Multiple Correspondence Analysis

The **Multiple Correspondence Analysis** procedure creates a map of the associations among categories of two or more variables. It generates a map similar to that of the *Correspondence Analysis* procedure. However, unlike that procedure which compares categories of each variable separately, this procedure is concerned with interrelationships amongst the variables.

Sample StatFolio: *mca.sgp*

Sample Data:

The file *survey.sgd* contains data that describe the response of 3,418 residents of Germany to four questions about attitudes toward working women (from Greenacre, 2007). The first several rows of the file are shown below:

Respondent	*Q1*	*Q2*	*Q3*	*Q4*	*C*	*G*	*M*
1	W	H	w	w	DW	M	ma
2	w	H	H	w	DW	M	ma
3	?	H	H	w	DW	M	ma
4	?	?	?	?	DW	F	si
5	?	?	?	?	DW	F	si
6	W	H	w	W	DW	M	ma
7	?	H	H	?	DW	M	ma
8	?	?	w	?	DW	F	si
9	W	H	H	w	DW	F	ma
10	?	H	H	?	DW	M	ma
…							

Columns *Q1-Q4* identify a respondent's answer to each of four questions using the coding:

W – work full-time
w – work part-time
H – stay at home
? – no response

Column *C* identifies the country of the respondent (*DW* for West Germany, *DE* for East Germany). Column *G* indicates the respondent's gender, while column *M* indicates the respondent's marital status (married, widowed, divorced, separated, or single).

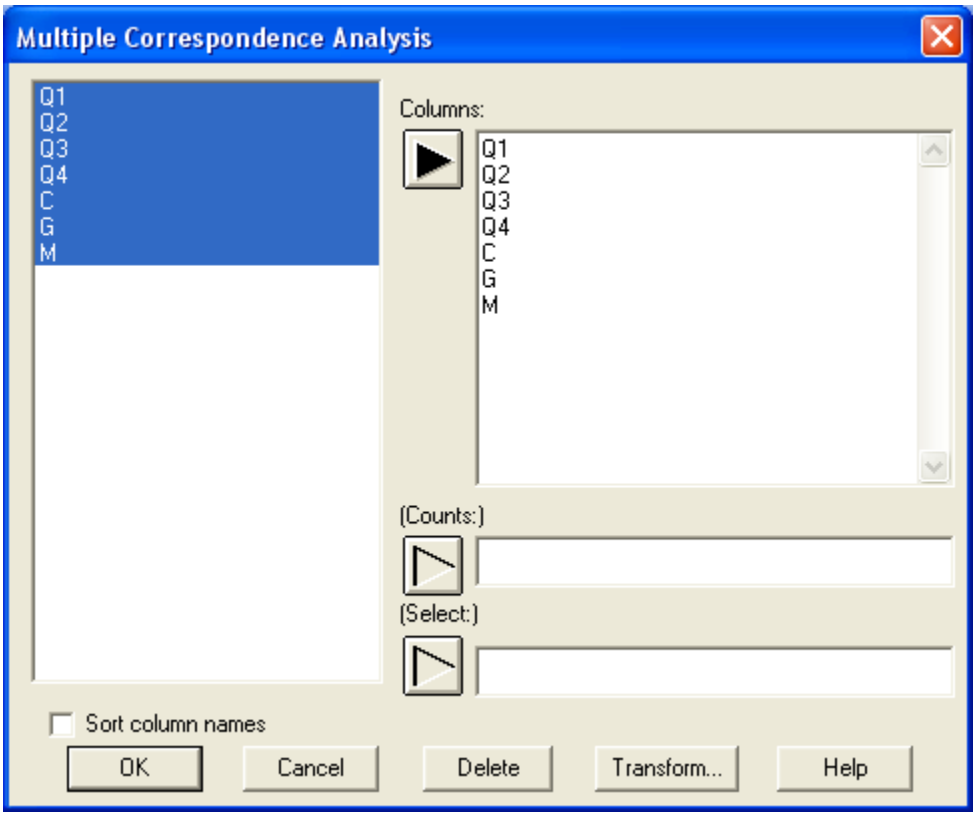
Data Input

The data for this procedure may be arranged in either of two manners:

1. A separate row may be created for each respondent, with a column for each variable of interest.

2. Rows may be created for each unique combination of categories, with an additional column indicating how often that combination occurred.

