Exploring Numerical and Textual Data in Practice with Dtm-Vic

(Version 6 of Dtm-Vic)

Ludovic Lebart

Marie Piron

ISBN 978-2-9537772-1-5

Table of Content

Dtm-Vic

Data and text mining

Visualization, inference, classification

Multivariate Descriptive Statistical Analysis Software for Numerical, Categorical and Textual Data

Freely downloadable from: www.dtmvic.com

Introduction

Dtm-Vic is a program devoted to the Multivariate Descriptive Statistical Analysis of numerical and textual data.

The exploratory analysis, as its name suggests, is a preliminary phase for getting acquainted with a data set. Exploration assumes that the data are complex, that a priori knowledge on these data is limited.

The multivariate analysis applies to cases where the dimensions (most often: the variables) are numerous, which is a factor of complexity, and therefore an incentive to start with an exploratory approach. Another more technical incentive to "explore" concerns the unrealistic assumptions of distributional statistics in the multidimensional case.

The exploratory analysis of multidimensional numerical data will be an important facet of Dtm-Vic. The basic tools are on the one hand principal axes analyses such as principal components analysis, simple and multiple correspondence analyses, on the other clustering methods (hierarchical clustering methods, partitioning, self-organizing maps). These techniques are systematically used as complementary tools providing several points of view on the statistical reality.

Textual data are both multidimensional and complex. They are therefore likely candidates for treatment given by the exploratory analysis. They are often associated with numerical or categorical data. It is the emblematic case of sample surveys including both closed questions (numerical and categorical data) and open questions (textual data). These survey data are a typical example around which grew Dtm-Vic. An important part of the methods used in the textual component of the software Dtm-Vic are presented and discussed in the book "Exploring Textual Data".

The exploratory analysis of multidimensional numerical data and text appears as an inevitable phase of treatment of these complex collections.

Explorers often discover something other than what they are looking for. Dtm-Vic users often have the opportunity to check that assumption. The exploratory analyses carried out are formidable tests of both consistency and quality of basic information. The results of these tests are neither necessarily pleasant for

those who have gathered this information, nor to those who have used it too quickly.

But for experienced users, and particularly for social scientists, these tests of global coherence are not accidental byproducts but one of their fundamental objectives, explicitly inserted in a critical approach that considers *a priori* a data set as a disputable construction.

* *
*

The current limitations of the software (revisable) regarding the size of the input data are as follows: 30,000 lines (individuals, observations), 1,000 columns (numerical variables, categories), 100 000 characters for the verbatim responses of an individual / observation. This format corresponds to the vast majority of applications to socio-economic surveys, management and satisfaction surveys, ecological surveys, sensory analyses, etc.. Note that a direct analysis of texts can ignore these limitations.

* *
*

After a brief introduction to the software (Chapter I), Chapter II presents three examples of analyses involving only numerical and categorical data, whereas Chapter III deals with three examples of analyses of textual data. The six examples of both these chapters are carried out on already prepared data sets, that is to say, data presented in a suitable format (*internal Dtm-Vic format*). All the example data sets are supplied with the software. These examples correspond to frequent use of Dtm-Vic. The user will learn to create by him/her/self a command file from the proposed interface. We find successively in chapter II a principal component analysis (linked with a clustering, and an automatic description of the clusters), a simple (or: two-way) correspondence analysis, a multiple-correspondence analysis (also supplemented by a classification). The first example of chapter III deals with a lexical correspondence analysis of a series of text. The second example, in the context of sample surveys data, carries out another lexical correspondence analysis of a contingency table built from an open question and a closed question. Finally, Chapter III ends with a direct analysis of responses to an open-ended question. Most applications of these two chapters result in visualizations validated by the *bootstrap* technique.

Chapter IV consists of three new types of examples relating to textual data, for advanced users who desire to edit directly the command file. A first example concerns again the direct processing of open-ended questions in a sample survey, without using closed questions. The second example shows how the

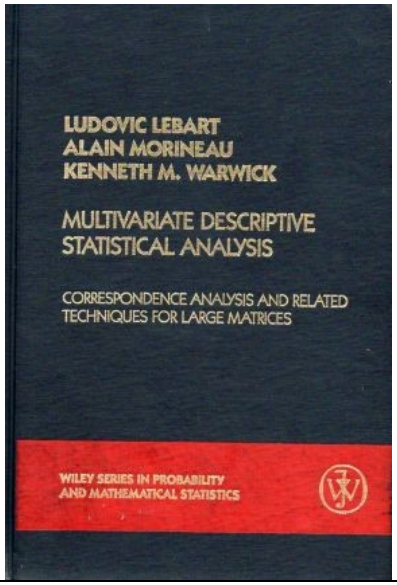
results of a Multiple Correspondence Analysis can be illustrated by the responses to open questions. The third example, making use of a thesaurus, deals with the visualization of the Semantic network of French verbs.

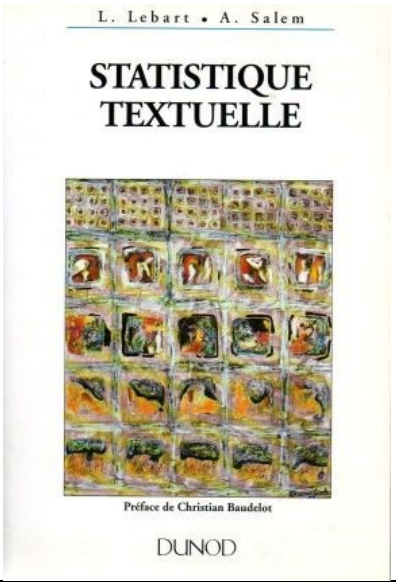
Chapter V presents more detailed applications concerning numerical data including the implementation of new options for visualizing these data. After dealing with *Semiometric data*, this chapter discusses the technique of Contiguity Analysis (applied to the famous Fisher's "Iris Data"). The following examples concern the descriptions of graphs, and eventually the ability of principal axes techniques to compress some images.

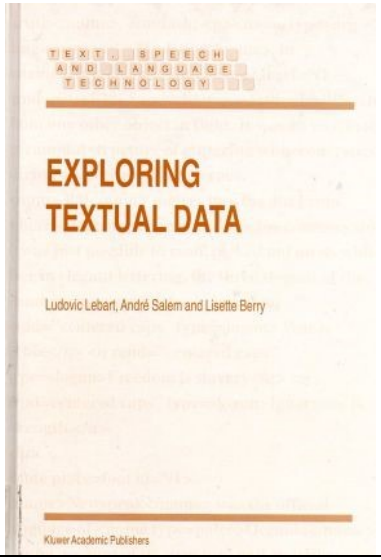
Finally, Chapter VI presents some procedures for importing external data. One can easily imagine that dealing with statistical units as disparate as a number, a category, a terse response to an open question, or a novel by Dickens can sometimes lead to inextricable situations. However, the full transparency of both input files and files generated by Dtm-Vic (all files are in non-proprietary text format) should (perhaps) comfort the user and limit the complexity of the process.

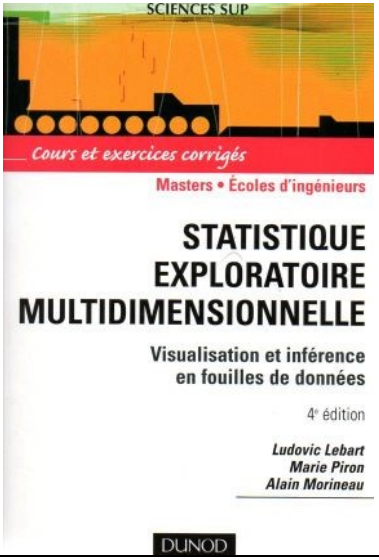
All these phases of learning assume that the software and the collection of examples have been copied or downloaded, which is possible from the website: http://www. DtmVic.com.

Four books describing approximately the perimeter of the package Dtm-Vic.









Chapter I

Overview of Dtm-Vic

To start running *Dtm-Vic*, just click on the shortcut icon:  put on the desktop of the computer by the user. We obtain the following screen (Fig. 1):

Figure 1. First screen: General information, Tutorial, Reminders…

Content: General information about the software (button "About DtmVic), about the internal Data format ("Data Format"), Tutorials based on the examples downloaded with DtmVic, ("Tutorials"), references and reminders. A gallery of portraits pays tribute to some pioneers of Multivariate Descriptive Statistical Analysis (*avant la lettre…*).

Clicking on "Start DtmVic", we obtain the second screen, similar to the first screen of the previous version of DtmVic (Fig. 2).

After the sub-menu Dtm-Vic: –Data Importation, Dtm-Vic is structured into two main sections:
I - The first section under the headline: Dtm – Data and Text Mining includes procedures for setting up the data (importation, capture, export) and data analysis procedures (creation and execution of the command file).
II - The second section: VIC – Visualization, Inference, Classification provides tools for visualization, validation and interpretation of results.

One can also see on the home screen two optional headings: the "toolbox" DtmVic Tools which offers different types of recoding, data storage, and the specific (essentially pedagogical) section DtmVic Images devoted to some particular image analyses.

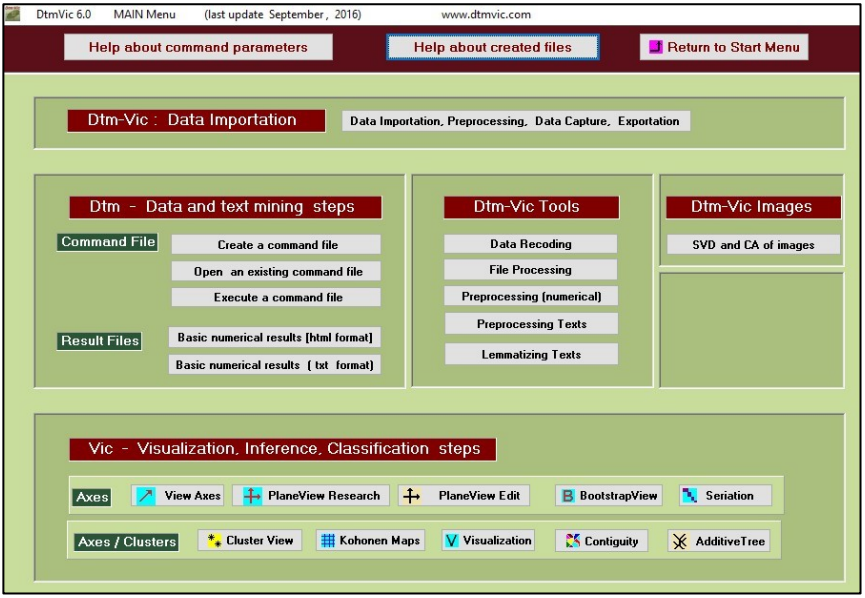


Figure 2. Main Screen of DtmVic

This manual should help to make an implementation of these different steps. Most analyses comprise the following sequence of steps:

1. Selection of the type of analysis
2. Opening of the various data files in Dtm-Vic format.

- Selection of variables (active, supplementary, discarded)
- Selection of the parameters specific to the analysis.

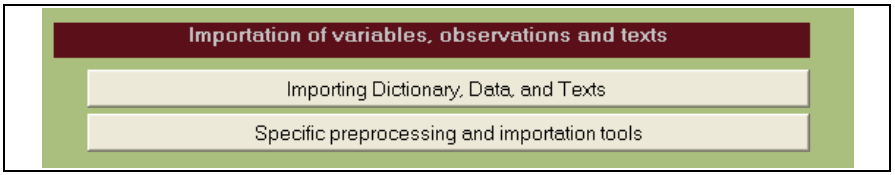
3. Creation of a command file
4. Execution of the command file
5. Display of results.

For assistance about the parameters or about the created files, click the Help menus in the top bar. A tutorial is available on this bar.

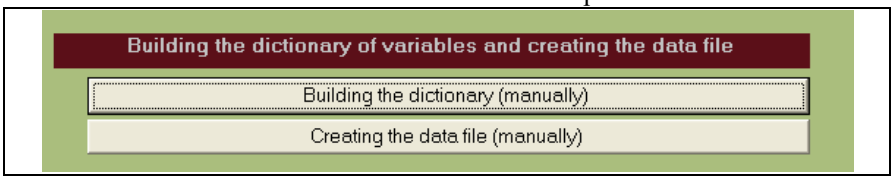
I.1 Setting up data files:

➢ Click on Data Importation, Preprocessing, Data Capture, Exportation in the section **Data File**.

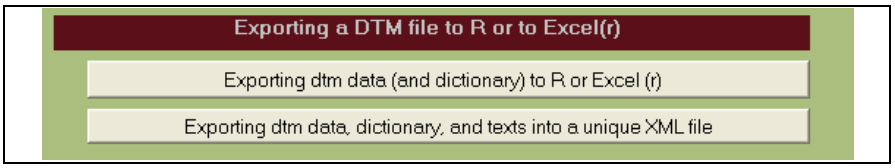
A window is displayed, suggesting various procedures. Here are the components of this window:



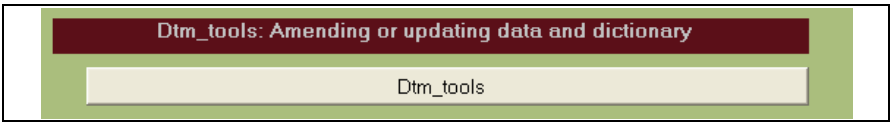
These items are dealt with in chapter VI



The capture of data (mainly for small-sized exercises for students) are described in chapter VI.



For exporting data files in Excel, XML or R , see Chapter VI



Creating new variables, selection of a sub-sample or concatenation of multiple files. There is a direct access to this section from the heading DtmVic Tools in the main menu.

I.2 Data Analysis Methods

Click on: Create a Command File in the section: **Command File.**

A window appears, displaying different available techniques. The upper part of this window deals with numerical or categorical data:

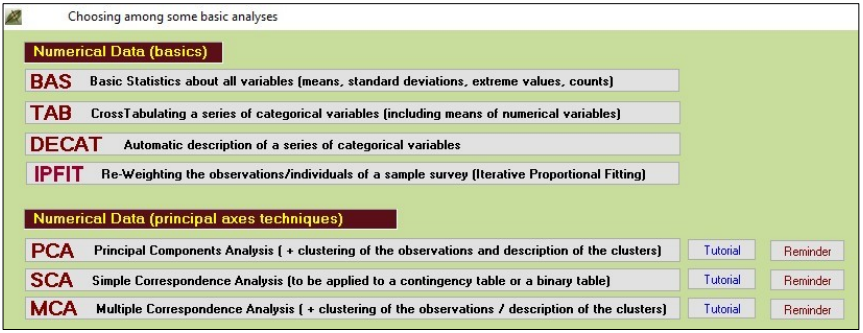


Figure 3. Univariate and multivariate basic processing.

Summary of treatments:

Univariate or elementary treatments.
BAS: Basic descriptive statistics;
TAB: Cross-tabulation;
DECAT: Automatic description of the categories of a categorical variable.
IPFIT: Iterative Proportional Fitting. Creating new weights for the observations.

Multivariate processing:

PCA: Principal Components Analysis,
SCA: (Simple) Correspondence Analysis,
MCA: Multiple Correspondence Analysis
These three principal axes methods are complemented with clustering (k-means clustering combined with agglomerative methods). (See Chapter II).

The bottom of the window (Fig. 4) concerns the processing of textual data (exploratory statistical analysis of a corpus of texts):

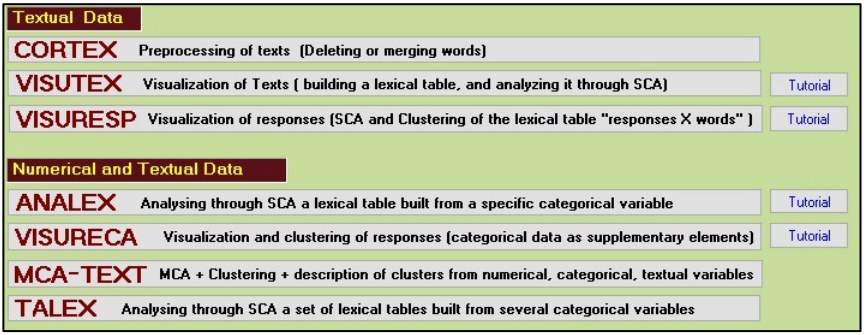
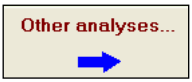


Figure 4. Processings involving textual data

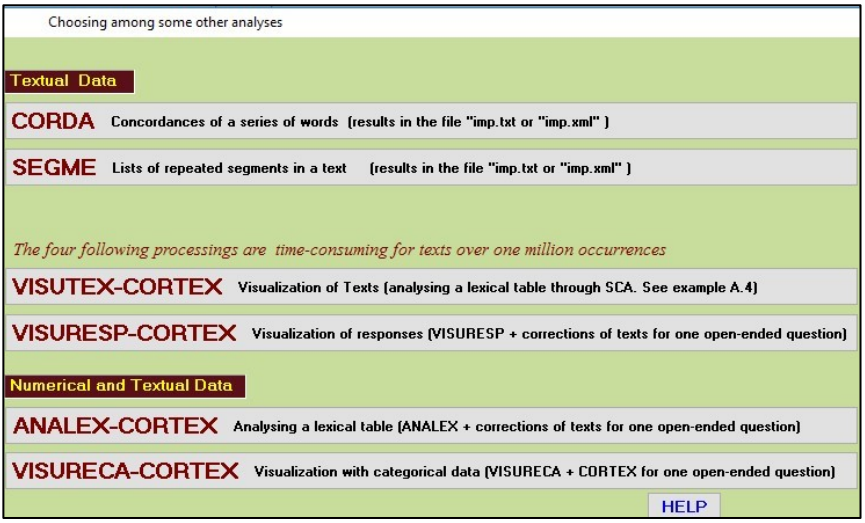
CORTEX: removes words, or agglomerates words (empirical lemmatization)
VISUTEXT: performs a simple correspondence analysis of a lexical table (rows = words, columns = texts or categories; see chapter III);
VISURESP: performs a direct analysis of responses to open-ended questions (rows = observations or respondents, columns = words; see chapter IV).
ANALEX: performs a simple correspondence analysis of lexical aggregated table (rows are the words, columns the categories of respondents corresponding to a selected categorical variable)(see chapter III);
VISURECA: performs an analysis similar to VISURESP, but illustrates it with categorical variables (see chapter III).
MCA-TEXT: performs a multiple correspondence analysis (complemented with a clustering of observations) on a set of categorical variables. The clusters and the visualizations as well are illustrated by lexical variables (words used in the responses to an open question).
TALEX: performs a simple correspondence analysis of a composite lexical table. This lexical table is obtained by juxtaposing several lexical tables: As for

ANALEX, the row elements are still the words, but the columns correspond to several categorical variables (instead of a unique categorical variable in the case of **ANALEX**).

Other techniques of textual analysis are available in the menu



If you click on this button, a new window appears (Fig.5).



CORDA and **SEGME** provide lists of concordances and repeated segments, while the following analyses include the processing **CORTEX** (text corrections) in the previous analyses VISUTEX, VISURESP, VISURECA, ANALEX.
VISUTEXT-CORTEX execute the step **VISUTEX** after corrections of texts (**CORTEX**).
VISURESP-CORTEX execute the step **VISURESP** after text corrections (**CORTEX**)
ANALEX- CORTEX execute the step **ANALEX** after corrections of texts (**CORTEX**)
VISURECA - CORTEX execute the step **VISURECA** after corrections of texts (**CORTEX**)
It is of course possible to use independently **CORTEX** and, afterwards, one of the analyses on the created files. But **CORTEX** covers the whole text file, and

sometimes one may wish to address individually each open-ended question. Moreover, in the composite analyses above, the modal responses, (typical responses for each text) will be the original raw responses, not the corrected responses (with missing or grouped words).

