package for Correspondence Analysis that is comprehensive and well organized in its output results. In addition, special emphasis was put in providing the capability for selective output at the level of factorial axes and plots, as well as in the visualization of the results.

Both input (data) and output (results) files are in MS Excel format. The complete software package (AFC97) was written under the graphical environment of MS Windows using Visual Basic for Applications. We believe that this software can be a practical tool for the data analyst towards a better understanding of the method's results.

The idea for the creation of this statistical software package was initiated from the need of the

author to analyze life insurance data during the period between 1992 and 1995 from Northern Greece [8].

For the sake of a better understanding of the characteristics and capabilities of the software, we will utilize a data set that refers to Greek occupational data for the year 1996. The example data set contains 10 rows (corresponding to types of occupation) and 8 columns (referring to the level of education). The following table describes the different types of occupations:

Occupation	Description
AM	Higher Administrative/Managerial
SA	Scientific/Artistic
TA	Technological/Tech Assistantship
CL	Clerks
SE	Services Personnel
SF	Skilled Farmers
ST	Skilled Technicians
MW	Machinery/Industrial Workers
UW	Unskilled/Manual Workers
AF	Armed Forces

For the example table the educational levels (columns) are: Uneducated, Primary School, Gymnacium, Lycee, Higher Private Education, Technological Educational Institute, University and Graduate as seen below (Table 1).

Table 1 : Example Data on Occupation and Level of Education.

E/O	Uned	Prim	Gymn	Lycee	HPriv	TEI	Univ	Grad
AM	2050	148144	43513	140713	13539	12482	43370	3707
SA	0	2986	1481	12300	19672	6527	388138	10557
TA	0	7184	7068	59399	60785	66661	25992	809
CL	0	21229	24277	273385	28898	12997	34405	492
SE	1745	130705	65415	196205	21540	12217	12958	225
SF	38166	619861	49593	50408	1661	2091	3199	0
ST	3675	317891	103998	161961	21139	7540	7599	0
MW	1099	153635	50789	76800	5671	4230	2835	0
UW	10158	147629	24410	40201	2953	1661	4241	0
AF	0	2137	1745	13646	953	12734	11773	75

SOURCE: Research Center for Gender Equality (www.kethi.gr, 1996)

Before the data processing, apart from the data file name itself, the user must provide the following information for both active and supplementary rows and columns:

- Number of rows and columns.
- Number of factorial axes and results tables.
- Specification of up to 3 factorial plots

Correspondence Analysis utilizes factorial diagrams (or factor score plots) in order to aid the interpretation of the analyzed phenomenon (Benzecri [1,2], Papadimitriou [10]). The results of the method include, along with the plots, absolute and relative contribution tables (Greenacre [5]).

The AFC97 software when applied on the above example data (Table 1) firstly produces information about the characteristic values of the data table (Figure 1).

A/A	CHARACT VAL	%	CUMULATIVE
1	0,664232	57,156530	57,156530
2	0,319422	27,485925	84,642455
3	0,143367	12,336623	96,979078
4	0,023141	1,991229	98,970307
5	0,009854	0,847889	99,818195
6	0,001411	0,121381	99,939576
7	0,000702	0,060424	100,000000