The *survey* file shown above has 3418 rows, one for each respondent. To analyze this data, complete the data input dialog box as shown below:



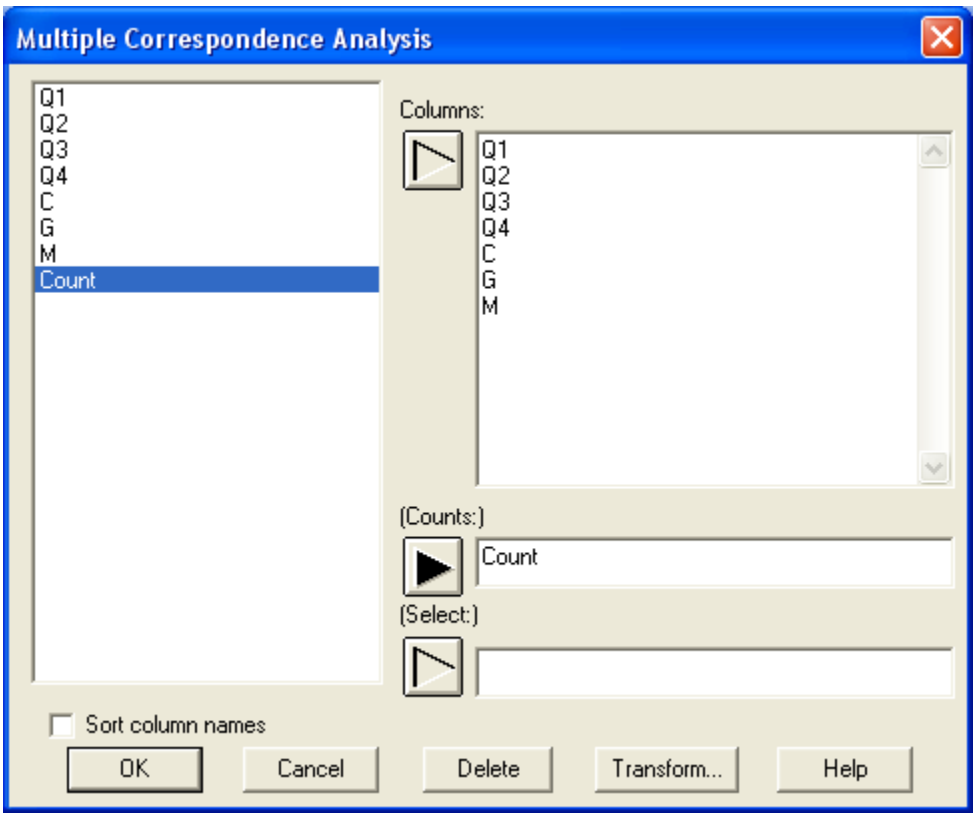
- **Columns:** names of numeric or non-numeric columns identifying the value of each variable for each case.

- **Select:** subset selection.

Alternatively, a data file could be constructed containing tabulated counts in a format similar to the file displayed below:

Q1	Q2	Q3	Q4	C	G	M	Count
W	H	w	w	DW	M	ma	1
w	H	H	W	DW	M	ma	3
?	H	H	W	DW	M	ma	1
?	?	?	?	DE	F	si	5
W	H	W	W	DE	M	se	3
?	H	H	?	DE	M	di	2
…							

This file has one row for each unique combination of the variables, with *Count* showing how often that combination occurs. To enter data in this format, complete the data input dialog box as shown below:



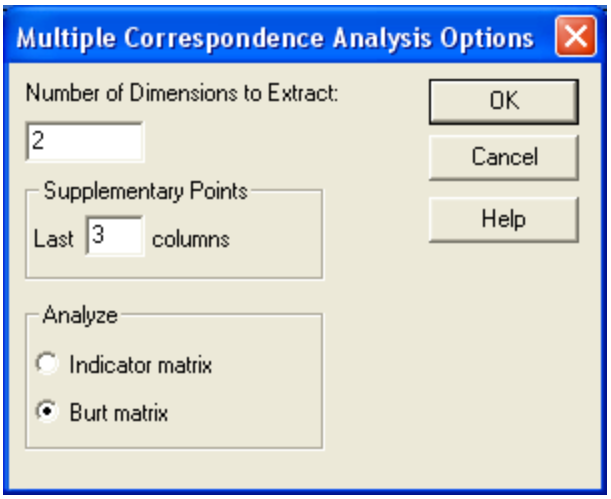
- **Columns:** names of numeric or non-numeric columns identifying each observed combination of the variables.

- **Count:** name of a numeric column tabulating how often each combination occurred.