Once the command file is created in the Create a command file procedure, it is possible, always under the heading: **Command File** to open the file directly (button: Open an existing command file) to directly modify certain parameters, save, and then execute (button : Execute a command file).

The procedures for exploratory analysis of numerical data or textual involve the concatenation of several techniques, PCA, CA, MCA, clustering, Kohonen Maps, Bootstrap validation. The numerical results of basic analyses can be viewed in either the heading: **Result Files** (Basic numerical results) allowing for navigating into a HTML file, in or text format (text format). Then, they can be visualized by the various tools offered under the heading: **VIC: Visualization, Inference, Classification**.

I.3 Visualization of results

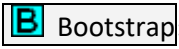
Under the heading: **VIC: Visualization, Inference, Classification**, a series of visualization tools serve to validate the results and facilitate their interpretation (see Chapters II and III).

To use these tools, click on the following buttons:

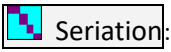
: Principal axes taken individually.
For each axis, the coordinates of individuals, active variables, supplementary variables, are sorted for a rapid assessment of the results from the principal axes analyses.

: Graphical displays of principal planes (pairs of axes). Description of axes for all types of elements involved in the analysis (up to 30,000 elements).

: Graphical displays of principal planes (pairs of axes). [with moveable tags] (up to 900 elements).



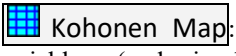
: Confidence areas (ellipses or convex hulls) in factorial designs for selected items.



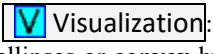
: The rows and columns of the contingency table are reordered according to the first axis of correspondence analysis of the table.

[Seriation techniques are based on simple permutations of rows and columns of the table studied. They have the practical cognitive advantage of showing the raw data (merely reordered) to the user. He/she does not have to learn complex rules of interpretations. These permutations can highlight homogeneous blocks in the contingency table. They can also pinpoint a gradual and continuous evolution of the column-profiles or the row-profiles. An optimal property of correspondence analysis is as follows: the first axis of correspondence analysis provides an optimal order for rows and columns of a contingency table.]

: Projection of centroids (mean points) of clusters on the plane spanned by pairs of principal axes. Description of the characteristic features for each cluster (numerical variables, categories, and also characteristic words or responses in the case of open questions).



: Kohonen maps. Self-organizing maps of individuals, variables, (and simultaneous representation of individuals and variables) (square grids of dimensions 3 x 3 up to 20 x 20).



: Additional graphical displays of axes and clusters. Density ellipses or convex hulls for the clusters. Plot of the minimum spanning tree, of the nearest neighbours on the principal planes. Visualization of the gradual computation of clusters (for procedure k-means / dynamic clustering). Visualization of Kohonen grids and of some particular graphs.

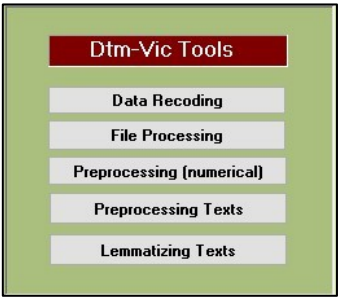
: Analysis of Contiguity. Local analysis, graph structure.

The contiguity analysis is a “local analysis technique” that is briefly presented in Chapter V. It considers the case in which observations have an *a priori* graph structure, but also the case in which the graph is intrinsic (i.e.: derived from the data themselves: e.g.: graph of nearest neighbours). It generalizes the Linear Fisher Discriminant analysis (particular case of a graph associated with a partition). Contiguity analysis is discussed in Section V.2 of Chapter V.

Building additive trees from the software SplitsTree [1]

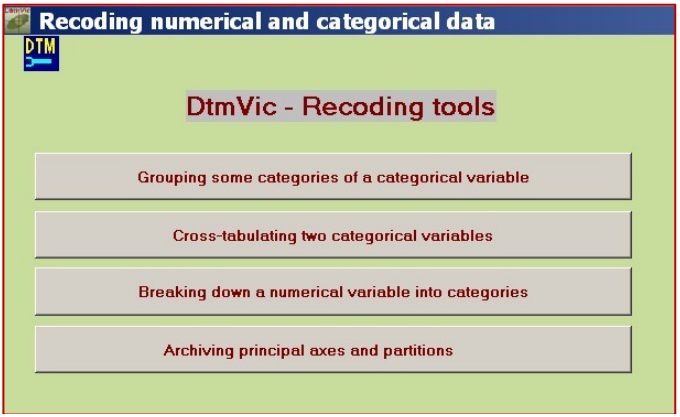
I.4. The toolbox

The toolbox: **DtmVic Tools** offers a variety of recoding, storage and processing of data.



1.4.1 Click on: Data Recoding

A first group of recoding is displayed:



[1] Huson D.H. and Bryant D. (2006). Application of Phylogenetic Networks in Evolutionary Studies, *Molecular Biology and Evolution*, vol. (23), 2: 254-267.

Creating or recoding nominal (categorical) variables:
1) Aggregation of categories of a variable.
2) Creation of a categorical variable by cross-tabulating two categorical variables;
3) Conversion of a continuous variable into a categorical variable;
4) Storage of principal axes and partitions.

The second group enables some elementary actions on the data file:

1.4.2 Click on: File Processing

Five buttons are displayed:

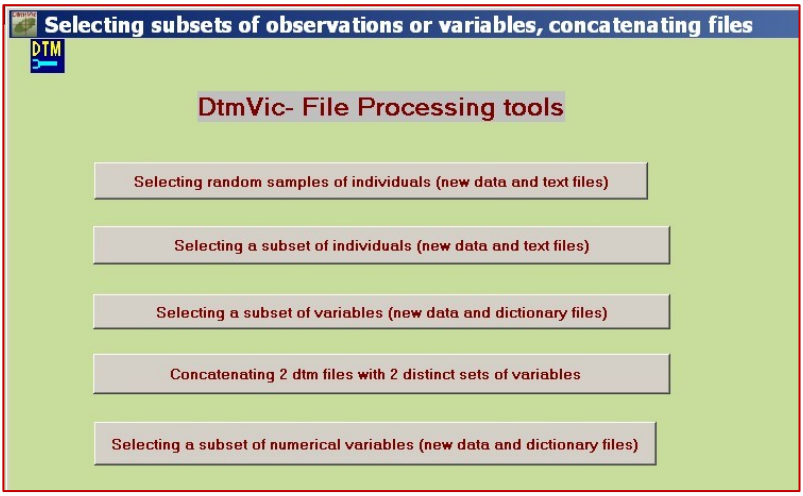
1) Selecting a random subset of individuals (lines); (to deal with a smaller sample, or to check, on different subsets, the stability of the results)

2) Selecting a specified subset of individuals (lines);

3) Selecting a subset of variables (columns);

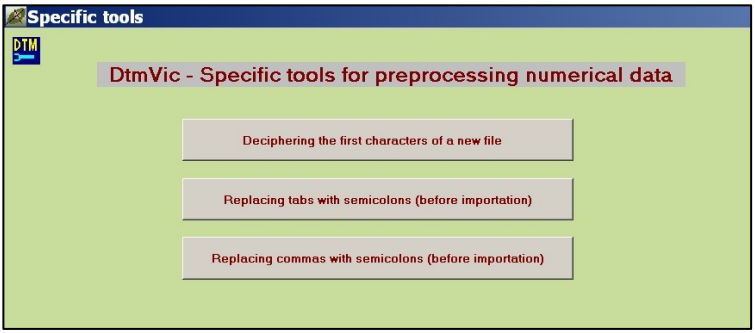
4) Concatenation of two databases (different variables, same individuals [rows]);

5) Selecting a subset of variables with largest weights (weight = sum of the values over the individuals).



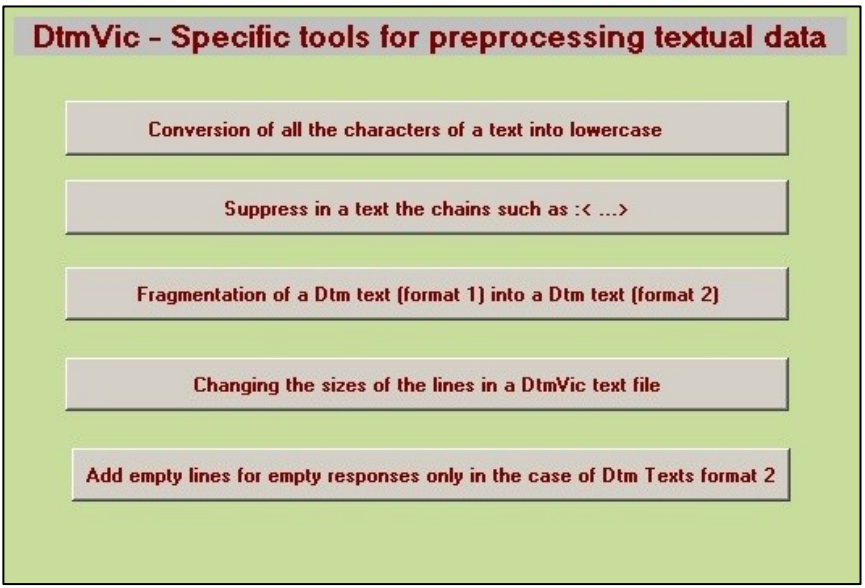
A third button in the small panel " DtmVic Tools" concerns a few basic tools for some elementary pre-processing of the data. These tools may be useful before an importation step.

1.4.3 Click on: Preprocessing (numerical data)



The last button of the small panel " DtmVic Tools" concerns the preprocessing of textual data.

1.4.4 Click on: Preprocessing texts



1) Conversion into lowercase of all the characters of a text .

2) Remove the " < " and ">" tags together with text that they may contain.

3) Fragmentation of a series of texts in format number 1 (texts separated by ****) into texts in format 2 (texts separated by ----). The new texts are small fragments (context units) of the initial text. They may consist of one line, two lines, ... of the initial text.). A dummy variable is created in a new Dtm-Vic data file (with its dictionary) to connect the numerous new texts to the original texts.

4) Change in the length of lines of text. We start with a DtmVic text file (format 1 or 2) without limitation for the length of the lines. At the end: text file whose lines have with a user-selected length (but < 200 characters). This procedure allows you to import text with very large lines, but also to format the context units (see paragraph 4 above) .

5) The latest (limited and specialized) procedure to enforce the constraint " an empty line for an empty open-response" for files that would use two consecutive delimiters (to be used after the re-importation of paragraph 4 above.

1.4.5 Click on Lemmatizing texts.

This step re- import a DtmVic text file (type 1 or 2) via the (free) software WinTreeTagger[2] . This allows for lemmatising a text and also for removing certain grammatical categories (prepositions , articles, ...) . Valid for the following languages: English, French, Spanish, Italian.

Under the heading: **DtmVic Images**, a series of processing shows the potential for image compression offered by correspondence analysis or simply by the singular value decomposition, with comparison with Discrete Fourier Series (section V.4 of Chapter V).

---

[2] Schmid H. (1994). Probabilistic part of speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

I.5 Internal DtmVic format for input data and texts

The aim of the importation procedures (Chapter VI) is to transform a pre-existing text file into the "Internal DtmVic format". The knowledge of the internal DtmVic format could be useful to some advanced users; it is not indispensable for the beginners (see also the button "Data Format" in the first menu).

Let us remind that DtmVic is a software devoted to exploratory analysis of multivariate numerical and textual data. The leading case that exemplifies all the possibilities of the software is a sample survey data set, comprising both responses to closed questions and responses to open-ended questions (the closed questions may lead to numerical [quantitative] or categorical [qualitative] data).

In the most general configuration, three files constitute the internal DtmVic input data set:

1) The **dictionary file** that provides the names (or identifiers) of the numerical and categorical variables It includes the names of the categories corresponding to each categorical variable. That latter feature is rather uncommon in statistical software, but seems indispensable to explore high dimensional categorical data sets.

2) The **data file** , that contains the values of these variables for a set of individuals (or: observations), together with the identifiers of the individuals.

3) The **text file** made (e.g.) of the responses to open ended questions. The text file (known as text file type 2) concerns the same respondents as those of the data file, in the same order. A simplified "text file format" (text file type 1) can be used when dealing only with a series of texts, without associated data file and dictionary file.

Some applications may involve only the text file (see for instance the example A4 of Chapter III), whereas others may need only the dictionary and the data files (application examples A1, A2, A3, C1, C2, of Chapters II and Chapter V).

Description of the internal "DtmVic format"

The format is specific, but not proprietary: The three types of files are in simple text format (extension ".txt", readable through a "notepad" or a text editor, or also with a word processor, provided that they are saved as simple text files).
As an introductory exercise, they can be recorded directly from the keyboard, or with the help of the menu "DataCapture" (see preliminary example D.0 in chapter VI below).

In most cases, however, they have to be imported from (often large) pre-existing files. The transformation into DtmVic format is then transparent to the user.

Table 1 shows an example of a small DtmVic dictionary, involving four variables.

Table 2 displays an example of a DtmVic data file (same four variables, three individuals or respondents).

Table 3 presents a text file (text format DtmVic 1) relating to a simple series of texts.

Table 4 presents a text file relating to three open-ended questions and three respondents (text format DtmVic 2).

Table 1: Example of an internal DtmVic dictionary for 4 variables:

Gender (2 categories); Age (0 categories = numerical variable); Age broken down into 4 categories; Educational level (3 categories). [fixed format, comments in *italic, blue. Note that column 6 is empty*].

```
   2 GENDER  (number of categories [2] in columns 1-4; blank; title)
MALE MALE  (short identifier [column 1-4]; blank; identifier [< 20]
FEMA FEMALE  (short identifier [column 1-4]; blank; identif. [< 20]
   0 AGE (number of categories [0] in col. 1-4; blank; numerical)
   4 AGE_CODE (number of categ. [4] in columns 1-4; blank; title)
AGE1 18_24  (short ident.[column 1-4]; blank; ident.[< 20 char]
AGE2 25_39 (short ident.[column 1-4]; blank; ident.[< 20 char]
AGE3 40_59  (short ident.[column 1-4]; blank; ident.[< 20 char]
AGE4 >60  (short ident.[column 1-4]; blank; ident.[< 20 characters]
   3 EDUCATION  (number of categories [3] in col 1-4; blank; title)
EDUL LOW   (short ident.[col 1-4]; blank; ident.[< 20 characters]
EDUM MEDIUM (short ident.[column 1-4]; blank; ident.[< 20 char]
EDUH HIGH   (short ident.[column 1-4]; blank; ident. [< 20 char]
```

Table 2: Example of an internal DtmVic data file for the previous 4 variables: Gender, Age broken down into 4 categories, Educational level. 9 respondents (individuals, observations)

```
  '1006' 1 76  12 1
  '1007' 2  20 2 2
  '1008' 2  29 3 2
  '1012' 1 77  10 1
  '101A' 2  20 2 2
  '1028' 2  29 3 3
```

Identifiers of the respondents: between quotes, less than 20 characters. No blank space, no quotes within the identifiers. Separators between values: at least one blank space [free format].

Table 3: Example of an internal DtmVic text file (type 1) for three texts (see: application example EX_A04.Text.Poems of Chapter III).

Free text format on less than 200 columns. Separator of texts : "**" followed, after four blank spaces, by the identifier (<= 20 characters); End of file: "====". All separators are in columns 1-4.**

```
****    S01_Sonnet_1
from fairest creatures we desire increase,
that thereby beauty's rose might never die,
but as the riper should by time decease,
his tender heir might bear his memory:
but thou, contracted to thine own bright eyes,
feed'st thy light'st flame with self-substantial fuel,
making a famine where abundance lies,
thyself thy foe, to thy sweet self too cruel.
thou that art now the world's fresh ornament
and only herald to the gaudy spring,
within thine own bud buriest thy content
and, tender churl, makest waste in niggarding.
pity the world, or else this glutton be,
to eat the world's due, by the grave and thee.

****    S02_Sonnet_2
when forty winters shall beseige thy brow,
and dig deep trenches in thy beauty's field,
thy youth's proud livery, so gazed on now,
will be a tatter'd weed, of small worth held:
then being ask'd where all thy beauty lies,
where all the treasure of thy lusty days,
to say, within thine own deep-sunken eyes,
were an all-eating shame and thriftless praise.
how much more praise deserved thy beauty's use,
if thou couldst answer 'this fair child of mine
shall sum my count and make my old excuse,'
proving his beauty by succession thine!
this were to be new made when thou art old,
and see thy blood warm when thou feel'st it cold.

****    S03_Sonnet_3
look in thy glass, and tell the face thou viewest
now is the time that face should form another;
.....................
=====
```

Such a format does not imply a specific importation procedure.

The original text, in MsWord, for instance, has to be saved in .txt format, with an option: Insert or save the Ends of Lines and Carriage Return (to obtain lines of less than 200 characters).
Check afterwards that separators and identifiers comply with the previous constraints.

Table 4: Example of an internal DtmVic text file (type 2) for three responses to three open-ended questions and for three respondents (see: examples A5, A6, chapter III and examples B1, B2, chapter IV).

Free text format on less than 200 columns. Separator of respondents: "----" followed by the identifier (<= 20 characters); Separator of question: "++++"; End of file: "====". All separators are in columns 1, 2, 3, 4.
Note the blank lines for empty responses (last respondent, second and third questions).

The first open-ended question was *"What is the single most important thing in life for you?"* . It was followed by the probe: " *What other things are very important to you?"* . A third question has also been asked: *"What means to you the culture of your own country?"* (see Chapter III, example A5 for more information about that international sample survey).

```
---- 1006
 my sons, my kids are very important to me,
being on my own I am responsible for their education
and moral standard
++++
 education and moral standard of the youngsters, law and order
++++
 basically, British culture is traditional,
people tend to keep themselves to themselves
---- 1007
 job, being a teacher I love my job, for the well being
of the children
++++
 law and order, drug abuse, child abuse
++++
 accommodating, of course people from different races
and culture have settled in here, (i.e., Irish, Jewish,
Asians) and  the British culture is working alright
---- 1008
 job, sometimes it is very hard to find a job
++++

++++

====
```

Chapter II

Three elementary examples to discover DtmVic (numerical data)

The following three examples aim at introducing DtmVic to the user in a pragmatic fashion. Each example corresponds to a directory included in the directory ***"DtmVic_Examples_A_Start"*** *that has been downloaded with DtmVic.*

II.1 Principal Components Analysis

Example A.1. EX_A01.PrinCompAnalysis.

Active and supplementary variables. Supplementary categories. Bootstrap validation. PCA is followed by a clustering of observations, and a description of the obtained clusters.

II.2 Simple Correspondence Analysis

Example A.2. EX_A02.SimpleCorAnalysis.

Correspondence Analysis of a small contingency table. Bootstrap validation.

II.3 Multiple Correspondence Analysis

Example A.3. EX_A03.MultCorAnalysis.

Active and supplementary categories. Bootstrap validation. MCA is followed by a clustering of observations, and a description of the obtained clusters.

II.1 Principal Components Analysis

Example A.1: EX_A01.PrinCompAnalysis

Example A.1 aims at describing a set of continuous variables through PCA. The principal axes visualization is complemented with a clustering, including an automatic description of the clusters. The importance of the dichotomy *Active variables - Supplementary variables* is stressed. The data are an excerpt from a " Multimedia time budget sample survey" (carried out by the CESP in 1992. [about the CESP, see: www.cesp.org]). They deal with the average responses of a (small) subset of 96 groups of respondents to 44 questions.