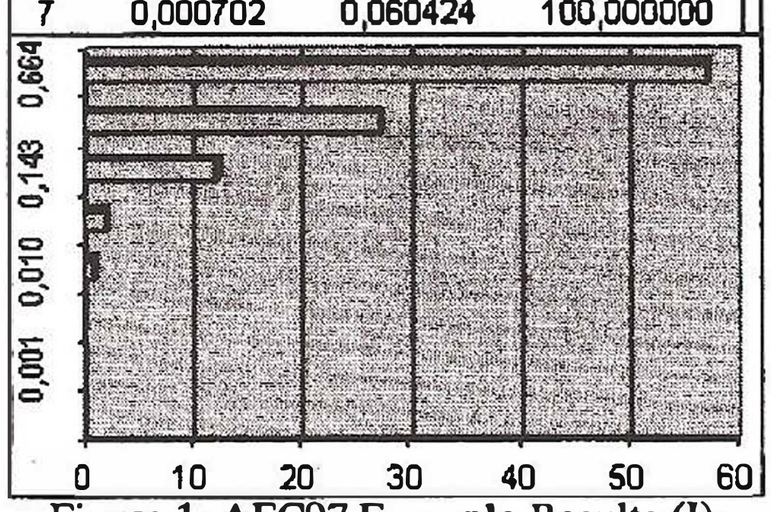

Figure 1: AFC97 Example Results (I)

In addition AFC97 creates interpretation tables of the following form (Figure 2).

	QTY	MASS	INER	1st FACT	COR	CTR	2nd FACT	COR	CTR
μ1	751	15	24	-652	228	8	852	389	31
μ2	998	401	172	-529	559	168	445	395	247
μ3	551	96	28	-384	432	21	-32	3	0
μ4	982	265	122	-156	45	9	-564	591	262
μ5	911	46	66	352	74	8	-1020	621	146
μ6	976	36	122	366	33	7	-1425	515	222
μ7	998	138	453	1908	953	755	419	45	75
μ8	925	4	12	1745	880	18	393	44	1

Figure 2 : AFC97 Example Results (II)

The above tables contain coordinates, masses, inertias and COR/CTR coefficients of row and column data. The first column (QTY) denotes the total inertia of the first axes, the second (MASS) gives the mass of each row and the next column

refers to the inertia (INER) of the corresponding row. Then, for each factorial axis, AFC97 produces three columns containing the coordinates (1st/2nd Fact), the contribution (COR) and quality of representation (CTR) coefficients for all rows.

Following the above results, AFC97 creates the factorial plots containing cow and/or column points in a graphical form (Figure 3).

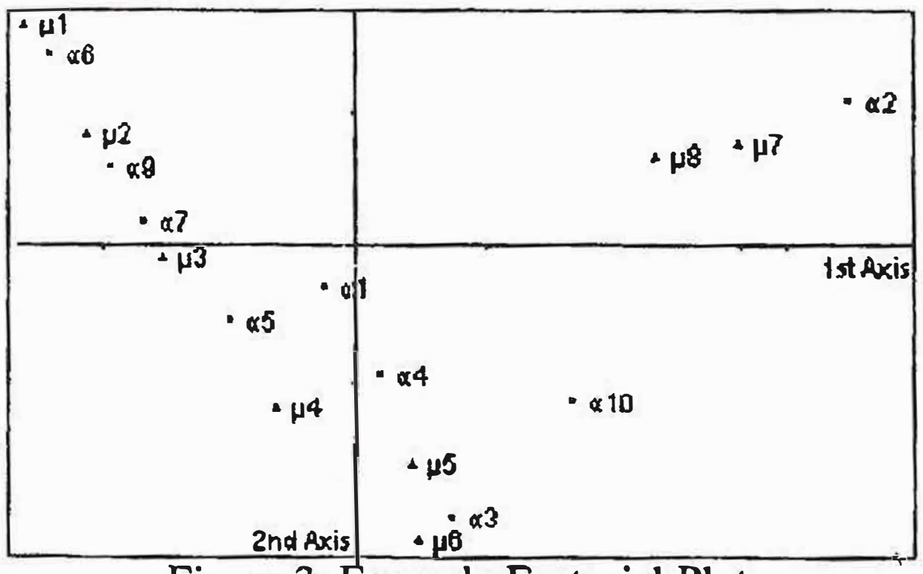

Figure 3: Example Factorial Plot

In the figure above we see the arrangement for both row (α) and column (μ) points conveniently allowing for a quick view on data formations, trends and groupings. Along with the above outputs, the AFC97 software keeps track of possible overlapping points on the factorial plots. These points are recorded on a separate table.