- **Select:** subset selection.

Analysis Options

The *Analysis Options* dialog box is shown below:



- **Number of Dimensions to Extract:** the number of dimensions that you wish to extract from the data. You must specify a number between 2 and one less than the total number of categories for the variables. Often, 2 or 3 dimensions are sufficient to explain most of the differences amongst the categories.

- **Supplementary Points:** the number of columns that you wish to exclude from the calculations. The categories in these columns will be plotted on the correspondence map but will not be used to determine the scaling.

- **Analyze:** whether to perform the correspondence analysis using an indicator matrix created from the data or using the Burt matrix. Both matrices have one column for each category in each variable. The indicator matrix has a row for each respondent. The Burt matrix tabulates how often each pair of categories (one from each variable) occurs together. The resulting analysis will be similar, except that the principal inertias of the Burt analysis are equal to the squares of the inertias from the indicator matrix analysis. For details, see Chapter 18 of Greenacre (2007).

For the sample data, the scaling will be calculated from the first 4 columns. Columns *C*, *G* and *M* will be treated as supplementary.

Analysis Summary

The *Analysis Summary* displays the names of the data columns together with the Burt matrix. The output below shows the results, omitting the supplementary variables:

Multiple Correspondence Analysis

Column variables:

- Q1
- Q2
- Q3
- Q4

Inertia calculated from: Burt matrix

Burt Table

	Q1.?	Q1.H	Q1.W	Q1.w	Q2.?	Q2.H	Q2.W	Q2.w	Q3.?	Q3.H	Q3.W	Q3.w	Q4.?	Q4.H	Q4.W	Q4.w
Q1.?	362	0	0	0	196	108	1	57	204	55	7	96	264	2	51	45
Q1.H	0	79	0	0	0	72	1	6	0	61	1	17	6	38	14	21
Q1.W	0	0	2501	0	91	1131	172	1107	91	345	355	1710	157	40	1766	538
Q1.w	0	0	0	476	5	335	7	129	18	181	16	261	38	17	128	293
Q2.?	196	0	91	5	292	0	0	0	229	4	9	50	203	0	62	27
Q2.H	108	72	1131	335	0	1646	0	0	60	573	24	989	186	84	760	616
Q2.W	1	1	172	7	0	0	181	0	2	4	127	48	1	0	165	15
Q2.w	57	6	1107	129	0	0	0	1299	22	61	219	997	75	13	972	239
Q3.?	204	0	91	18	229	60	2	22	313	0	0	0	234	0	49	30
Q3.H	55	61	345	181	4	573	4	61	0	642	0	0	81	73	202	286
Q3.W	7	1	355	16	9	24	127	219	0	0	379	0	4	1	360	14
Q3.w	96	17	1710	261	50	989	48	997	0	0	0	2084	146	23	1348	567
Q4.?	264	6	157	38	203	186	1	75	234	81	4	146	465	0	0	0
Q4.H	2	38	40	17	0	84	0	13	0	73	1	23	0	97	0	0
Q4.W	51	14	1766	128	62	760	165	972	49	202	360	1348	0	0	1959	0
Q4.w	45	21	538	293	27	616	15	239	30	286	14	567	0	0	0	897

Each row and column of the Burt table shows a possible response for a selected variable. The counts in the table show how many respondents gave both the answer indicated by the row and the answer indicated by the column. For example, 196 respondents answered *?* to both *Q1* and *Q2*.

Indicator Matrix

If you have selected to analyze the indicator matrix, this pane will display that matrix. A portion of the output is shown below:

Indicator Matrix

Row	Count	Q1.?	Q1.H	Q1.W	Q1.w	Q2.?	Q2.H	Q2.W	Q2.w	Q3.?	Q3.H	Q3.W	Q3.w	Q4.?	Q4.H	Q4.W	Q4.w
1	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1
2	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1
3	1	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1
4	2	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
6	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0
7	1	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0
8	1	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0
9	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	1
10	1	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0
11	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0
12	1	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	1
13	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0
14	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1
15	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1
16	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0

The table contains one row for each unique combination of responses. For example, row 1 corresponds to the following responses:

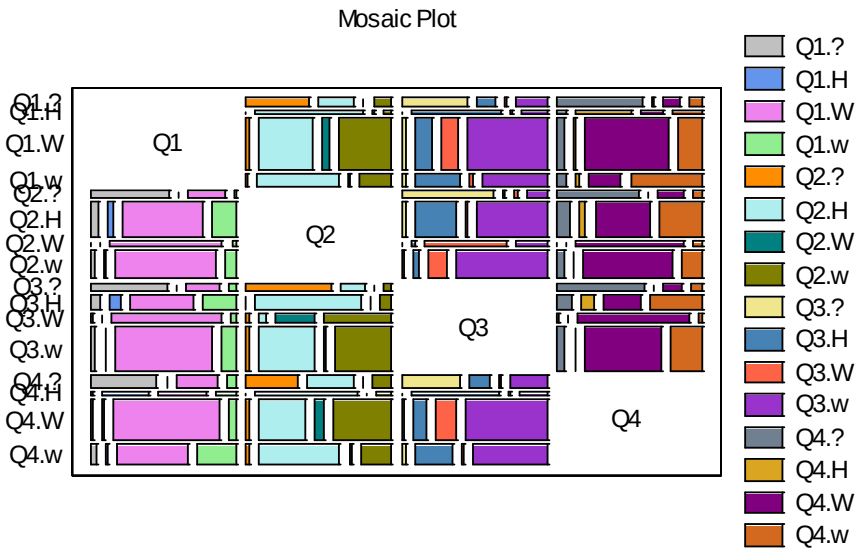
Q1 - ? Q2 – H Q3 – w Q4 – w

The *Count* column shows that a single respondent answered in that way.

NOTE: the indicator matrix that is actually analyzed has a separate row for each respondent, some of which are identical. The *Count* column is used in the output to save space.

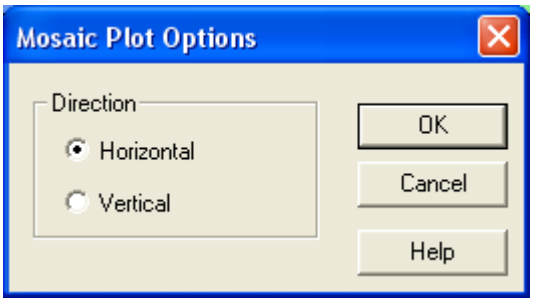
Mosaic Plot

An interesting way to illustrate the data is to plot the Burt table as a *Mosaic Plot*:



In this plot, the height of each row is proportional to the frequency of a single category. For example, the most common response to *Q1* was *W*. Within each row, the width of a bar represents the proportional distribution of the other variables. For example, most people who answered *W* to *Q1* answered either *H* or *w* to *Q2*, *w* to *Q3*, and *W* to *Q4*.

Pane Options



- **Direction:** the orientation of the bars. Since the Burt table is symmetric, the vertical plot is a rotated version of the horizontal plot.

Inertia and Chi-Square Decomposition

A multiple correspondence analysis seeks to find a small number of dimensions that describe most of the variability or inertia amongst the categories. (See the pdf file titled *Correspondence Analysis* for more details). The *Inertia and Chi-Square Decomposition* displays important information about those dimensions. The output for the sample data is shown below:

	Singular		*Chi-*		*Cumulative*	
Dimension	*Value*	*Inertia*	*Square*	*Percentage*	*Percentage*	*Histogram*
1	0.6934	0.4807	26291.2207	41.9787	41.9787	**********
2	0.5132	0.2634	14403.5727	22.9979	64.9766	*******
3	0.3647	0.1330	7273.7210	11.6138	76.5904	****
4	0.3074	0.0945	5167.9313	8.2515	84.8420	***
5	0.2176	0.0474	2589.6903	4.1349	88.9769	**
6	0.1815	0.0329	1801.9685	2.8772	91.8540	*
7	0.1648	0.0272	1484.8129	2.3708	94.2248	*
8	0.1430	0.0204	1118.2946	1.7856	96.0104	*
9	0.1363	0.0186	1016.2973	1.6227	97.6331	*
10	0.1137	0.0129	706.4424	1.1280	98.7610	*
11	0.1005	0.0101	552.1713	0.8816	99.6427	*
12	0.0640	0.0041	223.7854	0.3573	100.0000	*
TOTAL		1.1452	62629.908			