The 18 000 original respondents are grouped according to some combinations of five nominal (categorical) variables: gender (2 categories), age (3 categories) activity (2 categories), educational level (3 categories) and size of town (8 categories). Our "fictitious respondents" are in fact these 96 groups. The 39 questions corresponding to numerical variables (from V6 to V44) concern the "Time spent to various activities, including sleep, meals, reading, working, etc…" (expressed in minutes per day, measured for the day preceding the interview).

The 5 questions corresponding to nominal (or categorical) variables (from V1 to V5) are: gender, age, activity, educational level, size of town.

	Caract. socio-éco				Activités							Médias		
Ident	Sexe	Age	Activ	Educ	Sommeil	Repos	Travail	Enfants	Ménage	Relation	Loisirs	…	Presse	Quotid_Nat	…
1111	H	Jeun	Actif	Prim	463,8	23,8	306,5	27,9	21,3	70,2	100,6		20,9	0,8	
1115	H	Jeun	Actif	Prim	515,6	58,5	208,8	11,3	41,9	58,3	53,1		23,7	7,2	
1121	H	Jeun	Actif	Sec	463,3	34,2	317,0	22,3	18,1	66,8	94,3		24,7	1,6	
1122	H	Jeun	Actif	Sec	456,4	43,1	250,3	19,9	26,0	82,1	105,8		31,8	3,6	
1123	H	Jeun	Actif	Sec	478,0	44,2	217,9	29,6	22,3	80,4	81,1		29,3	1,9	
1124	H	Jeun	Actif	Sec	465,1	41,6	248,5	25,9	37,0	85,8	56,3		35,3	10,2	
1135	H	Jeun	Actif	Sup	458,4	47,4	328,2	24,4	25,3	72,5	65,0		45,8	10,9	
1133	H	Jeun	Actif	Sup	457,2	30,7	274,9	20,7	52,1	86,8	79,7		36,8	5,4	
1134	H	Jeun	Actif	Sup	465,2	40,2	280,0	16,5	36,3	97,5	64,1		51,8	14,9	
2111	H	Moy	Actif	Prim	449,0	42,1	316,6	5,7	15,1	46,7	133,8		28,0	1,2	
2112	H	Moy	Actif	Prim	450,2	63,1	249,6	18,1	40,4	78,0	99,1		23,5	1,2	
2115	H	Moy	Actif	Prim	455,2	47,4	251,6	15,7	30,4	53,7	82,1		31,9	4,9	
2121	H	Moy	Actif	Sec	461,9	39,3	337,1	15,1	14,9	49,6	105,3		33,3	2,0	
2122	H	Moy	Actif	Sec	453,7	44,7	274,9	23,5	23,1	72,1	106,9		37,2	3,3	
2123	H	Moy	Actif	Sec	433,1	49,8	299,7	22,6	22,4	51,4	98,9		49,4	4,1	

***Excerpt of the Data table : Time Budget** (first rows, identifiers in French)*

1) Looking at the two files: dictionary and data.

To look at these files, use your text editor outside DtmVic (Notepad, Notepad++, Ultraedit, TotalEdit) or simply a text editor within DtmVic: button "Open an existing command file" of the main menu.
(See also Chapter I or the button "Data Format" of the First menu for comments about the internal formats of DtmVic for dictionary and data)

1.1) Dictionary file:

To have a look at the internal DtmVic format for the dictionary, search for the example directory **DtmVic_Examples_A_Start**, and in that directory, open the directory of example A.1, named **"EX_A01.PrinCompAnalysis"** .

Open then the dictionary file: **"PCA_dic_Eng.txt "** . Do not use a text processor (such as "Word"). (For a dictionary in French, open **"PCA_dic_Fr.txt "**).

The dictionary file **"PCA_dic_Eng.txt "** contains the identifiers of 44 variables. In this internal format of the dictionary, the identifiers of categories must begin at: "column 6" [a fixed interval font - also known as teletype font - such as "courier" can be used to facilitate this kind of format]. Such a dictionary can be imported from a spreadsheet format (Excel ® for instance, see Chapter 6). The identifier of a categorical variable is preceded by the number N of its categories (columns 1 to 5); the N following lines identify the N response items. An optional "short identifier" occupies columns 1 to 5. A numerical variable has 0 category.

1.2) Data file:

In a similar fashion, open the data file **"PCA_dat.txt".** The data file **"PCA_dat.txt"** comprises 96 rows and 45 columns (identifier of rows [between quotes] + 44 values [corresponding either to numerical variables or to item numbers of categorical variables] separated by at least one blank space).

Note that in this particular case, the identifier of each group happens to be a summary of the characteristics of the group: The first digit (<= 6) describe the cross-tabulation "gender – age", the second digit (<= 2) the activity, the third digit (<= 3) the educational level and the fourth and last digit the size of town (or category of agglomeration).

2) Generation of a command file (or: "parameter file")

Click "Start DtmVic".

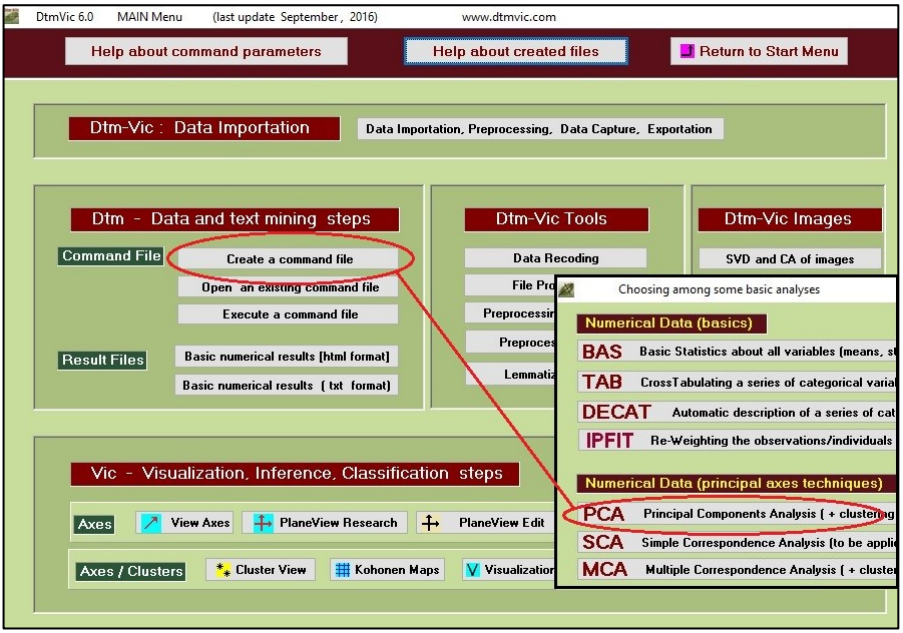
2.1) Click the button: "Create a command file" of the main menu,

The window **" Choosing among some basic analyses"** is displayed.

2.2) Click then the button:

"PCA_Principal Components Analysis"

This button is located in the paragraph: **"Numerical data"**.



2.3) Click the button: "Open a dictionary (Dtm format)"

To open the dictionary, search for the example directory **DtmVic_Examples_A_Start** , and in that directory, open the directory of example 1, named **"EX_A01.PrinCompAnalysis"** . Open then the dictionary file: **"PCA_dic_Eng.txt"** (or: **"PCA_dic_Fr.txt** " for a French version of the dictionary).

The DtmVic dictionary file in displayed in a window. Another window indicates the status of each variable (numerical or categorical).

2.4) Click the button: "Open a data file (Dtm format)"

Open the data file **"PCA_dat.txt"** .

As shown before, the data file "Dtm_PCA_dat.txt" comprises 96 rows and 45 columns (identifier of rows [between quotes] + 44 values [corresponding either to numerical variables or to item numbers of categorical variables] separated by at least one blank space.

2.5) Click the button:

"Continue (select active and supplementary elements)" .

A new window is displayed, allowing for the selection of active variables.
We suggest to select the following set of numerical variables as active variables [the reader is free to select another set of numerical variables]

Suggested set of active numerical variables

We suggest selecting the set ranging from variable V6 (duration of sleep) to variable V32 (time spent watching TV)

6 . Sleep_V6	16 . Housework_V16	26 . Errands_V26
7 . Rest_V7	17 . Contacts_V17	27 . Ambling_V27
8 . Wash_V8	18 . Call_friends_V1	28 . Errand2_V28
9 . Meal_V9	19 . Leisure_V19	29 . Moving_V29
10 . Breakfast_V10	20 . Game_V20	30 . Mov_Walk_V30
11 . Meal_home_V11	21 . Gardening_V21	31 . Mov_Car_V31
12 . Meal_rest_V12	22 . Ext_leisure_V22	32 . TV_V32
13 . Work_V13	23 . Records_V23	
14 . Work_H_V14	24 . Reading_V24	
15 . Children_V15	25 . Books_V25	

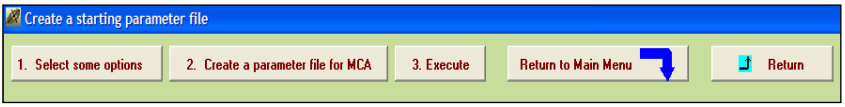
Suggested set of supplementary variables (socio-demographic characteristics):
We will characterize *a posteriori* the respondents by some socio-demographics:

1 . Gender_V1
2 . Age_V2
3 . Activity_V3
4 . Education_V4

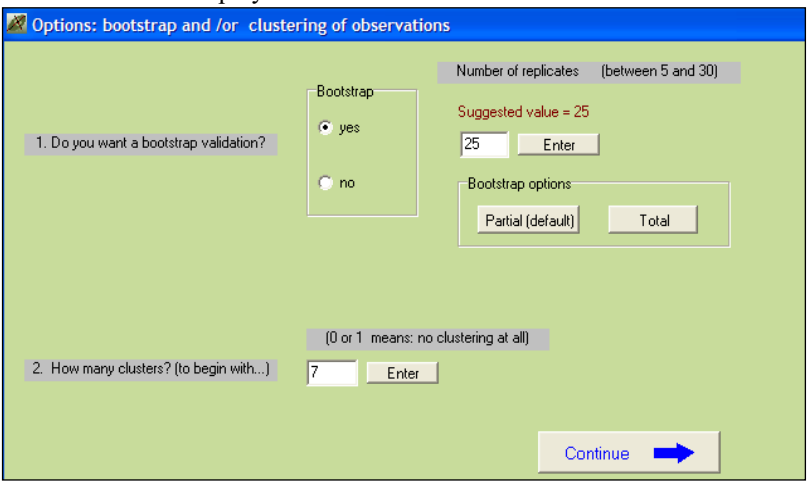
2.6) Click the button: "Continue"

A new window devoted to the selection of active observations (rows) is displayed.
Click on the button: **"All the observations will be active"** .

- The window **"Create a starting parameter file"** is displayed.



2.6.1 Click on: **"1) Select some options"** .
A new window entitled **"Options Bootstrap and/or clustering of observations"** is displayed.

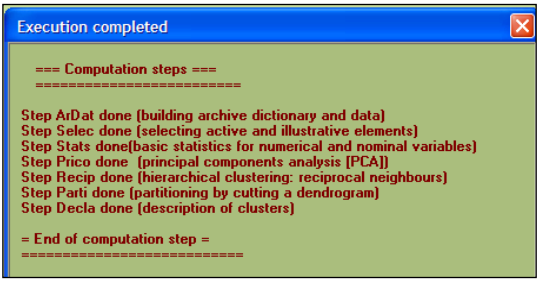


Click **"yes"** for the "Bootstrap validation", and then, click **"Enter"** for confirming the default number of replicates (25). Ignore the other suggested bootstrap options.
Select then the number of clusters (we suggest 7 clusters).
Click on: **"Enter"** and on: **"Continue"**. Back to the previous window.

2.6.2 Click on: **"2) Create a parameter file for PCA"** .
A parameter file is displayed in the memo [It can be edited by the advanced users. It allows for performing again the same analysis later on, if needed].

Important : *The parameter file is saved as* **"Param_PCA.txt"** *in the current directory*.
If you wish, you could now exit from DtmVic, and, later on, use the button of the main menu **"Open an existing command file"** (line: **"Command file"**) *to open directly the file* **"Param_PCA.txt"**, *and, in so doing, reach this point of the process, using afterwards the* "Execute" *command of the main menu.*

2.6.3 Click then on: **"3) Execute** .
This step will run the basic computation steps present in the command file: archiving data and dictionary, selection of active elements, principal components analysis of the selected data, bootstrap replications of the table, brief description of the axes, clustering procedure, thorough description of clusters. After the execution has taken place, a small window summarizes the different steps of computation:



3) Basic numerical results

Click: "Basic numerical results" button

The button opens a created (and saved) html file named **"imp.html"** which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu. Note that this file is also saved under another name. The name **"imp.html"** is concatenated with the date and time of the analysis (continental notation): **"imp_08.07.09_14.45.html"** means July 8 th, 2009, at 2:45 p.m. That file keeps as an archive the main numerical results whereas the file **"imp.html"** is replaced for each new analysis performed in the same directory.
This file in also saved under a simple text format, under the name **"imp.txt"** , and likewise with a name including the date and time of execution. **Return.**

4) Steps VIC (Visualization, Inference, Classification)

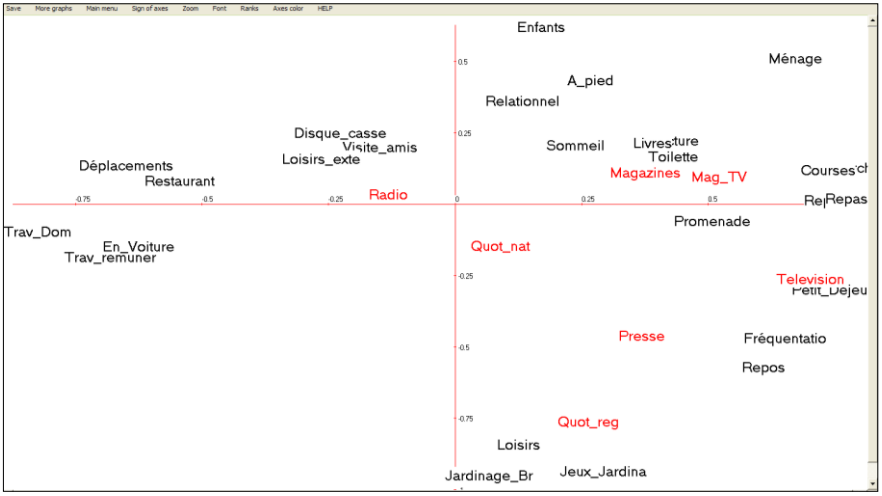
4.1) Click "ViewAxes " button

... and follow the sub-menus. In fact, only three tabs are relevant for this example: "Active variables" , "Individuals (observations)" and "supplementary categories" . After clicking on "View", the set of principal coordinates along each axis is obtained.

Clicking on a column header produces a ranking of all the rows according to the values of that column. In this particular example, this is somewhat redundant with the printed results of the step "DEFAC". Evidently, the use of the ViewAxes menu is justified when the data set is very large.
Note that for the tab: "Individuals (observations)", the procedure may help to detect possible outliers.
Return.

4.2) Click " PlaneView Research" button ... and follow the sub-menus.

In this example, six items of the menu are relevant "Active columns (variables or categories)" , "Supplementary categories", "Active rows (individuals, observations)", "Active columns + Active rows", "Active individuals (density)" and "Active columns + Supplementary categories". The graphical display of selected pairs of axes is then produced.



Plane of axes 1 and 2: Active variables (Time budget *in black*) and supplementary variables (medias: *red labels*). [the identifiers are in French]

In the "Active individuals (density)", the identifiers of individuals are replaced by a single character [case of very large set of individuals]. This display shows mainly the shape of the cloud of individuals, but the original identifiers can be produced by clicking the right button of the mouse. All the displays concern the planes spanned by the chosen pairs of axes.

In the case of PCA, the first menu item "Active columns (variables or categories)" contains in fact both active numerical variables (in black) and supplementary numerical variables (in red). The item "individuals (rows) contain our "individuals" that are, in this particular example, groups of respondents.

The roles of the different buttons are straightforward, except perhaps the button: "Rank", which is useful only in the case of very intricate displays, (which is far from being the case here!): this button converts the two coordinates of the current display into ranks. For instance, the n values of the abscissa are converted into n integers, from 1 to n, having the same order as the original values. Thus the two distributions are uniform, and the identifiers turn out to be much less overlapping, and more legible (at the cost of a substantial distortion of the display).

4.3) Click "BootstrapView" button.

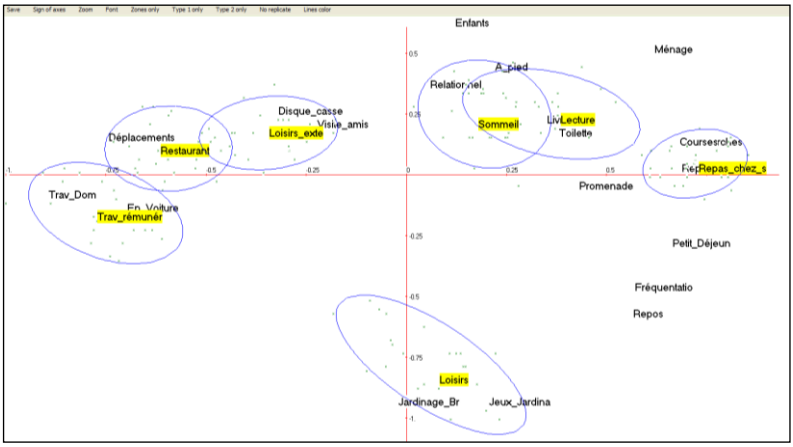
This button opens the **"DtmVic: Bootstrap - Validation - Stability – Inference"** windows. Please, click the Button "**Reminder about bootstrap methods**" for more information about the various options proposed here.

4.3.1 Click on: **"LoadData"** . In this case (partial bootstrap), the two replicated coordinates file to be opened are named **"ngus_var_boot.txt"** and **"ngus_sup_cat_boot.txt"** (see the panel reminding the names of the relevant files below the menu bar). The file **ngus_var_boot.txt** contains only active variables. The file **ngus_sup_cat_boot.txt** contains only supplementary categories, for which the bootstrap procedure is all the more meaningful.

4.3.2 Click on: " **Confidence Areas"** , submenu, and choose the pair of axes to be displayed (select axes 1 and 2 [default option] to begin with)[Enlarge the window if necessary].

4.3.3 In the window that appears then, displaying the dictionaries of variables, tick the chosen white boxes to select the elements the location of which should be assessed, and press the button **"Select"** .

4.3.4 Click on: **"Confidence Ellipses"** to obtain the graphical display of the active variable points (if the file **ngus_var_boot.txt** has been loaded), or of the supplementary category points (if the file **ngus_sup_cat_boot.txt** has been loaded).



Plane (1, 2). Confidence bootstrap ellipses for the selected (yellow) active variables.

[Note that the ellipses are large because of the small number of involved individuals (we remind that, in this example, "individuals" are in fact groups of respondents). To use bootstrap in this case leads to pessimistic confidence zones for the points. In a real application, the original individual file (comprising thousands of individuals) should be replicated before carrying out the grouping of individuals, leading then to much smaller confidences ellipses…]

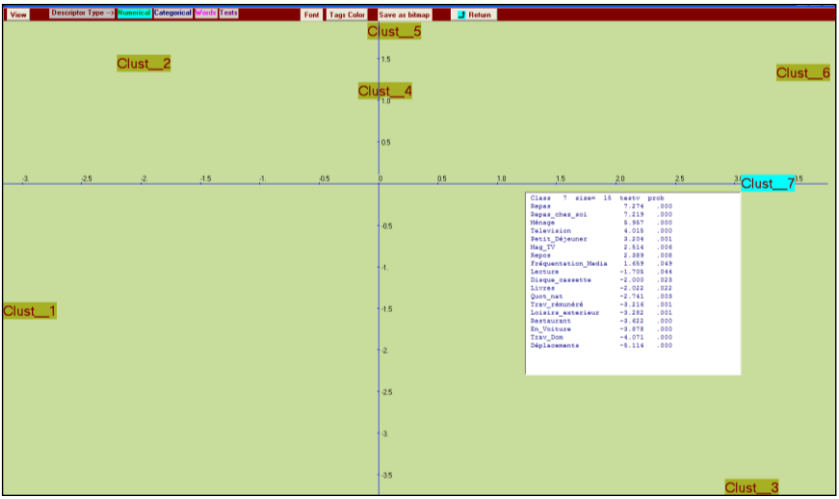
4.3.5 Close the display window (Return), and press **"Convex hulls"**. The ellipses are now replaced by the convex hulls of the replicates for each point. The convex hulls take into account the peripheral points, whereas the ellipses are drawn using the density of the clouds of replicates. The two pieces of information are complementary.