Moreover, AFC97 provides the means for a definition of the level of importance of contribution (COR) and quality of representation (CTR) of rows and columns, for all factorial axes. The data analyst is prompted for the desirable (minimum) values of the above COR/CTR indices (Figure 4).

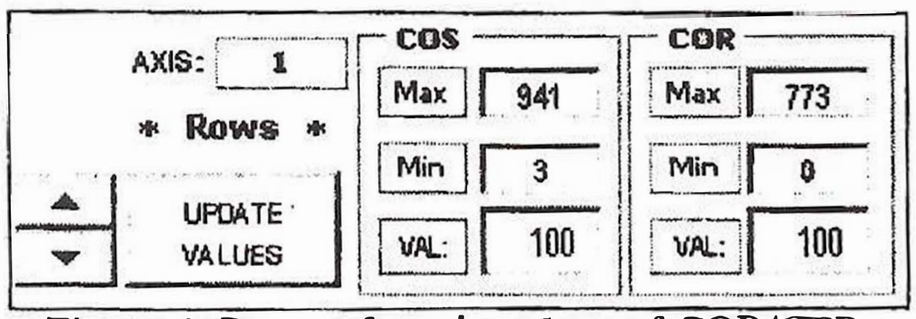

Figure 4: Prompt for min values of COR/CTR

In this way, the factorial axes are reproduced carrying only the row and/or column points that satisfy the criteria (minimum values) posed by the data analyst. The resulting axis (for rows and columns), using the example data and setting the value 100 as a minimum for both COR and CTR appears as follows:

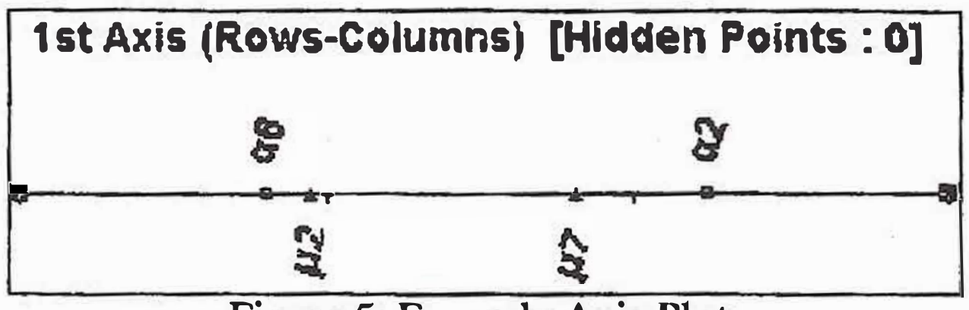

Figure 5: Example Axis Plot

Here, as in the factorial plots described above, AFC97 provides information about the number of possible hidden (or overlapped) row/column points, as well as a detailed description of the visible points that cover the hidden ones. The produced factorial axes provide an easy way to distinguish the main trends in the data set.

The next step in the execution of AFC97 is the combination of any two axes produced during the previous phase. That is, the user is prompted for the specification of the characteristics of factorial plots with points that satisfy the axis criteria given earlier. In this way, the plot of the first factorial plane (for axes 1 and 2) contains only those points that characterize that plane (Figure 6).