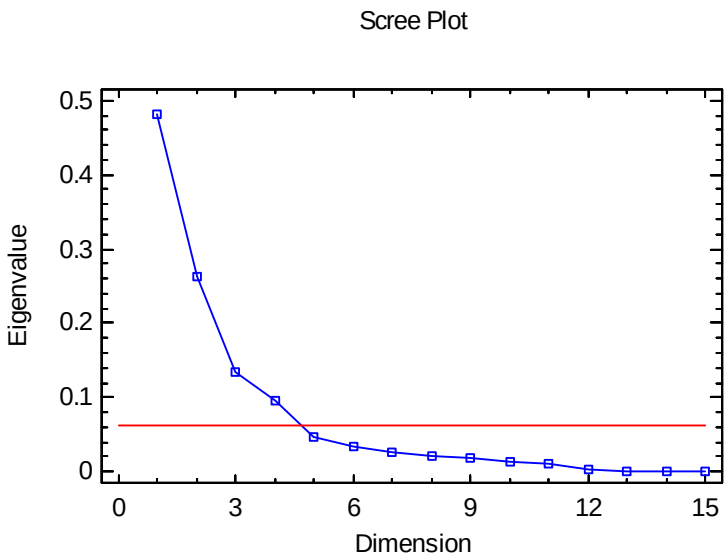
This table displays the following information for each dimension:

- **Singular Value** - the square roots of the eigenvalues of a square, symmetric matrix calculated from the Burt table or indicator matrix.
- **Inertia** - the eigenvalues of that matrix. The largest the inertia for a particular dimension, the more variability amongst the categories that dimension represents.
- **Chi-Square** – the contribution of a particular dimension to the chi-squared statistic.
- **Percentage** - the percentage of the total inertia or total chi-square statistic represented by each dimension.
- **Cumulative Percentage** - the percentage of total inertia represented by a selected dimension and those extracted earlier. Often, a small number of dimensions are sufficient to represent a large percentage of the total.
- **Histogram** – a graphical representation of each percentage.

In the current example, the first three dimensions account for approximately three-quarters of the total inertia.

Scree Plot

A useful plot in helpful in determining how many dimensions are necessary to adequately represent the data is the *Scree Plot*, which plots the eigenvalues in decreasing order:



When analyzing the Burt matrix, it has been suggested that only dimensions with eigenvalues in excess of $1/Q^2$ are interesting, where Q equals the number of variables (Q=4 for the sample data). When analyzing the indicator matrix, the cutoff equals 1/Q. The *Scree Plot* includes a horizontal line at this value. For the sample data, 4 dimensions appear to be interesting.

Category Contributions

Once the principal dimensions have been calculated, the coordinates of the categories in those dimensions can be examined. The *Category Contributions* pane for the sample data is shown below:

Category Contributions

						Dim #1			*Dim #2*	
		Quality	*Mass*	*Inertia*	*Coord*	*Corr*	*Contr*	*Coord*	*Corr*	*Contr*
1	Q1.?	0.923	0.026	0.097	1.970	0.922	0.214	-0.074	0.001	0.001
2	Q1.H	0.318	0.006	0.067	-0.194	0.003	0.000	2.038	0.315	0.091
3	Q1.W	0.740	0.183	0.022	-0.247	0.435	0.023	-0.207	0.305	0.030
4	Q1.w	0.367	0.035	0.056	-0.169	0.016	0.002	0.804	0.351	0.085
5	Q2.?	0.940	0.021	0.105	2.279	0.922	0.231	-0.318	0.018	0.008
6	Q2.H	0.709	0.120	0.039	-0.117	0.037	0.003	0.500	0.673	0.114
7	Q2.W	0.255	0.013	0.064	-0.483	0.042	0.006	-1.086	0.213	0.059
8	Q2.w	0.494	0.095	0.043	-0.297	0.169	0.017	-0.411	0.324	0.061
9	Q3.?	0.955	0.023	0.108	2.260	0.945	0.243	-0.225	0.009	0.004
10	Q3.H	0.693	0.047	0.061	-0.080	0.004	0.001	1.010	0.688	0.182
11	Q3.W	0.438	0.028	0.066	-0.433	0.069	0.011	-0.999	0.369	0.105
12	Q3.w	0.311	0.152	0.028	-0.236	0.267	0.018	-0.096	0.044	0.005
13	Q4.?	0.906	0.034	0.092	1.671	0.906	0.198	-0.003	0.000	0.000
14	Q4.H	0.327	0.007	0.067	-0.242	0.005	0.001	1.860	0.321	0.093
15	Q4.W	0.785	0.143	0.036	-0.304	0.322	0.027	-0.364	0.463	0.072
16	Q4.w	0.440	0.066	0.050	-0.177	0.036	0.004	0.596	0.405	0.088