Go back to the main menu.

4.4) Click **"ClusterView"**

4.4.1 Choose the axes (1 and 2 to begin with), and **"Continue"**.

4.4.2 Click on: **"View"**. The centroids of the 7 clusters appears on the first principal plane.



Display of the 7 clusters in the plane (1, 2) with description of the cluster 7.

4.4.3 Activate the button **"Categorical"**, and, pointing with the mouse on a specific cluster, press the right button of the mouse. A description of the cluster involving the most characteristic response items appears. This description is somewhat redundant with that of the Step DECLA (see files **"imp.html" or "imp.txt"** using the buttons "Basic numerical results" of the main menu **)** . But we do have in front of us the pattern of clusters and their relative locations. One can easily imagine the usefulness of the tool for a survey with thousands of individuals, hundreds of variables, and more than 20 clusters.

4.4.4 Activate the button **"Numerical"**. We will observe the link between the numerical variables (both active and supplementary variables) of the data file and the 7 clusters. Owing to the small number of individuals, some clusters do not produce significant results.

In the context of this example, the other items of the main menu are not relevant.

General remark.

As you can observe when looking at the content of the example's directory, several files have been created and saved [these files are briefly described in the memo "Help about files" in the toolbar of the main menu]. If you need to continue using again the buttons of the paragraph VIC of the main menu after having closed DtmVic, just click on the button "Open an existing command file" from the line "Command file" , select and open the saved command file: **"Param_PCA.txt"** , and close it. It is not necessary to click on: "Execute a command file" again. You can then continue your investigation (axes views, graphs, maps, etc.).
The advanced users can also edit the parameter file **"Param_PCA.txt"** , (using the memo "Help about parameters" in the toolbar of the text editor activated by the button "Open") to perform a new analysis in which the parameters are given new values. All the intermediate files will be replaced (except the files **"imp_*date_time*.txt"** and **"imp_*date_time*.html"** which are the only saved archives)

End of example A1

II.2 Simple Correspondence Analysis

Example A.2: EX_A02.SimpleCorAnalysis

Example A.2 aims at describing a contingency table through Correspondence Analysis (CA) *aka* Simple Correspondence Analysis (SCA).

DtmVic generates numerous intermediate text-files related to a specific application. It is recommended to use one specific directory for each application.

The small data table of Example A.2 (**SCA _dat_Eng.txt**) comes from a "multi-media sample survey" (carried out by the CESP in 1992 [about the CESP, see: www.cesp.org]). It describes the distribution of six media (Radio, Television, National and regional diaries, magazines, TV magazines) among eight socio-economic categories of respondents (first eight rows). The six media are the columns, the eight categories being the rows of the contingency table. The cell (i, j) of the contingency table contains the number of contacts, during the previous day, between respondents belonging to category i and media j. Some supplementary rows give the number of contacts according to three new categorical variables: gender, age, educational level.

1) Looking at the two files: dictionary and data.

In the example directory " **DtmVic_Examples_Start"** , the sub-directory of example A.2 is named **"EXA02.SimpleCorAnalysis"** . At the outset, such directory must contain at least two files:

a) the dictionary file,

b) the data file.

To look at these files, use your text editor outside DtmVic (Notepad, Notepad++, Ultraedit, TotalEdit) or simply a text editor within DtmVic: button "Open an existing command file" of the main menu.

1.1) Dictionary file:

The dictionary name is **"SCA_dic_Eng.txt"**. (**"SCA_dic_Fr.txt"** for a French version)
This particularly simple example of dictionary file contains the identifiers of the 6 categories that are the columns of the contingency table. In this internal format of DtmVic, the identifiers of categories must begin at: "column 6" [a fixed interval font such as "courier new" should be used to facilitate the use of this kind of format].

1.2) Data file:

In a similar fashion, open the data file **"SCA_dat_Eng.txt" . ("SCA_dat_Fr.txt"** for a French version)

The data file **"SCA_dat_Eng.txt"** comprises 8 rows and 7 columns. Each row contains the identifier of rows [between quotes] + 6 values corresponding to the absolute frequencies of 6 media-categories, separated by at least one blank space.

2) Generation of a command file (or: "parameter file")

2.1) Click the button: "Create a command file" of the main menu, line "Command File".

A window "Choosing among some basic analyses" appears.

2.2) Click then the button : "SCA – Simple correspondence analysis" – located in the paragraph "Numerical data" .

2.3) Click the button "Open a dictionary (Dtm format)"

To open the dictionary, search for the examples directory **DtmVic_Examples_Start** , and in that directory, open the directory of example A.2 named **"EXA02.SimpleCorAnalysis"** . Open then the dictionary file: **"SCA_dic_Eng.txt"** (or: **"SCA_dic_Fr.txt"**. The dictionary file in displayed in a window. Another window indicates the status of each variable (all the variables have the status: "numerical" in this case).

2.4) Click the button "Open a data file (Dtm format)"

Open the data file **"SCA_dat_Eng.txt"** (or: **"SCA_dat_Fr.txt "** for the French version).

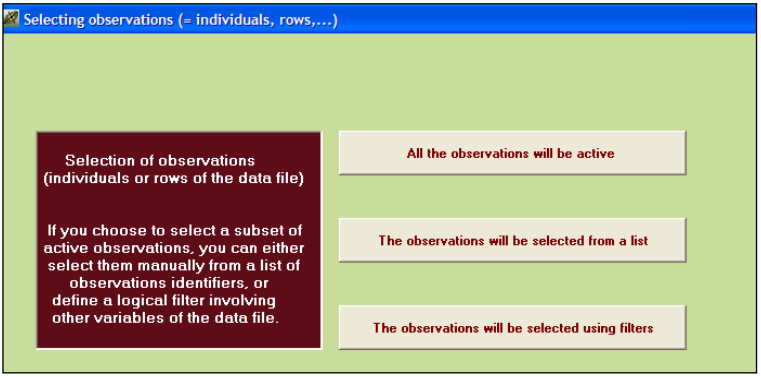
A new window displays the data file (The button "more data" is of no use in this case of small sized data set).

2.5) Click the button "Continue (select active and supplementary elements)"

A new window is displayed, allowing for the selection of active variables. In this simple case, we should select all the variables in the "memo" named "Variables to be selected" , and tick the upper blue arrow to give to the selected subset the status of "active variables" (no supplementary variables in this example).

2.6) Click the button "Continue"

A new window devoted to the selection of active observations (rows) is displayed. Click on the button: **" The observations will be selected from a list"** . Select then the first eight rows (occupations) as "active observations" (upper blue arrow) and the remaining rows as "supplementary observations". Click then on **"Continue"**.



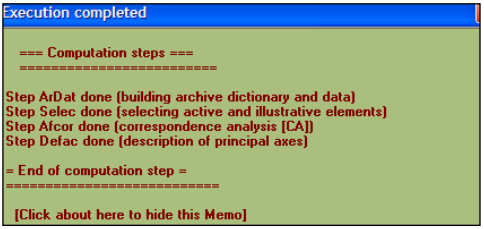
2.7) The window "Create a starting parameter file" is displayed.

2.7.1 Click on the button: **"1) Select some options"**.
A new window entitled **"Options Bootstrap and/or clustering of observations"** is displayed. Click **"yes"** for the "Bootstrap validation", and then, click **"Enter"** for confirming the default number of replicates (25). Ignore the suggested bootstrap options. Click **"Enter"** for 0 cluster, and click then on **"Continue"**.

2.7.2 Back to the previous window, click on the button: **"2) Create a parameter file for SCA"**.
A parameter file is displayed in the memo [That parameter file can be edited by the advanced users. It allows for performing again the same analysis later on, if needed].

Important: *The parameter file is saved as* **"Param_SCA.txt"** *in the current directory. If you wish, you could now exit from DtmVic, and, later on, use the button of the main menu* **"Open an existing command file"** **(line: "Command file")** *to open directly* **"Param_SCA.txt"** *, and, in so doing, reach this point of the process, using the "Execute" command of the main menu..*

2.7.3 Click then on the button: **“3) Execute a command file”**.
The basic computation steps mentioned in the command file are: archiving data and dictionary, selection of active elements, correspondence analysis of the selected table, bootstrap replications of the table to build confidence regions for column-points and row-points, brief description of the axes. After the execution has taken place, a small window summarizes the different steps of computation.



3) Basic numerical results

Click “Basic numerical results” button.

The button allows the user to browse a created (and saved) html file named **“imp.html”** which contains the main results of the previous basic computation steps. After perusing these numerical results, **“Return”** to the main menu.

Note that this file is also saved under another name. The name **“imp.html”** is concatenated with the date and time of the analysis (continental notation): **“imp_08.07.13_14.45.html”** means July 8 th, 2013, at 2:45 p.m. That file keeps as an archive the main numerical results whereas the file **“imp.html”** is replaced for each new analysis performed in the same directory.

This file in also saved under a simple text format , under the name **“imp.txt”** , and likewise with a name including the date and time of execution.

Return.

4) Steps VIC (Visualization, Inference, Classification)

4.1) Click the "ViewAxes" button …

and follow the sub-menus. In fact, only two tabs are relevant for this first simple example: "Active variables" and "Individuals (observations)". After clicking on "View" in both cases, one obtains the set of principal coordinates along each axis.
Clicking on a column header produce a ranking of all the rows according to the values of that column. In this particular example, this is somewhat redundant with the printed results of the step "DEFAC" printed in the log-file **"imp.txt"**.
Evidently, the use of the ViewAxes menu makes sense when the data set is very large.
Return.

4.2) Click the "PlaneView Research" button …

and follow the sub-menus.
In this example, only three items are relevant "Active columns (variables or categories)", "Active rows (individuals, observations)", "Active columns + Active rows" (respectively columns, rows of the data table, and simultaneous representation of rows and columns). The graphical displays of chosen pairs of axes are then produced.
The roles of the different buttons are straightforward, except perhaps the button: "Rank", which is useful only in the case of very intricate displays, (which is far from being the case here!): this button converts the two coordinates of the current display into ranks. For instance, the n values of the abscissa are converted into n integers, from 1 to n, having the same order as the original values. Thus the two distributions are uniform, and the identifiers turn out to be much less overlapping, and more legible (often at the cost of a substantial distortion of the display).
Return.

4.3) Click the "BootstrapView" button...

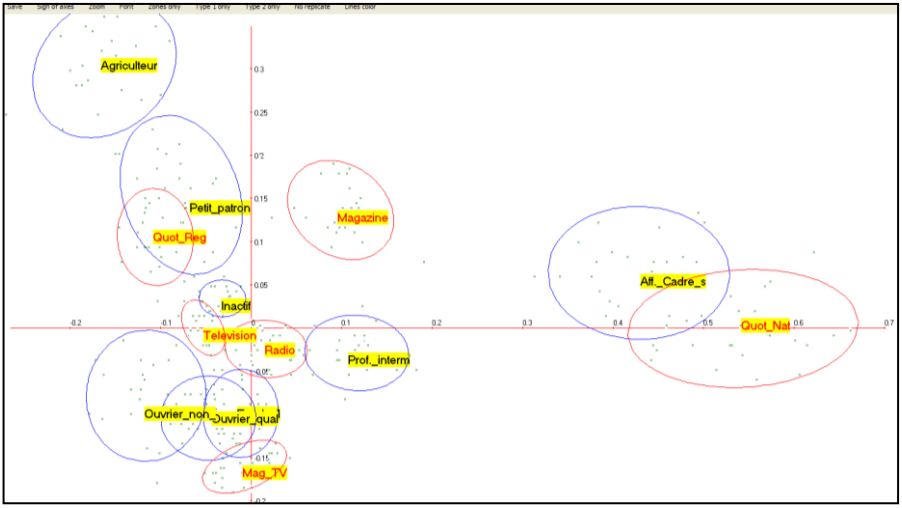
This button opens the **DtmVic-Bootstrap-Stability** windows.

4.3.1 Click on **"LoadData"** . In this case (partial bootstrap), the replicated coordinates file to be opened is named **"ngus_var_boot.txt"** .
4.3.2 Click on: **"Confidence Areas"**, submenu, and choose the pair of axes to be displayed (select axes 1 and 2 [default option] to begin with).

4.3.3 In the window that appears then, displaying the dictionaries of variables, tick the chosen white boxes to select the elements the location of which should be assessed, and press the button **"Select"** .

4.3.4 Click on **"Confidence Ellipses"** to obtain the graphical display of the column points (or variable points) in red colour, and of the row points (or individuals or observations) in blue.

In this display, we learn for example that all the occupation groups (row points) have distinct "media-contact-profiles", except the categories "Skilled worker" and "Unskilled worker" on the one hand, and "Skilled worker" and "Employees" on the other, whose confidence areas are largely overlapping.



Plane (1, 2): Bootstrap confidence ellipses for columns (medias, in red) and rows (occupations, in blue). [labels in their French version].

4.3.5 Close the display window, and, again in the blue window, press **"Convex hulls"**. The ellipses are now replaced with the convex hulls of the replicates for each point. The convex hulls take into account the peripheral points, whereas the ellipses are drawn using the density of the clouds of replicates. The two pieces of information are complementary…

In the context of this example, the other items of the main menu are not relevant.

General remark.
As you can observe when looking at the content of the example's directory, several files have been created and saved [these files are briefly described in the memo "**Help about files"** in the toolbar of the main menu]. If you need to continue using again the buttons of the paragraph VIC of the main menu after having closed DtmVic, just click on the button **"Open an existing command file"** from the line "command file" , select and open the saved command file: **"Param_SCA.txt"** , and close it. It is not necessary to click on: **"Execute a command file"** again. You can then continue your investigation (axes views, graphs, maps, etc.).

The advanced users can also edit the parameter file **"Param_SCA.txt"** , (using the memo **"Help about parameters"** in the toolbar of the editor) to perform a new analysis in which the parameters are given new values.. All the intermediate files will be replaced (except the file **"imp_date_time**.txt" which is the only saved archive).

End of example A2

II.3 Multiple Correspondence Analysis

Example A.3: EX_A03.MultCorAnalysis

Example A.3 aims at describing a set of categorical variables through MCA. The corresponding data sets are located in the subdirectory named: **EX_A03.MultCorAnalysis** within the directory **DtmVic_Examples_A_Start** . The data are an excerpt from a "sample survey about living conditions and aspirations of the French" [carried out by the CREDOC (www.credoc.fr) in 1986]. They deal with the responses of a (small) subset of 315 individuals to 49 questions. Some questions concern objective characteristics of the respondent or his/her household (age, status, gender, facilities). Other questions relate to attitude or opinions.

The principal axes visualization will be complemented with a clustering, including an automatic description of the clusters. The importance of the dichotomy: *Active variables* - *Supplementary variables* is stressed.

1) Looking at both files: dictionary and data.

1.1) Dictionary file:

To have a look at the dictionary, search for the examples directory **DtmVic_Examples_A_Start** , and, in that directory, open the directory of example A.3 named **"EX_A03.MultCorAnalysis"** .

Open then the dictionary file: **"MCA_Eng_dic.txt** " (for a dictionary in French, open: **"MCA_Fr_dic.txt"**).

The dictionary file **MCA_Eng_dic.txt** contains the identifiers of the 51 variables. About the internal data and dictionary format of DtmVic, please refer to the first examples above, or to Chapter 1.

1.2) Data file:

That data file comprises 315 rows and 50 columns (identifier of rows [between quotes] + 49 values [corresponding either to numerical variables or to item numbers of categorical variables] separated by at least one blank space).

2) Generation of a command file (or: "parameter file")

Open DtmVic.

2.1) Click the button: "Create a command file" of the main menu, section: " Command File".

A window "Choosing among some basic analyses" appears.

2.2) Click then the button: "MCA– Multiple correspondence analysis"

– located in the paragraph: **"Numerical data (principal axes techniques)"** .

2.3) Click the button: "Open a dictionary (Dtm format)"

To open the dictionary, search again for the examples directory " **DtmVic_Examples_A_Start" ,** and, in that directory, open the directory of example A3, named **"EX_A03.MultCorAnalysis"** . Open then the dictionary file: **"MCA_Eng_dic.txt** " (for a dictionary in French, open: **"MCA_Fr_dic.txt"**). The dictionary file in displayed in a window. Another window indicates the status of each variable (numerical or categorical).

2.4) Click the button: "Open a data file (Dtm format)"

Open the data file **"MCA_dat.txt".**
A new window displays the data file.

2.5) Click the button: "Continue (select active and supplementary elements)".

A new window is displayed, allowing for the selection of active variables.
We suggest to select the following set of categorical variables as active variables [of course, the reader is free to select another set of categorical variables]

Suggested set of active categorical variables (sample of opinions)

8 . family_is_the_only_place.. 9 . opinion_about_marriage 10 . house_work 11 . satisfaction_dwelling 12 . satisfaction_envir.	21 . headache 22 . backache 23 . nervousness 24 . depression 25 . health_satisfaction	34 . society_needs_changes? 48 . About_justice 49 . People_like_me_feel_alone

Suggested set of supplementary variables (socio-demographic characteristics)

3 . gender 50 . Age_categ 51 . Educ_3_categ

The active categorical variables are in this case 13 questions (opinions and attitudes) about family, housing expenditure, society, health problems, and anxiety.

Click the button: "Continue"

A new window devoted to the selection of active observations (rows) is displayed. Click on the button: **"All the observations will be active"**.

2.7 The window "Create a starting parameter file" is displayed.

2.7.1 Click on: **"1) Select some options"**.

A new window entitled **"Options Bootstrap and/or clustering of observations"** is displayed. Click **"yes"** for the "Bootstrap validation", and then, click **"Enter"** for confirming the default number of replicates (25). Ignore the other suggested bootstrap options.

Select then the number of clusters (we suggest 5) then click on: **"Enter"** and on: **"Continue"**.

Back to the previous window,

2.7.2 Click on: **"2) Create a parameter file for MCA"**.

A parameter file is displayed in the memo [such a parameter can be edited by advanced users. It allows for performing again the same analysis later on, if needed].

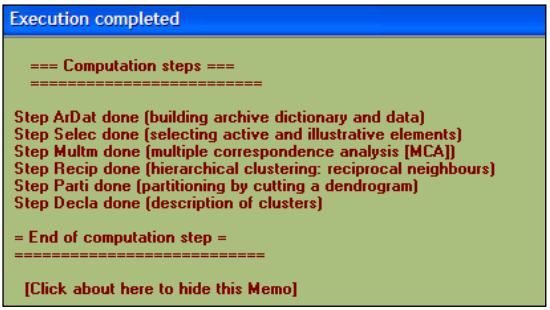
Important: *The parameter file is saved as* **"Param_MCA.txt"** *in the current directory.*

If you wish, you could now exit from DtmVic, and, later on, use the button of the main menu **"Open an existing command file"** (line: **"Command file"**) *to open directly* **"Param_MCA.txt"** *, and, in so doing, reach this point of the process.*

2.7.3 Click then on: **"3) Execute a command file"**.

This step will run the basic computation steps present in the command file: archiving data and dictionary, selection of active elements, multiple correspondence analysis of the selected table, bootstrap replications of the table, brief description of the axes, clustering procedure with a thorough descriptions of clusters.