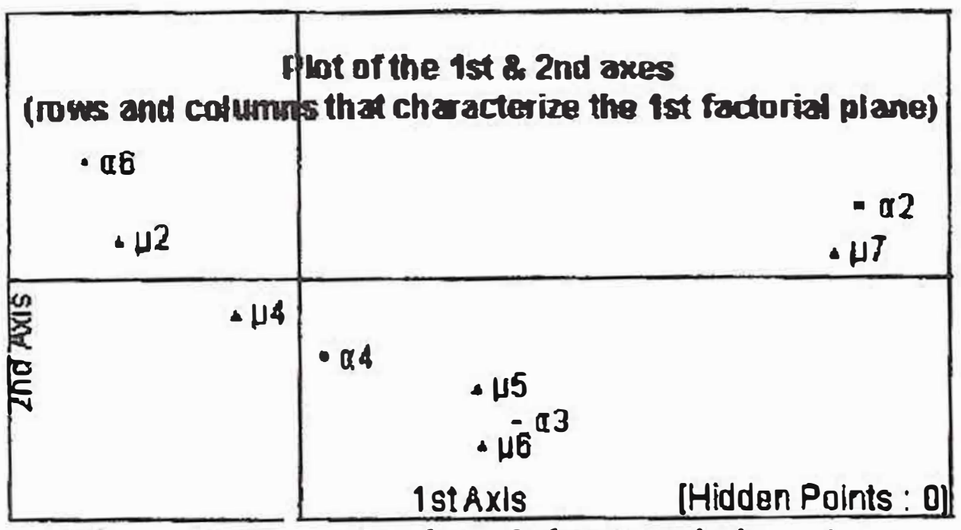

Figure 6: Example plot of characteristic points

In this way, it is clear to understand which ones are the main trends and relationships in the data set examined.

3. Conclusion

The statistical data analysis software package (AFC97) described above, is a tool for a better understanding and interpretation of the results of the statistical data analysis method of Correspondence (Factor) Analysis. It runs under a graphical environment and, we believe, significantly improves the data visualization process of the method. Special interest was put on the plotting of factorial planes and the row/column points that are overlapped by visible

ones. Furthermore, the user of AFC97 is able to define criteria for both axes and plane points. In this way, the software promotes the interpretation process, since the plots include only to significant information in the form of points that characterize each factorial axis and/or plane.

References:

[1] Benzécri, J.-P. Pratique ed l' Analyse des Données (T.1: Analyse des Correspondances, exposé élémentaire), Dunod, Paris, 1980.

[2] Benzecri, J.-P., Analyse des Données (T. 2: Correspondances), Dunod, Paris, 1973.

[3] Blumenthal, S., Microstat: un logiciel conversationnel de traitements statistiques pour micro-ordinateour Apple II, *Data Analysis and Informatics III*, Diday E. et. al. (eds.), North-Holland, Amsterdam, 1984, pp.517-528.

[4] Felsenstein, J., *PHYLIP Manual Version 3.2*, University of California Herbarium, Berkeley, California., 1989, pp. 285-293.

[5] Greenacre, M., *Correspondance Analysis in Practice, Academic Press*, London, 1993 .

[6] Karakos, A., A software program for the exploitation of Data Analysis Techniques v.3.0, Xanthi, 1988-1994.

[7] Karapistolis, D., *Creation of software for the establishment of a solvent portfolio with Data Analysis methods*, Phd thesis, Thessaloniki, 1996.

[8] Koutsoupias N., *Statistical Analysis of the Northern Greek Life Insurance Market using Relational Databases (RDBMs) and Structure Query Language (SQL)*, Phd. thesis, Thessaloniki, 1999.

[9] Kovach, W.L., MVSP - A MultiVariate Statistical Package, ver. 2. *INQUA Working Group on Data-Handling Methods Newsletter* vol.4, 1990, pp.1-3.

[10] Papadimitriou, J., Data Analysis Methods (University Lectures), Publications of University of Macedonia, Thessaloniki, 1994.

[11] SAS/STAT, *User's Guide*: Ver.6, 4th Edition, Vol. 1, 2, SAS Inst. Inc., Cary NC, 1998.

[12] S-Plus 4, *Guide To Statistics,* MathSoft Inc., July 1997, pp. 483-497.

[13] SPSS, *Applications Guide,* SPSS Inc., 1998, pp. 317-353.

[14] Statgraphics, *Multivariate Methods,* Manugistics

[15] Thioulouse, J., Statistical analysis and graphical display of multivariate data on the Macintosh, *Computer Applications in the Biosciences*, vol. 5, no. 4, 1989, pp. 287-292.