Supplemental Columns

						Dim #1			*Dim #2*	
		Quality	*Mass*	*Inertia*	*Coord*	*Corr*	*Contr*	*Coord*	*Corr*	*Contr*
17	C.DE	0.860	0.080	0.017	-0.188	0.147	0.006	-0.413	0.712	0.052
18	C.DW	0.860	0.170	0.008	0.089	0.147	0.003	0.195	0.712	0.025
19	G.F	0.763	0.124	0.001	-0.034	0.151	0.000	-0.068	0.611	0.002
20	G.M	0.763	0.126	0.001	0.033	0.151	0.000	0.067	0.611	0.002
21	M.di	0.772	0.015	0.000	-0.077	0.327	0.000	-0.090	0.446	0.000
22	M.ma	0.633	0.159	0.001	-0.042	0.460	0.001	0.026	0.173	0.000
23	M.se	0.390	0.004	0.000	-0.143	0.390	0.000	-0.004	0.000	0.000
24	M.si	0.826	0.051	0.002	0.144	0.468	0.002	-0.126	0.359	0.003
25	M.wi	0.748	0.021	0.001	0.043	0.043	0.000	0.174	0.705	0.002

The output displays:

- **Quality** – a measure of how well a category can be represented by the number of dimensions that have been extracted relative to its representation using all dimensions. The closer quality is to 1, the better the representation.

- **Mass** – the proportion of data in each category.

- **Inertia** – the relative inertia of each category. The values sum to 1 for those categories that were used in the calculations. For supplemental categories, their size relative to the other categories is of interest.

- **Coord**. – the principal coordinate or location of the category along the indicated dimension. These values may be plotted using either the *Correspondence Map* or the *Category Coordinate Plot*.

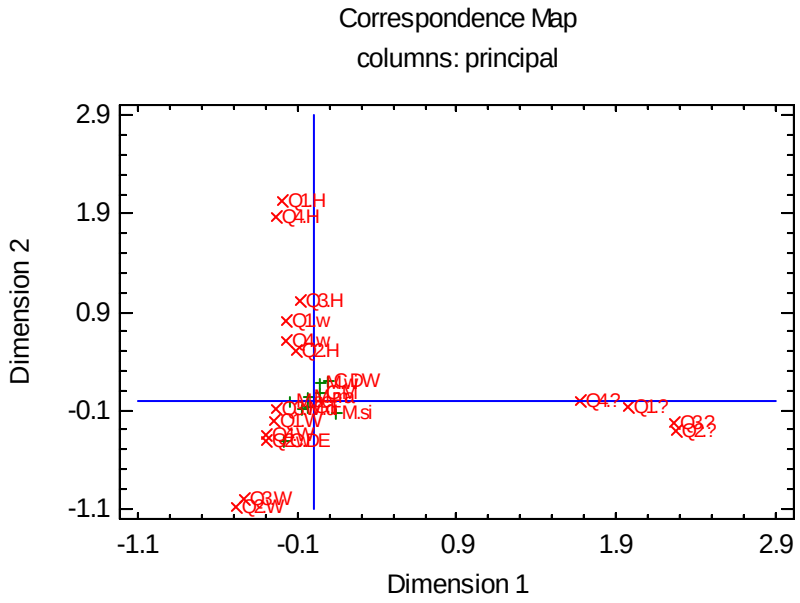
- **Corr**. – the correlation between a selected category and the axis defining a given dimension.

- **Contr**. – the relative contribution of an individual category to the inertia of a selected dimension. The higher the contribution, the more that category contributes to variability along that dimension.

For example, *Q2.?* has the highest contribution to dimension #1. On the other hand, the quality of the representation for *Q2.W* based upon the two dimensions is relatively low.