After the execution has taken place, a small window summarizes the different steps of computation:



3) Basic numerical results

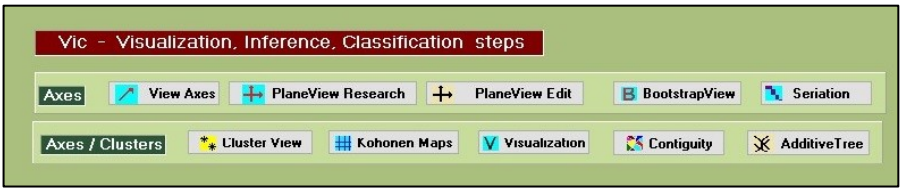
Click "Basic numerical results" button

The button opens a created (and saved) *html* file named **"imp.html"** which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu. [Note that this file is also saved under another name. The name **"imp.html"** is concatenated with the date and time of the analysis (continental notation): **"imp_08.07.13_14.45.html"** means July 8 th, 2013, at 2:45 p.m. That file keeps as an archive the main numerical results whereas the file **"imp.html"** is replaced for each new analysis performed in the same directory].

This file in also saved under a simple text format, under the name **"imp.txt"** , and likewise with a name including the date and time of execution.

Return.

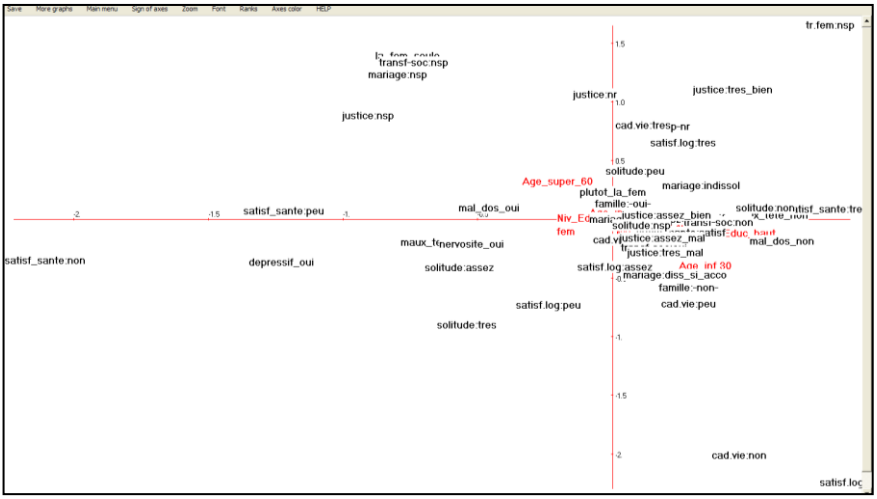
4) Steps VIC (Visualization, Inference, Classification)



4.1) Click "ViewAxes" button … and follow the sub-menus. In fact, only three tabs are relevant for this example: **"Active variables"** , **"Supplementary categories"** and **"Individuals (observations)"** . After clicking on **"View"** for each case, one obtains the set of principal coordinates along each axis.

Clicking on a column header produces a ranking of all the rows according to the values of that column. In this particular example, this is somewhat redundant with the printed results of the step "DEFAC" (see files "imp.txt" or "imp.html" through the buttons "Basic numerical results"). The use of the ViewAxes menu is profitable when the data set is very large.
Return.

4.2) Click "PlaneView Research" button … and follow the sub-menus. In this example, six items of the menu are relevant **"Active columns (variables or categories)", "Supplementary categories", "Rows (individuals, observations)" , "Active columns + Rows" , "Rows, Individuals (density)"** and **"Active columns + Supplementary categories"**.
The graphical displays of the selected pairs of axes are then produced.



Plane (1, 2): Active variables (in black) and supplementary categories (in red)
[Labels in their French version]

In the **"Rows, Individuals (density)"**, the identifiers of individuals are replaced by a single character [case of very large set of individuals... useful to detect and identify outliers]. This display shows mainly the shape of the cloud of individuals, but the original identifiers can be produced by clicking the right button of the mouse. All the displays concern the planes spanned by the chosen pairs of axes.

The roles of the different buttons are straightforward, except perhaps the button: **"Rank"** , which is useful only in the case of very intricate displays, (which is far from being the case here!): this button converts the two coordinates of the current display into ranks.

The n values of the abscissa are converted into n integers, from 1 to n, having the same order as the original values. Thus the two distributions are uniform, and the identifiers turn out to be much less overlapping, and more legible (at the cost of a substantial distortion of the display). This example is in fact a counterexample of that property: MCA derived from a few active categorical variables produces a lot of superimposed points, that are perfectly superimposed in the display of "individuals" and slightly different in the display of ranks (according to the option chosen here, they occupy consecutive or neighbouring ranks).

Return.

4.3) Click "BootstrapView" button.

This button opens the **DtmVic-Bootstrap-Stability** windows.

4.3.1 Click on: **"LoadData".** In this case (partial bootstrap), the two replicated coordinates file to be opened are named **"ngus_var_boot.txt"** and **"ngus_sup_cat_boot.txt"** (look at the small panel reminding the names of the relevant files below the menu bar).
In fact, in this version, the file **ngus_var_boot.txt** contains both active and supplementary categories. The file **ngus_sup_cat_boot.txt** contains only supplementary categories, for which the bootstrap procedure is more meaningful.

4.3.2 Click on: **"Confidence Areas"**, submenu, and choose the pair of axes to be displayed (select axes 1 and 2 [default option] to begin with).

4.3.3 In the window that appears then, displaying the dictionaries of variables, tick the chosen white boxes to select the elements the location of which should be assessed, and press the button **"Select"**.

4.3.4 Click on: **"Confidence Ellipses"** to obtain the graphical display of the active category points (in blue colour), and of the supplementary category points (in red).
In this display, we learn for example that in this principal space (built as a "space of opinions", due to the selection of active questions), male and

female [two supplementary categories that did not participate in building the axes] occupy distinct locations (ellipses no overlapping at all).
To test such a hypothesis (independence between the pattern of opinions and the gender) it is convenient (i.e. more legible) to tick only the two categories "male" and "female", and select the options of the upper bar: "Zones only" and "No replicate".
In the same vein, we can tick the classes of ages, and observe that the extreme categories ("under 30" and "over 60" correspond to confidence ellipses clearly separated).

4.3.5 Close the display window, and, press **"Convex hulls"**. The ellipses are now replaced by the convex hulls of the replicates for each point. The convex hulls take into account the peripheral points, whereas the ellipses are drawn using the density of the clouds of replicates. The two pieces of information are complementary.

Go back to the main submenu "VIC".

4.4. Click "ClusterView"

4.4.1 Choose the axes (1 and 2 to begin with), and **"Continue"**.

4.4.2 Click on the button **"View"** . The centroids of the 5 clusters appear on the first principal plane (Steps RECIP and PARTI of the created command file **"Param_MCA.txt"** , i.e.: Clustering using reciprocal neighbours (RECIP), then cut of the dendrogram and optimization of the cut through *k-means* (PARTI)).

4.4.3 Activate the button **"Categorical"**, and, pointing with the mouse on a specific cluster, press the right button of the mouse. A description of the cluster involving the most characteristic response items appears. This description is somewhat redundant with that of the Step DECLA (see files **"imp.txt"** or **"imp.html"** using the button **"Basic numerical results"**). But we do have in front of us the pattern of clusters and their relative locations. One can easily imagine the usefulness of the tool for a survey with 3000 individuals, hundreds of variables, and, say, 20 clusters.

In the context of this example, the other items of the main menu are not relevant.

General remark.
As you can observe when looking at the content of the example's directory, several files have been created and saved [these files are briefly described in the memo

“Help about files” in the toolbar of the main menu]. If you need to continue using again the buttons of the paragraph VIC of the main menu after having closed DtmVic, just click on the button **“Open an existing command file”** from the line “Command file” , select and open the saved command file: **“Param_MCA.txt”** , and close it. It is not necessary to click on: **“Execute a command file”** again. You can then continue your investigation (axes views, graphs, maps, etc.).
The advanced users can also edit the parameter file **“Param_MCA.txt”** , (using the memo **“Help about parameters”** in the toolbar of the main menu) to perform a new analysis in which the parameters are given new values. All the intermediate files will be replaced (except the file **“imp_*date_time*.txt”** which is the only saved archive).

End of example A3

Chapter III

Three more elementary examples to discover DtmVic (textual data)

The following three examples aim at introducing DtmVic to the user in a pragmatic fashion. Each example corresponds to a directory included in the directory ***"DtmVic_Examples_A_Start"*** *that has been downloaded with DtmVic.*

III.1 Correspondence Analysis of a lexical table

Example A.4. EX_A04.Text-Poems.

Processing of a simple series of texts (20 first Shakespearian Sonnets). Numerical coding. Correspondence Analysis of the lexical table words - poems. Bootstrap validation. Characteristic words and verses. Kohonen maps. Seriation.

III.2 Open questions in a sample survey: Agglomerations of responses

Example A.5. EX_A05.Text-Responses_1.

Using both numerical and textual data. Processing of the responses to an open-ended question using a specific categorical variable to agglomerate the responses. Numerical coding of the responses. Correspondence Analysis of the lexical table words x categories. Bootstrap validation. Description of the categories through their characteristic words and responses. Kohonen map for words and categories.

III.3 Open questions in a sample survey: Direct analysis, link with closed-end questions

Example A.6. EX_A06.Text-Responses_2

First processing of individual responses to an open-ended question. Numerical coding of the responses. Correspondence Analysis (CA) of the sparse lexical table words x respondents, clustering of the responses, description of the obtained clusters through their characteristic words, responses, and also through their characteristic categories (closed questions).

III.1 Correspondence Analysis of a lexical table

Example A.4: EX_A04.Text-poems

This elementary example deals with the simplest form of text analysis: The data set comprises a series of texts separated by the separator **** (columns 1 to 4). The dataset serving as an example, “Sonnet_LowerCase.txt”, contains the first 20 Sonnets from Shakespeare. For a larger set of sonnets and for comments, see, among many other websites, www.shakespeare-online.com/sonnets/ .
In this simple format, DtmVic can process up to 1200 texts without limitation of size for each text. Our corpus serving as an example is thus a “small scale model”, emphasizing only the functionalities (but not the power) of DtmVic. The conversion to lower case characters is meant to avoid typifying the first word of each verse or sentence.

The general methodology underlying the processing is presented in the book: “Exploring Textual data” (L. Lebart, A. Salem, L. Berry; Kluwer Academic Publisher, 1998). That textbook is an upgraded translation of the book: ”Statistique Textuelle” (Ludovic Lebart and André Salem, Dunod, Paris, 1994). This latter book (in French) can be freely downloaded from the site: www.dtmvic.com (section “publication”).

1) Looking at the text file (see table 3 of chapter 1)

Search for the examples directory: “ **DtmVic_Examples_A_Start”**
In that directory, open the folder of example A.4 named : **“EX_A04.Text-poems“** .
As mentioned in the previous examples, it is recommended to use one directory for each application, since DtmVic produces a lot of intermediate ”txt” files related to the application. At the outset, such directory must contain at least one text file: **“Sonnet_LowerCase.txt”** . Look at this text file using a text editor such as Notepad, Notepad++, TotalEdit, Ultraedit, TextEdit, or simply a text editor within DtmVic: button **"Open an existing command file"** of the main menu.

Excerpts of the data (see also: table 3 of chapter 1)

```
****    S_1
from fairest creatures we desire increase,
that thereby beauty's rose might never die,
but as the riper should by time decease,
his tender heir might bear his memory:
but thou, contracted to thine own bright eyes,
.....................................
****    S_2
when forty winters shall beseige thy brow,
and dig deep trenches in thy beauty's field,
thy youth's proud livery, so gazed on now,
.....................................
```

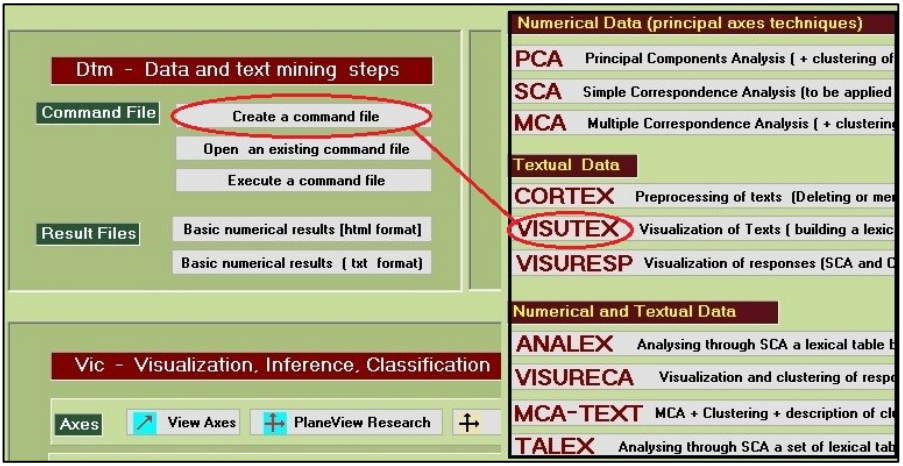
```
but since she prick'd thee out for women's pleasure,
mine be thy love and thy love's use their treasure.
.....

====
```

– The format is specific (DtmVic internal text format type 1. See Chapter 1, or the button "Data Format" of the first menu).
– Since the texts may have very different lengths, separators **** (at the beginning of a line) are used to distinguish between texts.
– The length of the lines is limited to 200 characters.
– The identifiers of texts must follow the separator "****" after 4 blank spaces.
– The symbol "====" indicates the end of the file.
– Like all the data files involved in DtmVic as input files, that file is a raw text file (.txt). If the text file comes from a text processing phase, it must be saved beforehand as a ".txt file".

2) Generation of a command file (or: "parameter file")

2.1) Click the button: "Create a command file" of the main menu, line: "Command File"

A window **"Choosing among some basic analyses"** appears.

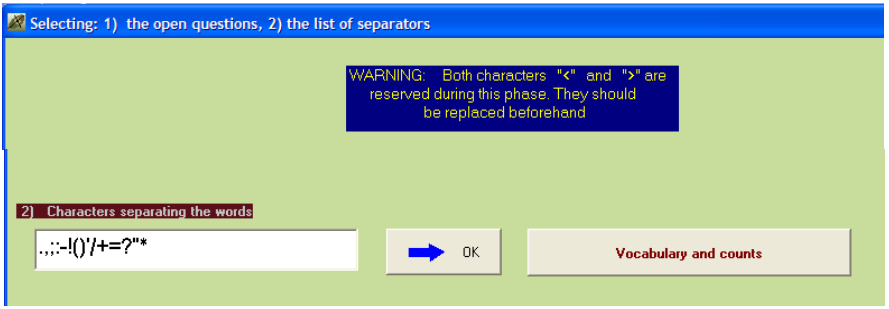


2.2) Click then the button: "VISUTEX – Visualization of Texts"

This button is located in the paragraph **"Textual and numerical data"**.

2.3) Press the button: "Open a text file", then search for the directory: **"DtmVic_Examples_A_Start"**. In that directory, open the directory of example A.4, named: **"EX_A04.Text-poems"** .
Open the file: **"Sonnet_LowerCase.txt"** .
A message box indicates then that the corpus contains 20 texts totalizing 321 lines.

2.4) Click: "Select open questions and separators"
The next window allows for the selection of open questions (not relevant here) and the selection of separators of words (the produced default separators suffice in this example) (lower part of the window below).

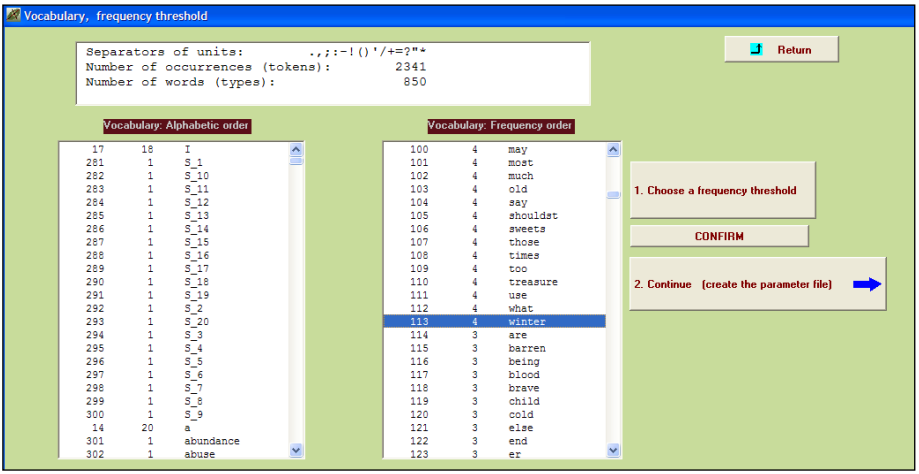


2.5) Click directly "Vocabulary and counts".
The next window presents the vocabulary (alphabetic and frequency orders).

We must select a threshold of frequency by selecting a line in the right hand side memo (frequency order).

The line number 113 corresponds to the frequency 4 (It is a very small frequency, adapted to a very small corpus.

This example is just an opportunity of exploring the sequence of commands, without meaningful linguistic interpretation).



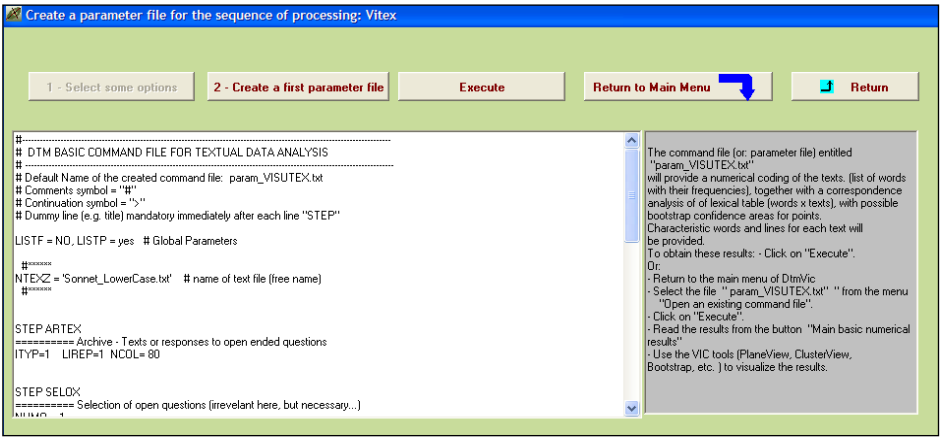
After selecting that line, click then on: **"Confirm".**

2.6) Click on: "Continue. (Create the parameter file)"

Continuing our visit, we have to **"Select some options"** . Click **"yes"** for the bootstrap validation, and **"Enter"** to confirm the default number of replicates (25). Click then on **"Continue"**.

2.7) Click on: "Create a first parameter file"

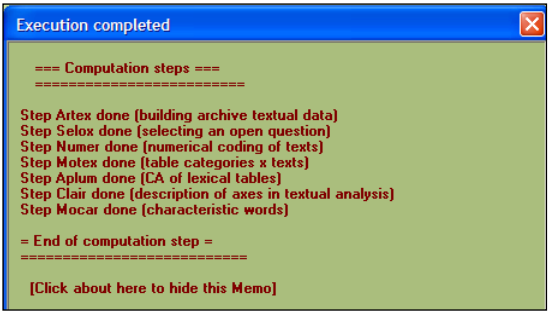
A parameter file is displayed in the memo [It can be edited by the advanced users. It allows for performing again the same analysis later on, if needed].
Important : *The parameter file is saved as* **"Param_VISUTEX.txt"** *in the current directory.*



If you wish, you could now exit from DtmVic, and, later on, use the button of the main menu **"Open an existing command file"** (line: **"Command file"**) *to open directly* **"Param_VISUTEX.txt"** *, and, in so doing, you reach this point of the process. You can then use afterwards the* **"Execute a command file"** *command of the main menu.*

2.8) Click on: "Execute a command file".

This step will run the basic computation steps present in the command file: archiving data and text, characteristic words and responses, correspondence analysis of the lexical table.

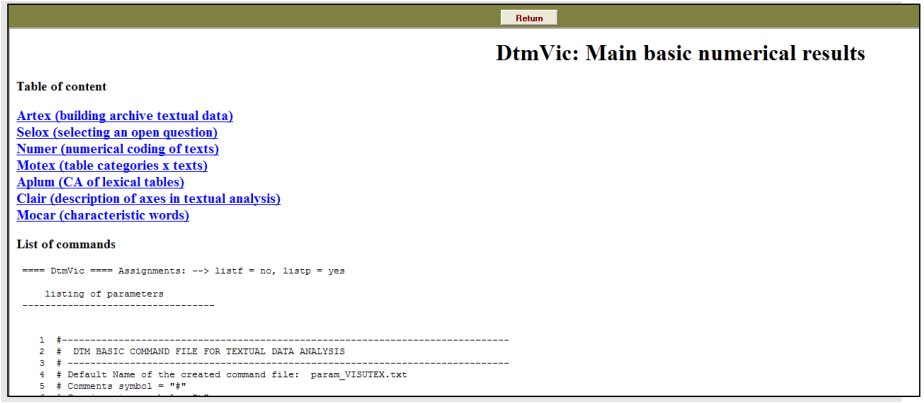


3) Basic numerical results

Click on: "Basic numerical results" button

The button opens a created (and saved) *html* file named **"imp.html"** which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu. Note that this file is also saved under another name.

[The name **"imp.html"** is concatenated with the date and time of the analysis (continental notation): **"imp_08.07.13_14.45.html"** means July 8 th, 2013, at 2:45 p.m. That file keeps as an archive the main numerical results whereas the file **"imp.html"** is replaced for each new analysis performed in the same directory. This file in also saved under a simple text format , under the name **"imp.txt"** , and likewise with a name including the date and time of execution].

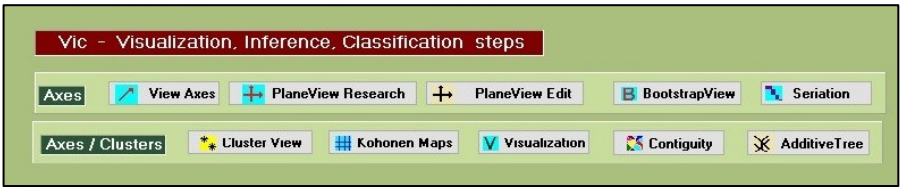


From the step NUMER, we learn for instance that we have 280 responses (lines), with a total number of words (occurrences or token) of 2321, involving 830 distinct words (or: types).

Using a frequency threshold of 3 (it means here keeping the words with frequency over three) the total number of kept words reduces to 1384, whereas the number of distinct kept words reduces to 114. (Note some – provisional– notational differences: the minimal selected frequency 4 corresponds to the frequency 3 in the listing meaning, equivalently, that all the words appearing more than three times are kept).

Return.

4) Steps VIC (Visualization, Inference, Classification)

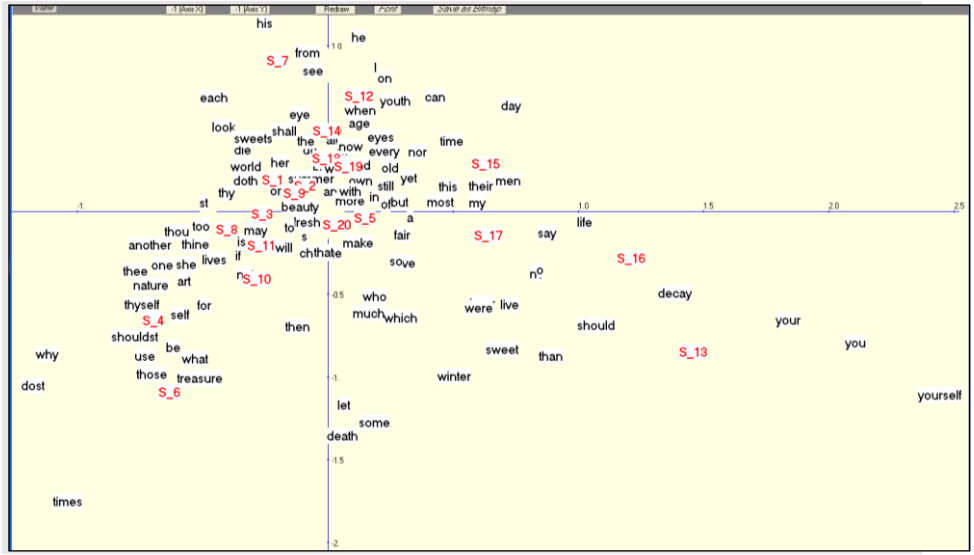


4.1) Click the button: "ViewAxes"

... and follow the sub-menus. In fact, only two tabs are relevant for this example: **"Active variables"** [= poems] and **"observations"** [words]. After clicking on **"View"**, the user obtains the set of principal coordinates along each axis.
Clicking on a column header produce a ranking of all the rows according to the values of that column.
As mentioned in the previous examples, the use of the ViewAxes menu is justified when the data set is large, which is not the case here.
Return.

4.2) Click the button: "PlaneView Edit" , and follow the sub-menus...

In this example, only one item of the menu is relevant **"Active columns + Rows"**. This item concerns both rows and columns of the contingency table (lexical table). The graphical displays of selected pairs of axes are then produced. Normally, the active categories (columns of the lexical table) are printed in red, while the active words (rows) are printed in blue.
The roles of the different buttons are straightforward, except perhaps the button: "Rank", which is useful only in the case of very intricate displays (which is not the case here) (see comments in the previous examples).
Return.



Example of "PlaneView Edit" (with moveable tags) (Sonnets + words)

4.3) Click the button: "BootstrapView"

This button opens the **"DtmVic: Bootstrap - Validation - Stability – Inference"** windows.

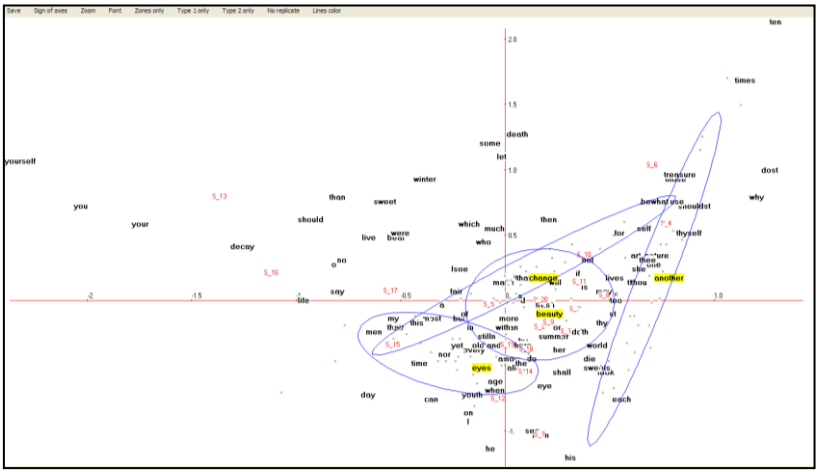
4.3.1 Click on: **"LoadData"**. In this case (partial bootstrap), the replicated coordinates file to be opened is named **"ngus_par_boot1.txt"** . (The set of possible files is given by a background panel).

4.3.2 Click on: **"Confidence Areas"** submenu, and choose the pair of axes to be displayed (select axes 1 and 2 to begin with).

4.3.3 The window that appears (enlarge it if necessary) contains the list of identifiers of active rows and columns (identifiers of columns [Sonnets in this case] are at the end of the list). Tick some white boxes to select some poems, select also some words, and press the button **"Select"**.

4.3.4 Click on: **"Confidence Ellipses"** to obtain the graphical display of the chosen column points in red colour, and of the row points (here: words) in

blue. We can see that several sonnets occupy significant locations (several confidence ellipses do not overlap) whereas the locations of the words are far from being as accurate.



4.3.5 Close the display window, and, again in the blue window, press **"Convex hulls"**. The ellipses are now replaced by the convex hulls of the replicates for each point. The convex hulls take into account the peripheral points, whereas the ellipses are drawn using the density of the clouds of replicates. The two pieces of information are complementary.

4.4) Click "ClusterView" (in this case, the clusters are the texts themselves)

4.4.1 Choose the axes (1 and 2 to begin with), and **"Continue"**.

4.4.2 Click on **"View"**. The locations of the 20 categories (texts) appear on the first principal plane. Thanks to some possible change of signs for the axes, the display is the same as that provided by the **"PlaneView"** procedures.

4.4.3 Activate the button **"Words"**, and, pointing with the mouse on a specific category, press the right button of the mouse. A description of the

category involving the most characteristic words of the category appears. This description is again redundant with that of the Step MOCAR (see files **"imp.txt"** or **"imp.html"** using the button **"Basic numerical results"**). But we can appreciate here the pattern of categories and their relative locations.

4.4.4 Activate the button **"Texts"**. Pointing with the mouse on a specific category, and pressing the right button of the mouse, we can read the most characteristic lines (verses) of the selected category. The concept of characteristic line is not obviously relevant in the case of poetries. It is in fact a particular case of the concept of "characteristic responses", extremely useful in the case of open questions.

More explanation about the corresponding methodology can be found in the already quoted book: " *Exploring Textual data* " (L. Lebart, A. Salem, L. Berry; Kluwer Academic Publisher, 1998).

Return.

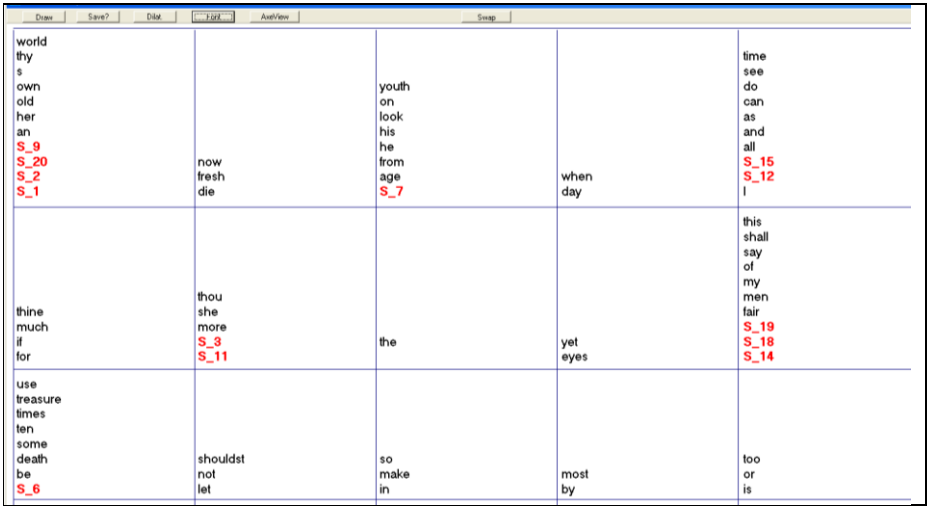
4.5) Click "Kohonen map"

4.5.1 Select: **"variables + observations (rows + columns)"**: these active variables are the words **and** the texts (poems) in this example.

4.5.2 Select a (5 x 5) map, and **"Continue"**.

4.5.3 Press **"Draw"** on the menu of the large green windows entitled "Kohonen map".

4.5.4 You can change the font size (**"Font"**) and dilate the obtained Kohonen map (**"Dilat."**) to make it more legible. The words appearing in the same cell are often associated in the same responses. This property holds, at a lesser degree, for contiguous cells.



Self Organizing Map plotting both words and texts (Sonnets).

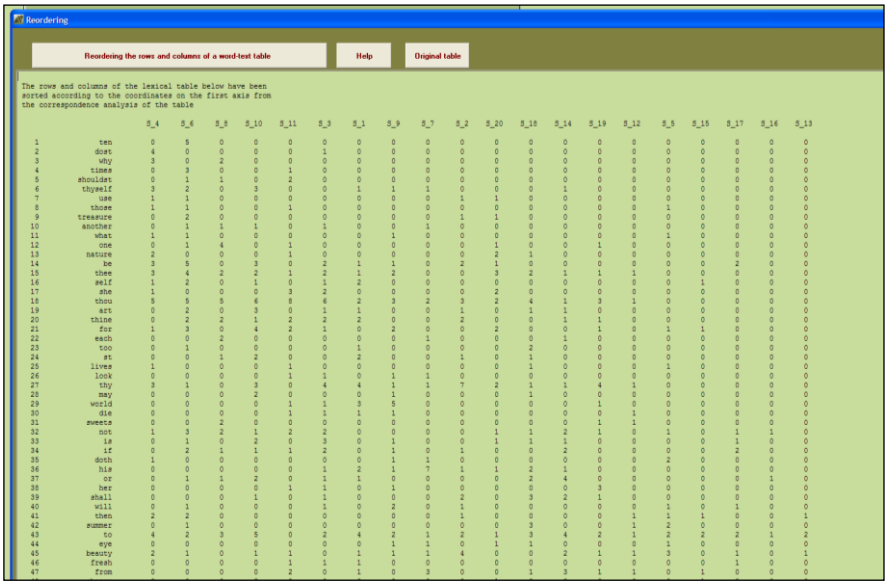
4.5.5 Note that we have obtained a simultaneous Kohonen representation of rows and columns, owing to the use, as an input file, of the coordinates from the correspondence analysis of the lexical table (coordinates allowing for a simultaneous representation).

4.6) Click "Seriation"

Seriation techniques as well as Block Seriation techniques are widely used by practitioners. Seriation is based on simple row and column permutations of the table under study; they have the great practical and cognitive advantage of showing the raw data to the user and therefore allowing the user to forego the use of intricate interpretation rules. These permutations can display homogenous blocks of high values or on the contrary, of small or null values. They can also pinpoint a continuous and progressive evolution of profiles.

An optimal property of correspondence analysis is the following: the first axis of a correspondence analysis provides us with a ranking of the row-points and of the column-points. That ranking can be used to sort the rows and columns of the analysed data table. The new obtained data table has then undergone an optimal seriation. Seriation will be applied here to the lexical table cross-tabulating the 20 sonnets and the selected words (words appearing at least 4 times in the corpus).

A new window named **"Reordering"** appears.



Click on the button: "Reordering the rows and the column of a word-text table". The reordered table cross-tabulating the 20 sonnets and the selected words is then displayed. It can be seen that the first words of the reordered list of words characterize (sometimes exclusively) the first sonnets in the reordered list of sonnets. The last words of the same list are either absent or rarely observed among these sonnets. However, they are frequent among the last sonnets (right hand side of the table). That reordered printing of the raw data is a useful tool of communication with the practitioners, since it can be interpreted without prior knowledge of data analysis techniques.

Remark. The columns are identified by the first four characters of the text identifiers. It is then helpful to use identifiers whose first four characters are all distinct.

General remark.
As you can observe when looking at the content of the example's directory, several files have been created and saved [these files are briefly described in the memo **"Help about files"** in the toolbar of the main menu]. If you need to continue using again the buttons of the paragraph VIC of the main menu after having closed DtmVic, just click on the button **"Open an existing command file"** from the line **"Command file"** , select and open the saved command file:

“Param_VISUTEX.txt” , and close it. It is not necessary to click on: **“Execute a command file”** again. You can then continue your investigation (axes views, graphs, maps, etc.).

The advanced users can also edit the parameter file **“Param_VISUTEX.txt”** , (using the memo **“Help about parameters”** in the toolbar of the main menu) to perform a new analysis in which the parameters are given new values. It is advised to give it a new name (such as **“Param_VISUTEX2.txt”** , for example). All the intermediate files will be replaced (except the files **“imp_date_time.txt”** and **“imp_date_time.html”** which are the only saved archives).

End of example A4

III.2 Open questions in sample surveys: agglomeration of responses

Example A.5: EX_A05.Text-Responses_1

Example A.5 aims at describing the responses to an open-ended question in a sample survey in relation with the responses to a specific closed-end question.
After archiving dictionary, data and texts, the numerical coding of the text allows us to build a lexical table cross-tabulating the words with a selected categorical variable. A correspondence analysis is then performed on that lexical table. Bootstrap confidence areas (ellipses or convex hulls) can be drawn around words and categories. Characteristics words/responses are computed for each category.

About the data

The open questions were included in a multinational survey conducted in seven countries (Japan, France, Germany, United Kingdom, USA, Netherlands, Italy) in the late nineteen eighties (Hayashi *et al* ., 1992). It is the United Kingdom survey which is presented here. It deals with the responses of 1043 individuals to 14 closed-end questions and three open-ended questions. Some questions concern objective characteristics of the respondent or his/her household (age, status, gender, facilities). Other questions relate to attitude or opinions. The first open-ended question was *"What is the single most important thing in life for you?"* It was followed by the probe: " *What other things are very important to you?"* . A third question (not analysed in this manual, but included in the example data set) has also been asked: *"What means to you the culture of your own country?"*

We will focus on the first open question and its probe. Being interested with the relationships between these responses and both the age and educational level of the respondent, we will use a specific categorical variable to agglomerate the open responses: a variable with nine categories cross-tabulating three categories of age with three educational levels. More explanations about this particular example and the corresponding methodology can be found in the book: " *Exploring Textual data*" (L. Lebart, A. Salem, L. Berry; Kluwer Academic Publisher, 1998).
This example corresponds to the directory **"EX_A05.Text-Responses_1"** included in: **"DtmVic_Examples_A_Start"**.

1) Looking at the three files: data, dictionary and texts.

1.1) Data file: " TDA_dat.txt" *(Excerpt below)*

The data file comprises 1043 rows and 15 columns (identifier of rows [between quotes] + 14 values [corresponding either to numerical variables or to item numbers of categorical variables] separated by at least one blank space).