2D Correspondence Map

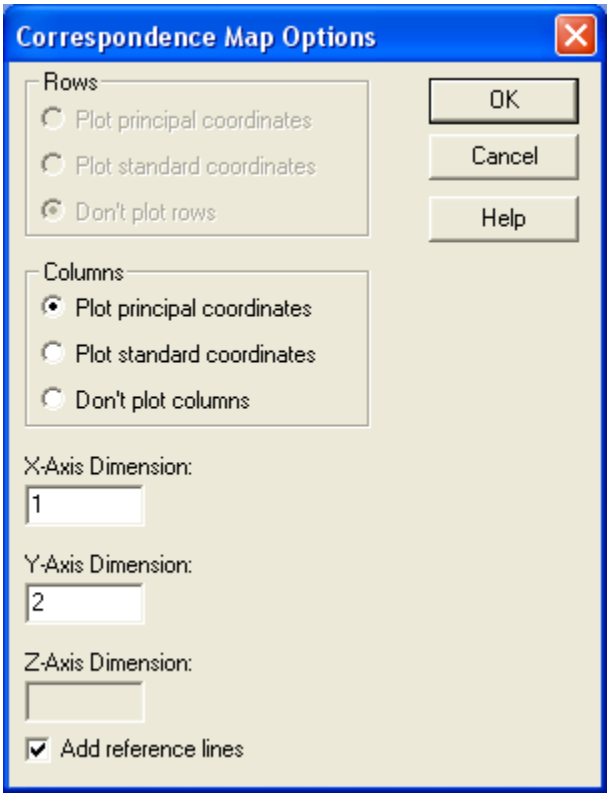
The coordinates of the caetgories may be plotted for any 2 dimensions by selecting the *2D Correspondence Map*. By default, the coordinates for the first two dimensions are displayed:



Different point symbols are used for the categories used in the calculations and for the supplementary points.

For the sample data, note that dimension 1 constrasts the no response (“?”) categories with the others.

Pane Options



- **Rows** – the type of row coordinates to plot, if any. The weighted average of the squared principal coordinates equals the eigenvalue of that dimension, while the weighted average of the squared standard coordinates equals 1. This option is only available when analyzing the indicator matrix rather than the Burt matrix.

- **Columns** – the type of column coordinates to plot, if any.

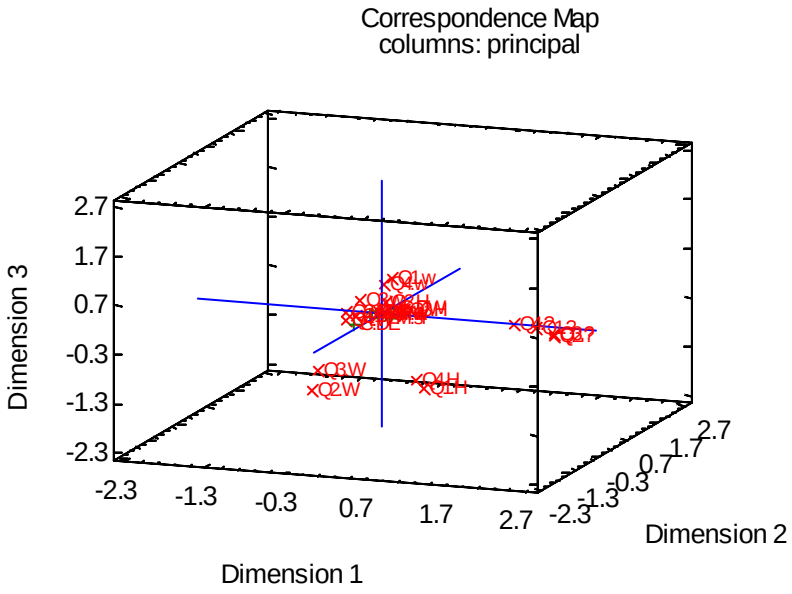
- **X-Axis Dimension** – the dimension to plot along the horizontal axis.

- **Y-Axis Dimension** – the dimension to plot along the vertical axis.

- **Add reference lines** – whether to add vertical and horizontal lines through the origin.

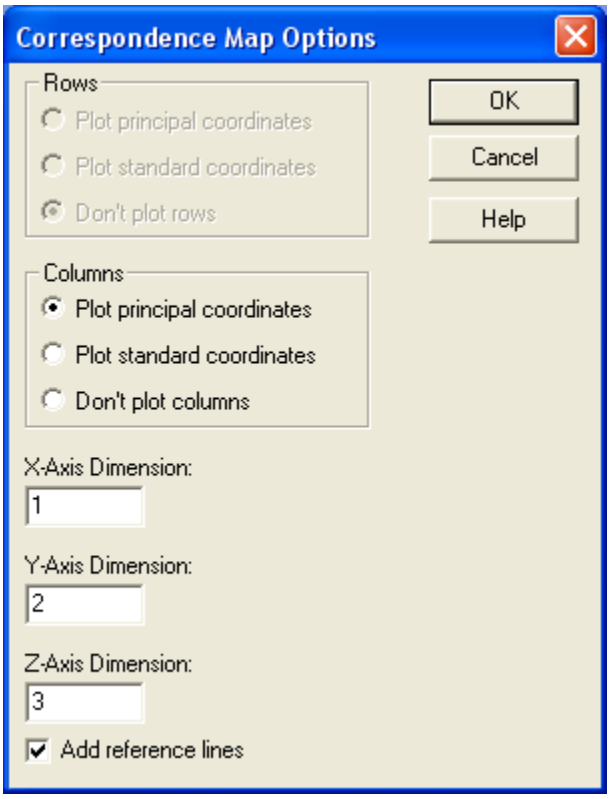
3D Correspondence Map

The coordinates of the categories may be plotted for any 3 dimensions by selecting the *3D Correspondence Map*. By default, the coordinates for the first three dimensions are displayed:



When viewing this plot, it may be helpful to use the graphics pan and zoom feature.