First (5) and last (5) rows of the data table.

```
'___1'   1  12  80   1   2   3   3   3   2   1   3   3   1   3
'___2'   1   8  54   1   1   1   3   1   1   1   2   2   1   2
'___3'   1   6  40   1   1   2   1   2   2   2   2   2   1   2
'___4'   2   3  27   2   1   2   1   1   1   1   1   4   5   4
'___5'   2   5  39   2   2   1   3   1   1   1   2   5   5   5

 . . . . . .
'1039'   1   8  54   2   2   4   2   0   0   1   2   2   2   5
'1040'   2   3  27   2   5   4   2   1   1   1   1   4   5   4
'1041'   1   2  23   3   3   2   1   2   2   1   1   1   3   7
'1042'   1   9  57   2   4   3   1   1   2   2   3   3   2   6
'1043'   2   5  38   1   5   3   5   2   2   2   2   5   4   2
```

1.2) Dictionary file: "TDA_dic.txt" *(Excerpt below)*

The dictionary file " **TDA _dic.txt"** contains the identifiers of the 14 variables. In this internal dictionary of DtmVic, the identifiers of categories must begin at: "column 6" The identifier of a categorical variable is preceded by the number N of its categories (columns 1 to 5); the N following lines identify the N response items. An optional "short identifier" could be located in columns 1 to 5. A numerical variable (such as "age") has 0 category. Note that blank spaces are not allowed within the identifiers (about DtmVic formats, see: chapter 1).

```
   2 GENDER
MALE MALE
FEMA FEMALE
  12  AGE_CODE
AGE1  18_19
AGE2  20_24
AGE3  25_29
AGE4  30_34
AGE5  35_39
AGE6  40_44
AGE7  45_49
AGE8  50_54
AGE9  55_59
AG10  60_65
AG11  65_70
AG12  71_et_+
   0 AGE
   3 EDUCATION
EDUL LOW
EDUM MEDIUM
EDUH HIGH
   3 WILL_PEOLE_BE_HAPPIER?
HAP1 Happier
HAP2 LESS_happy
HAP3 About_the_same
   4 PEOLE_PEACE_OF_MIND...
PEA1 INCREASES
PEA2 DECREASES            PEA3
NOT_CHANGES
PEA4 OTHER
   3 MORE_OR_LESS_FREEDOM
FRE1 MORE_FREEDOM
FRE2 LESS_FREEDOM
FRE3 THE_SAME
   3 Age_3_ classes
 -30 less_than_30
3055 from_30_to_55
+ 55 over_55
```

1.3) Text file: "TDA_tex.txt" *(Excerpt below)*

This file contains the free responses of 1043 individuals to three open-ended questions mentioned earlier.

The DtmVic internal format of the text file is very specific. Since the responses may have very different lengths, separators are used to distinguish between questions and between individuals (or: respondents). Individuals are separated by the chain of characters "----" (starting column 1) possibly followed by an identifier. Within each individual data, the open questions are separated by "++++" (column 1). The symbol "====" indicates the end of the file. Like all the data files involved in DtmVic as input files, that file is a raw text file (.txt). If the text file comes from a text processing phase, it must be saved as a ".txt file".

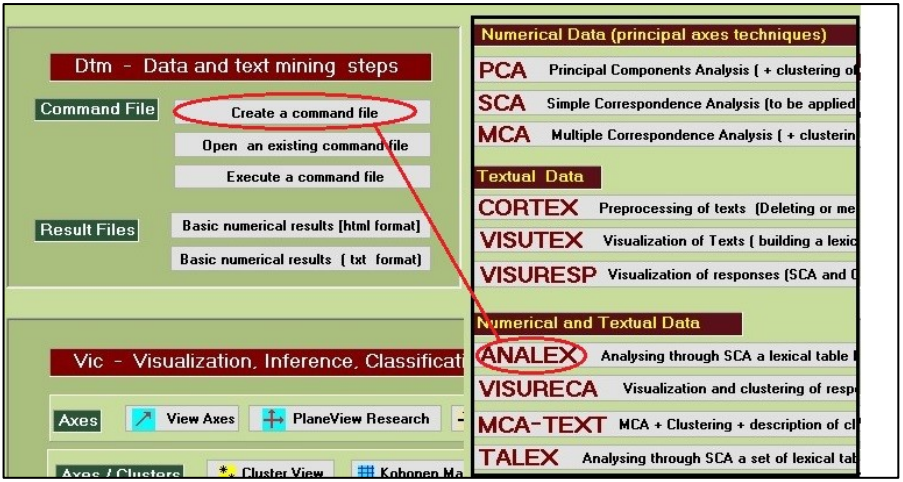
```
----'___1'
 good health
++++
 happiness
++++

----'___2'
 happiness in people around me, contented family, would make me happy
++++
 contented with life as a whole
++++
education
----'___3'
 contentment
++++
 family
++++
 arts
……

----1043
 contentment
++++
 my children's health and happiness
++++

====
```

2) Generation of a command file (or: "parameter file")

2.1) Click the button: "Create a command file" of the main menu, line "Command File".

A window "Choosing among some basic analyses" appears.

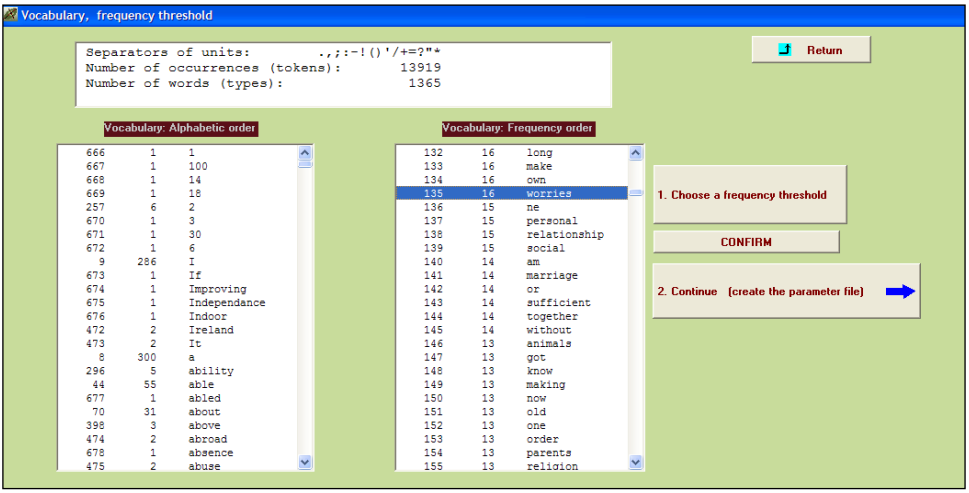
2.2) Click then on the button: "ANALEX" – located in the paragraph **"Textual and numerical data"**.



2.3) Press the button: "Open a text file", then search for the directory: **"DtmVic_Examples_A_Start"**. In that directory, open the directory of example A.5, named **"EX_A05.Text-Responses"** .
Open then the text file: **"TDA_tex.txt".** A message box indicates then that the corpus comprises 7329 lines, 1043 observations and 3 open questions.

2.4) Click on: "Select Open questions and separators"
The next window allows for the selection of open questions and the selection of separators of words (the default list of separators suffices in this example).
We will select questions 1 and 2 (that means that the two responses will be merged). It is licit here to merge the two responses because question 2 is a probe for question 1.

2.5) Click directly on: "Vocabulary and counts".
The next window presents the vocabulary (alphabetic and frequency orders). We must select a threshold of frequency by selecting a line in the right hand side memo frequency order). The line number 135 corresponds to the frequency 16. After selecting that line, click on: **"Confirm".**



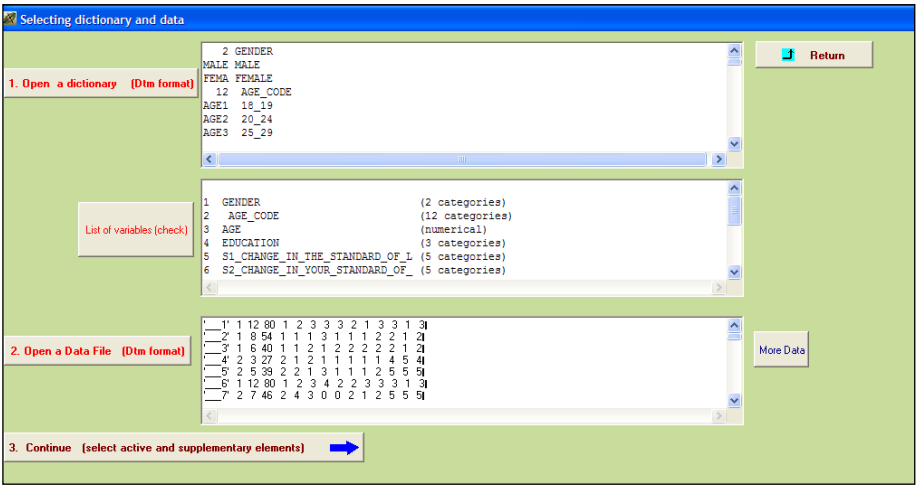
Then click on: **"Continue"**.

2.6) Click the button: "Open a dictionary (Dtm format)"

Open then the dictionary file: **"TDA_dic.txt "** .

The dictionary file **TDA_dic.txt** contains the identifiers of the 14 variables.

The dictionary file in displayed in a window. Another window indicates the status of each variable (numerical or categorical).



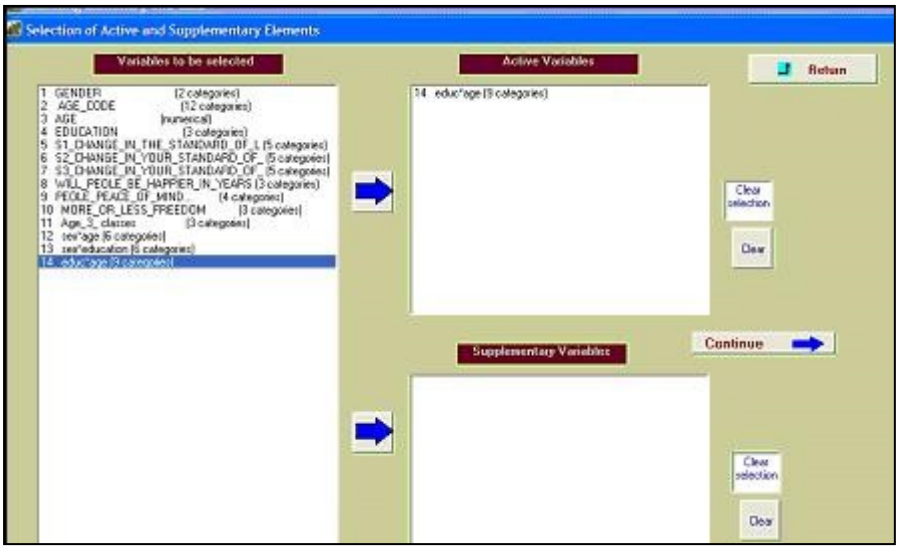
2.7) Press the button: "Open a data file (Dtm format)"

Open the data file: **"TDA_dat.txt"** .
That data file comprises 1043 rows and 15 columns (identifier of rows [between quotes] + 14 values [corresponding either to numerical variables or to item numbers of categorical variables] separated by at least one blank space).
A new window displays the data file.

2.8) Click the button:

"Continue (select active and supplementary elements)"

A new window is displayed, allowing for the selection of active variables.
We suggest to select the categorical variable number 14, (age - education). Only one active variable can be selected in the ANALEX case.
All the remaining variables could be selected as supplementary elements. They will serve to describe the categories of the active variable.



2.9) Click then on the button: "Continue"

A new window devoted to the selection of active observations (rows) is displayed.
Click on the button: **"All the observations will be active"** .
The window **"Create a starting parameter file"** is displayed.

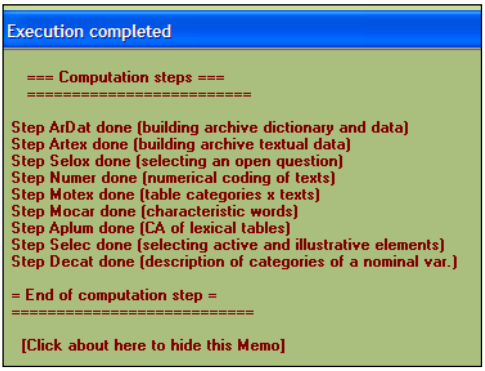
2.10) Click on: "1-Select some options". A new window entitled **"Options Bootstrap and/or clustering of observations"** is displayed. Click **"yes"** for the "Bootstrap validation", and then, click **"Enter"** for confirming the default number of replicates (25). Ignore the other suggested bootstrap options.
Back to the previous window.

2.11) Then click: "2-Create a first parameter file"

A parameter file is displayed in the memo [It can be edited by the advanced users. It allows for performing again the same analysis later on, if needed].
Important : *The parameter file is saved as* **"Param_ANALEX.txt"** *in the current directory. If you wish, you could now exit from DtmVic, and, later on, use the button of the main menu* **"Open an existing command file"** (line: **"Command file"**) *to open directly the file* **"Param_ANALEX.txt"** , *and, in so doing, reach directly this point of the process, using the* **"Execute a command file"** *command of the main menu.*

2.12) Click: "3-Execute" .

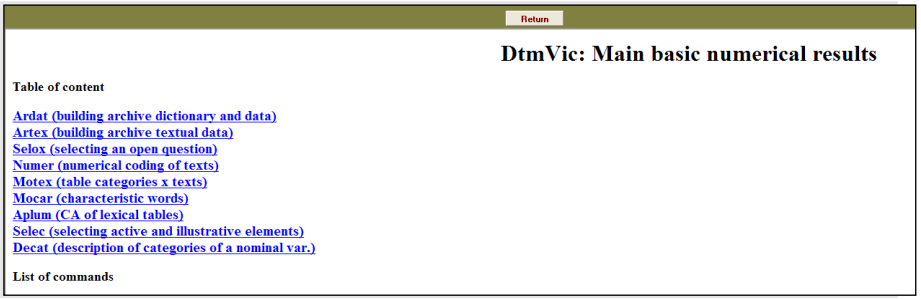
This step will run the basic computation steps present in the command file: archiving data and text, characteristic words and responses, correspondence analysis of the lexical table, thorough descriptions of categories using other variables.



3) Basic numerical results

Click on "Basic numerical results" button

The button opens a created (and saved) html file named **"imp.html"** which contains the main results of the previous basic computation steps.



After perusing these numerical results, return to the main menu. Note that this file is also saved under another name.

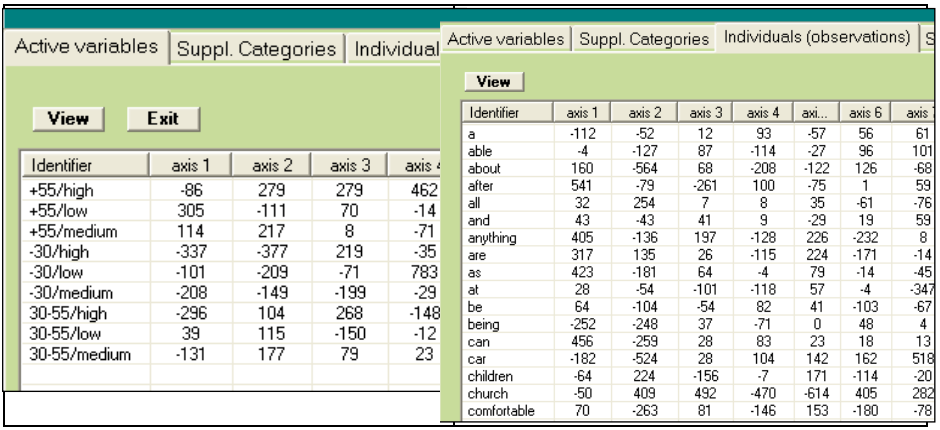
[The name **"imp.html"** is concatenated with the date and time of the analysis (continental notation): **"imp_08.07.09_14.45.html"** means July 8th, 2009, at 2:45 p.m. That file keeps as an archive the main numerical results whereas the file: **"imp.html"** is replaced for each new analysis performed in the same directory].

This file in also saved under a simple text format , under the name **"imp.txt"** , and likewise with a name including the date and time of execution.
From the step NUMER, we learn for instance that we have 1043 responses, with a total number of words (occurrences or token) of 13 919, involving 1 365 distinct words (or: types). Using a frequency threshold of 16, [the same threshold id denoted 15 in the result file: first neglected frequency] the total number of kept words reduces to 10738, whereas the number of distinct kept words reduces (more drastically) to 136.
The book "Exploring textual data" (*op. cit.*) deals in details with this pre-processing and with all the results that follow.

4) Steps VIC (Visualization, Inference, Classification)

4.1) Click the button: "ViewAxes" ... and follow the sub-menus.

In fact, two tabs are relevant for this example: **"Active variables"** [= categories, in the case of ANALEX], and **"Individuals"** [words]. After clicking on: **"View"**, one obtains the set of principal coordinates along each axis. Clicking on a column header produces a ranking of all the rows according to the values of that column. Evidently, the use of the ViewAxes menu is justified when the data set is large, which is the case here. **Return.**



Active variables | Suppl. Categories | Individual

View | Exit

Identifier	axis 1	axis 2	axis 3	axis
+55/high	-86	279	279	462
+55/low	305	-111	70	-14
+55/medium	114	217	8	-71
-30/high	-337	-377	219	-35
-30/low	-101	-209	-71	783
-30/medium	-208	-149	-199	-29
30-55/high	-296	104	268	-148
30-55/low	39	115	-150	-12
30-55/medium	-131	177	79	23

Active variables | Suppl. Categories | Individuals (observations) | S

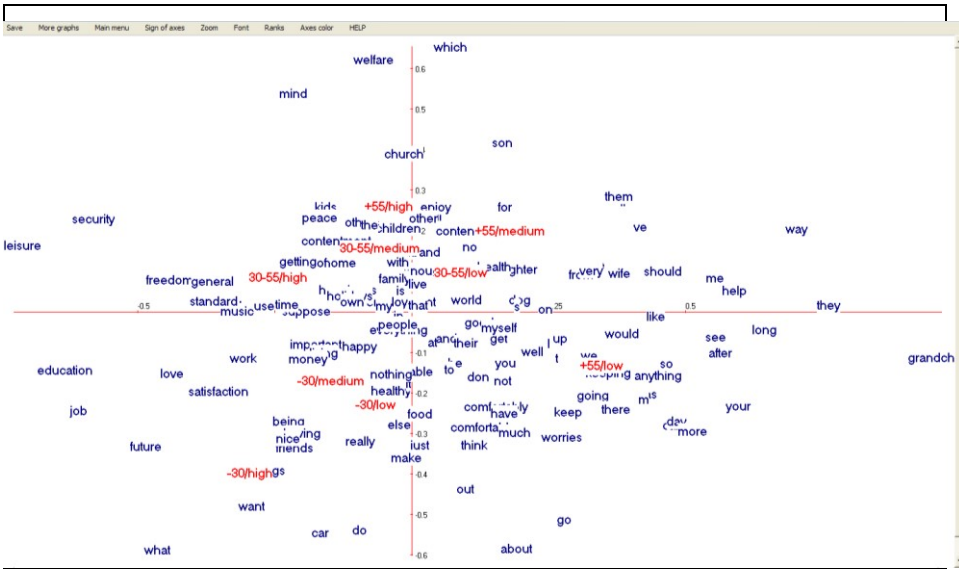
View

Identifier	axis 1	axis 2	axis 3	axis 4	axi...	axis 6	axis
a	-112	-52	12	93	-57	56	61
able	-4	-127	87	-114	-27	96	101
about	160	-564	68	-208	-122	126	-68
after	541	-79	-261	100	-75	1	59
all	32	254	7	8	35	-61	-76
and	43	-43	41	9	-29	19	59
anything	405	-136	197	-128	226	-232	8
are	317	135	26	-115	224	-171	-14
as	423	-181	64	-4	79	-14	-45
at	28	-54	-101	-118	57	-4	-347
be	64	-104	-54	82	41	-103	-67
being	-252	-248	37	-71	0	48	4
can	456	-259	28	83	23	18	13
car	-182	-524	28	104	142	162	518
children	-64	224	-156	-7	171	-114	-20
church	-50	409	492	-470	-614	405	282
comfortable	70	-263	81	-146	153	-180	-78

Active variables (categories) and Individuals (words)

4.2) Press the button: "PlaneView Resarch"

In this example, three items of the menu are relevant: **"Active columns (variables or categories)"**, **"Rows (Individuals)"**, and **"Active columns + Rows"**. This last item concerns both rows and columns of the contingency table (lexical table).



Plane (1, 2): Categories of respondents (red) and words (blue)

The graphical displays of the selected pairs of axes are then produced. The active categories (columns of the lexical table) are printed in red, while the active words (rows) are printed in blue.