Pane Options



- **Rows** – the type of row coordinates to plot, if any. This option is only available when analyzing the indicator matrix rather than the Burt matrix.

- **Columns** – the type of column coordinates to plot, if any.

- **X-Axis Dimension** – the dimension to plot along the X axis.

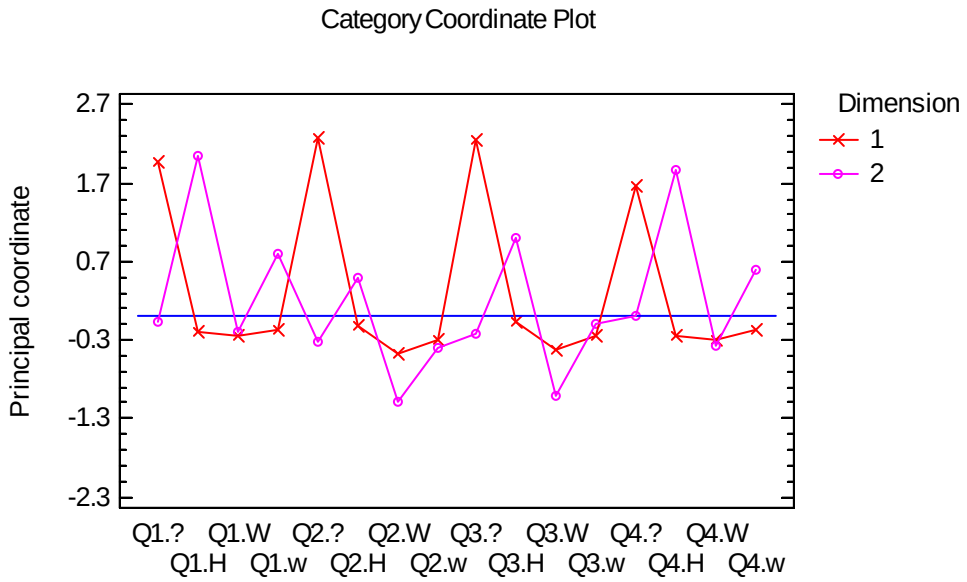
- **Y-Axis Dimension** – the dimension to plot along the Y axis.

- **Z-Axis Dimension** – the dimension to plot along the Z axis.

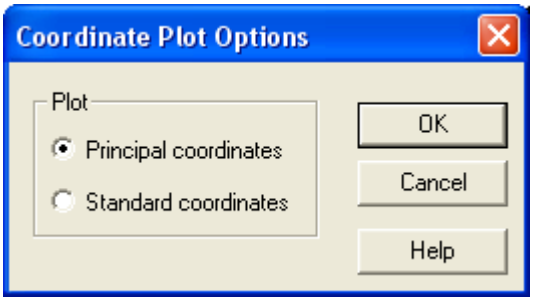
- **Add reference lines** – whether to add reference lines through the origin.

Plot of Category Coordinates

This pane displays the category coordinates for each extracted dimension:



Pane Options



- **Plot** – the type of coordinates to display.

Save Results

You may save the following results to the datasheet:

1. **Point Labels** – labels identifying each category.
2. **Point Quality** – the quality of the representation for each category.
3. **Point Mass** – the mass of each category.
4. **Point Inertia** – the relative inertia of each category.
5. **Principal Coordinates** – the principal coordinates for each category, for each extracted dimension.
6. **Standard Coordinates** – the standard coordinates for each category, for each extracted dimension.
7. **Point Correlations** – the correlations for each category, for each extracted dimension.
8. **Point Contributions -** the contributions of each category, for each extracted dimension.

Calculations

The procedure performs a correspondence analysis on either the indicator matrix or the Burt matrix. The details of the calculations are contained in the pdf document titled *Correspondence Analysis*. The counts n_{ij} are replaced by the cells of the selected matrix.