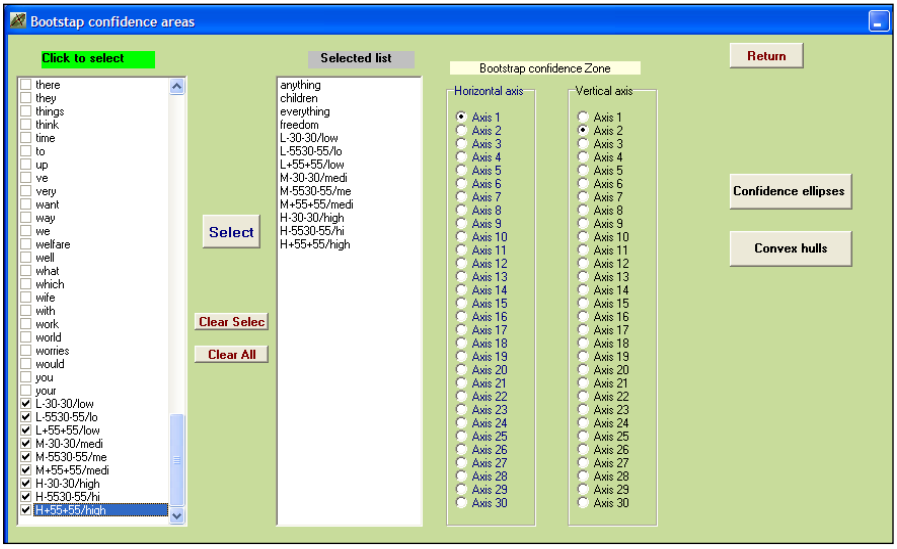
The roles of the different buttons are straightforward, except perhaps the button: "Rank", which is useful only in the case of very intricate displays (which is not the case here). (See comments in the texts relating to examples A.1 and A.2).

4.3) Click on the button: "BootstrapView"

This button opens the **"DtmVic: Bootstrap - Validation - Stability – Inference"** windows.

4.3.1 Click on: **"LoadData"** . In this case (partial bootstrap), the replicated coordinates file to be opened is named: **"ngus_par_boot1.txt"** . (The set of possible files is given by the panel).

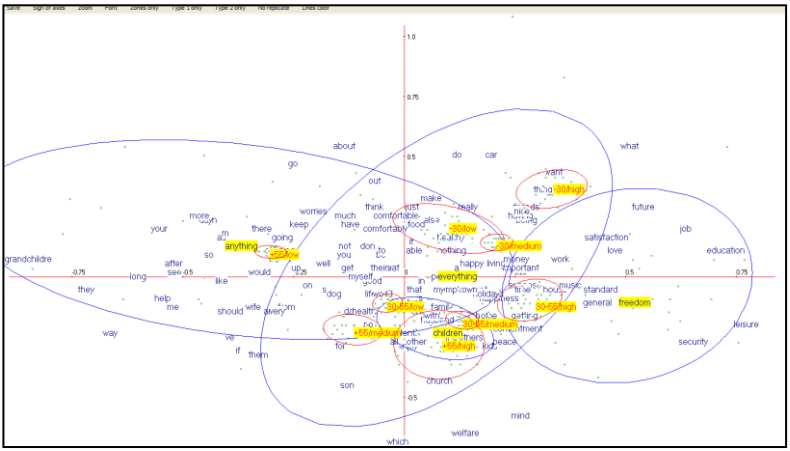
4.3.2 Click on: **"Confidence Areas"** submenu, and choose the pair of axes to be displayed (select axes 1 and 2 to begin with).



Selection of ellipses: Categories are at the end of the words file.

4.3.3 We obtain the list of the identifiers of active rows and columns (identifiers of columns [categories age x education] are at the end of the list). Since the column set is quite small, tick all the white cases to select all the categories, select also some words, and press the button: **"Select"** .

– Click on: **"Confidence Ellipses"** to obtain the graphical display of the chosen column points in red colour, and of the row points (or individuals or observations) in blue.



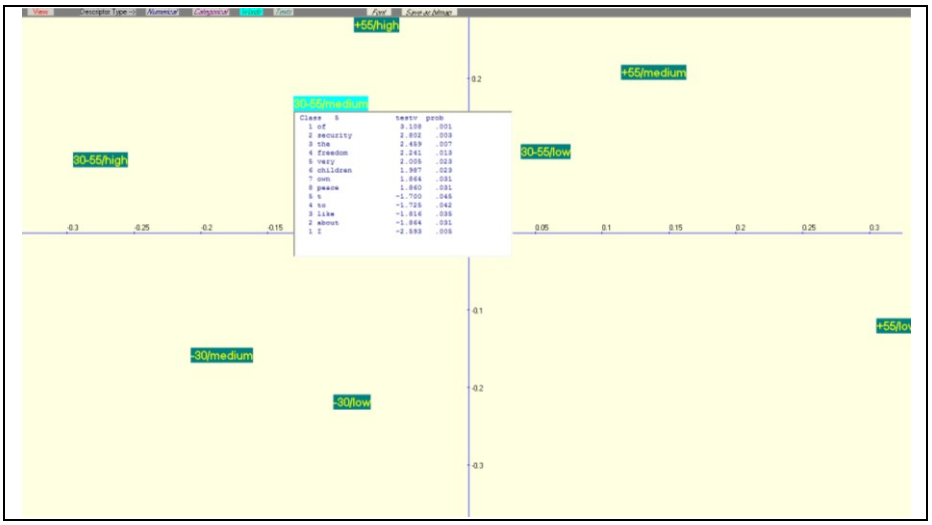
We can see that, individually, some words have no significant position. In this display, we learn for example that almost all the age-education groups (column points - red) have distinct "lexical profiles", except the categories "-30-low" [less than 30 years old, low level of education] and "-30-medium" [less than 30 years old, medium level of education] whose confidence areas are largely overlapping.
We can see, for instance, that some flections of the verb "to be", such as "is, be, are, being" may have locations that differ significantly.

– Close the display window, and, again in the blue window, press: **"Convex hulls"**. The ellipses are now replaced with the convex hulls of the replicates for each point. The convex hulls take into account the peripheral points, whereas the ellipses are drawn using the density of the clouds of replicates. The two pieces of information are complementary.
Return.

4.4) Click on: **"ClusterView "** [in this case, clusters are categories]

4.4.1 Choose the axes (1 and 2 to begin with), and **"Continue"**.

4.4.2 Click on **"View"**. The locations of the 9 categories (variable 14: age-education) appears on the first principal plane. Thanks to some possible change of sign for the axes, the display is the same as that provided by the **"PlaneView"** procedure.



4.4.3 Activate the button: **"Words"**, and, pointing with the mouse on a specific category, press the right button of the mouse. A description of the category involving the most characteristic words of the category appears. This description is again redundant with that of the Step MOCAR (file **"imp.txt"**). But we can observe here the pattern of categories and their relative locations.

4.4.4 Activate the button: **"Texts"**. Pointing with the mouse on a specific category, and pressing the right button of the mouse, we can read the most characteristic responses of the selected category.
[More explanation about the corresponding methodology can be found in the book: "*Exploring Textual data*" (L. Lebart, A. Salem, L. Berry; Kluwer Academic Publisher, 1998)].
Return.

4.5) Click **"Kohonen map"**

4.5.1 Select: **"variables + observations (rows + columns)"** : these active variables are the words **and** the texts (categories) in this example.

4.5.2 Select a (5 x 5) map, and **"Continue"**.

4.5.3 Press **"Draw"** on the menu of the large green windows entitled **"Kohonen map"** .

4.5.4 You can change the font size (**"Font"**) and dilate the obtained Kohonen map (**"Dilat."**) to make it more legible. The words appearing in the same cell are often associated with the same responses. This property holds, at a lesser degree, for contiguous cells.

4.5.5 Note that we have obtained a simultaneous Kohonen representation of rows and columns, owing to the use, as input file, of the coordinates from the correspondence analysis of the lexical table.

Self Organizing Map plotting both words and texts

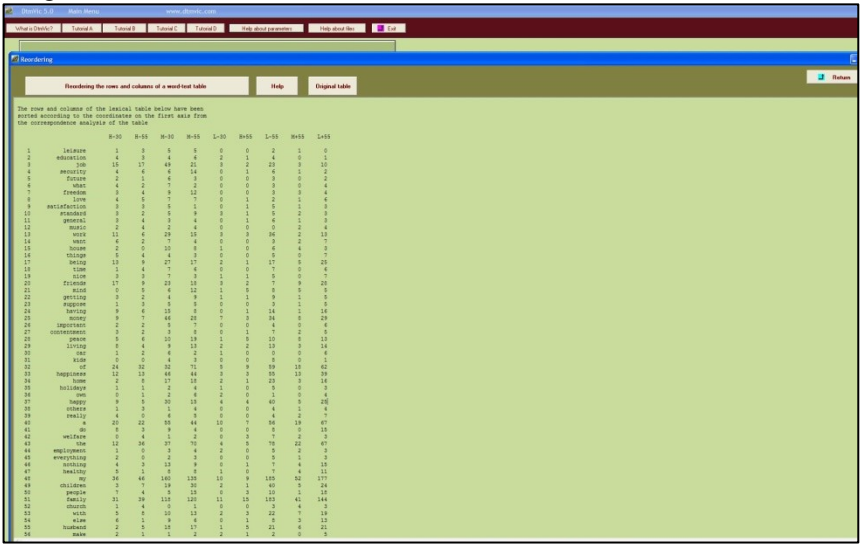
4.6) Click: "Seriation"

The aim of seriation techniques has been briefly described in the section 4.6 of example 4. Seriation will be applied here to the lexical table cross-tabulating the 9 categories of respondents and the selected words (words appearing at least 16 times in the corpus). In this version of DtmVic, Seriation can be obtained only after the three types of analysis: SCA, VISUTEX and ANALEX. All these approaches involve Correspondence Analysis of contingency tables.

A new window named **"Reordering"** appears.

Click on the button: "Reordering the rows and the column of a word-text table" .

The reordered table cross-tabulating the 9 categories and the selected words is then displayed. It can be seen that the first words of the reordered list of words characterize (sometimes exclusively) the first categories in the reordered list of categories.



The last words of the same list are either absent or rarely observed among these categories. However, they are frequent among the last categories (right hand side of the table).

General remark: As you can observe when looking at the content of the example's directory, several files have been created and saved [these files are briefly described in the memo **"Help about files"** in the toolbar of the main menu]. If you need to continue using again the buttons of the paragraph VIC of the main menu after having closed DtmVic, just click on the button **"Open an existing command file"** from the line "command file" , select and open the saved command file: **"Param_ANALEX.txt"** , and close it. It is not necessary to click on: **"Execute a command file"** again. You can then continue your investigation (axes views, graphs, maps, etc.). The advanced users can also edit the parameter file: **"Param_ANALEX.txt"** , (using the memo **"Help about parameters"** in the toolbar of the internal editor) to perform a new analysis in which the parameters are given new values. It is advised to give it a new name (such as: **"Param_ ANALEX2.txt"** , for example). All the intermediate files will be replaced (except the files **"imp_date_time.txt"** and **"imp_date_time.html"** which are the only saved archives).

End of example A.5

III.3 Open questions in a sample survey: Direct analysis and link with closed-end questions

Example A.6: EX_A06.Text-Responses_2.

Example A.6 aims at describing directly the responses to an open ended question in a sample survey, without prior agglomeration, in relation with a set of categorical variables[3]. The survey and the responses are the same as in examples A.5.

We can take advantage of the presence of closed-end questions to describe the clusters, not only with characteristic words and responses, but also with categories, selected after a step SELEC, and analysed through the step DECLA .
Another new step of the command file, POSIT, describes the location of these supplementary categories in the plane spanned by the first principal axes.

1) Looking at the three files: data, dictionary and texts.

To have a look at the data, search for the directory: **DtmVic_Examples.**
In this directory, open the sub-directory **DtmVic_Examples_B_Texts** .
In that directory, open the directory of Example A.6, named: **"EX_A06.Text-Responses_2 "** .

It is recommended to use one directory for each application, since DtmVic produces a lot of intermediate txt-files related to the application.

At the outset, such directory must contain at least 3 files :

a) the data file,
b) the dictionary file,
c) the text file,

a) Data file: TDA_dat.txt (same as that of Example A.5)
This file contains responses to questions which were included in the multinational survey conducted in seven countries (Japan, France,

[3] More explanation about this type of example and the corresponding methodology can be found in the book: "Exploring Textual data" (L. Lebart, A. Salem, L. Berry; Kluwer Academic Publisher, 1998).

Germany, United Kingdom, USA, Netherlands, Italy) in the late nineteen eighties (Hayashi *et al*., 1992). It is the United Kingdom survey which is presented here.
It deals with the responses of 1043 individuals to 14 questions. Some questions concern objective characteristics of the respondent or his/her household (age, status, gender, facilities). Other questions relate to attitude or opinions. The data file: **" TDA_dat.txt"** comprises 1043 rows and 15 columns (identifier of rows [between quotes] + 14 values [corresponding either to numerical variables or to item numbers of categorical variables] separated by at least one blank space).

b) Dictionary file: TDA_dic.txt (same as that of Example A.5)
The dictionary file **"TDA_dic.txt"** contains the identifiers of these 14 variables. In this version of Dtm-Vic, the identifiers of categories must begin at: "column 6" [a fixed interval font - also known as teletype font - such as "courier" should be used to facilitate this kind of format].

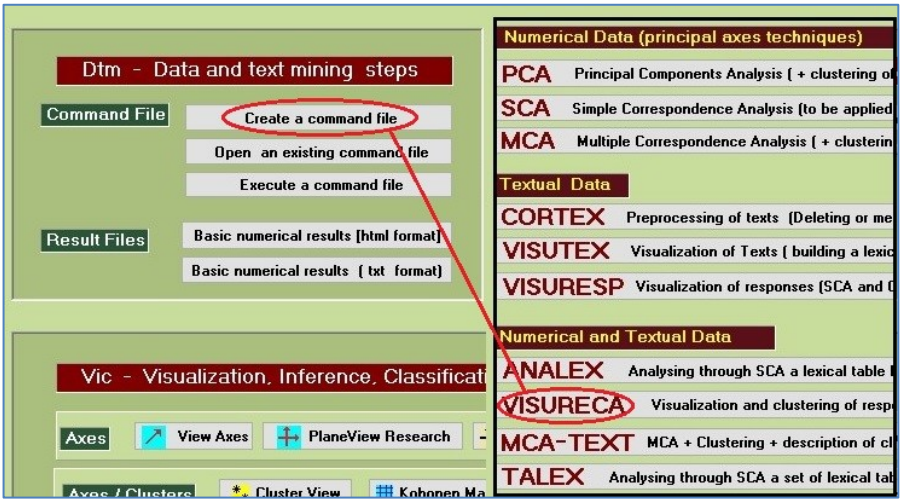
c) Text file: TDA_TEX.txt (same as that of examples A.5)
Let us remind its characteristics. It contains the free responses of 1043 individuals to three open-ended questions.
Firstly, the following open-ended question was asked: *"What is the single most important thing in life for you?"* It was followed by the probe: *"What other things are very important to you?"*.
A third question (not analysed here) has also been asked: *" What means to you the culture of your own country"*. We refer to the previous example (example A.5) for comments about the data format.

2) Generation of a command file (or: "parameter file")

2.1)Click the button: "Create a command file" of the main menu, line: " Command File".

A window: **"Choosing among some basic analyses"** appears.

2.2) Click then on the button: "VISURECA analysis - (Visualization and Clustering of responses with supplementary categorical data) " – located in the section **"Textual and numerical data"**. A window **"Opening a text file"** is displayed



2.3) Press the button:"Open a text file", then search for the directory: " **DtmVic_Examples_A_Start"** . In that directory, open the directory of example A.6, named **"EX_A06.Text-Responses_2"** .
Open then the text file: **"TDA_tex.txt "**.
A message box indicates then that the corpus comprises 7329 lines, 1043 observations and 3 open questions.

2.4) Click on: "Select Open questions and separators"
The next window allows for the selection of open questions and the selection of separators of words (the default separators suffice in this example).
We will select questions 1 and 2 (that means that the two responses will be merged). It is licit here to merge the two responses because question 2 is a probe for question 1.

2.5) Click directly on: "Vocabulary and counts" .
The next window presents the vocabulary (alphabetic and frequency orders). We must select a threshold of frequency by selecting a line in the right hand side memo frequency order). The line number 397 [first column] corresponds to the frequency 4 [second column]. (We took a threshold of 16 in the previous example A5. For individual responses, lexically very poor, it takes more words not to generate too many empty answers after choosing the threshold). We keep the 397 most frequent

words. After selecting that line, click on: **"Confirm".** The frequency appears in a message box. Reply: "OK".
Then click on: **"Continue"** . A window **dictionary and data files** appears.

2.6) Click the button: "Open a dictionary (Dtm format)"

Open then the dictionary file: **"TDA_dic.txt"** .
The dictionary file: **TDA_dic.txt** contains the identifiers of the 14 variables.
The dictionary file in displayed in a window. Another window indicates the status of each variable (numerical or categorical).

2.7) Press the button: "Open a data file (Dtm format)"

Open the data file: **"TDA_dat.txt"**
That data file comprises 1043 rows and 15 columns (identifier of rows [between quotes] + 14 values [corresponding either to numerical variables or to item numbers of categorical variables] separated by at least one blank space).
A new window displays the data file.

2.8) Click the button: "Continue (select active and supplementary variables)".

A new window is displayed, allowing for the selection of active variables. There is no active variable, since the responses to the 2 open questions are active here. We actually chose the active variables by selecting the open-ended questions 1 and 2.
All the remaining variables could be selected as supplementary elements. They will serve to describe the categories of the active variable.

2.9) Click then on the button: "Continue"

A new window devoted to the selection of active observations (rows) is displayed.
Click on the button: **"All the observations will be active"** .

The window: **"Create a starting parameter file"** is displayed.

2.10) Then click directly on: "Create a first parameter file".

For this type of analysis, there is an implicit bootstrap validation. The selected bootstrap option is the specific partial bootstrap (see the reminders). The clustering is automatic, and the number of clusters is selected (default) depending on the number of responses (30 clusters in this case). [This number of cluster can be changed by editing the command file (or parameter file) before the execution, the parameters to be altered belong to the "STEP PARTI" and "STEP DECLA"].

A parameter file is displayed in the memo [It can be edited by the advanced users. It allows for performing again the same analysis later on, if needed].

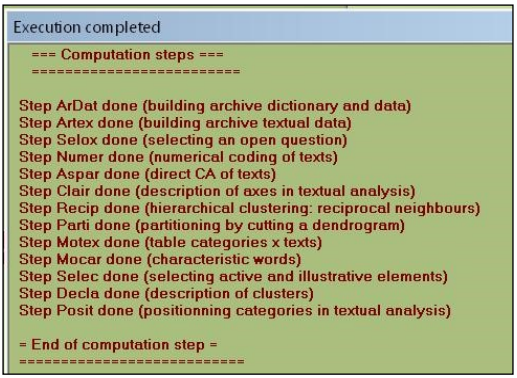
> Important: *The parameter file is saved as* **"Param_VISURECA.txt"** *in the current directory. If you wish, you could now exit from DtmVic, and, later on, use the button of the main menu* **"Open an existing command file"** (Section: **"Command file")** *to open directly the file* **"Param_VIRURECA.txt"**, *and, in so doing, reach directly this point of the process, using* ***the*** **"Execute a command file"** *command of the main menu.*

Let us remind that this set of commands comprises 14 steps:

ARDAT (archiving data),
ARTEX (Archiving texts)
SELOX (selecting the open question),
NUMER (numerical coding of the text),
ASPAR (correspondence analysis of the [sparse] contingency table "respondents - words"),
CLAIR (Brief description of factorial axes),
RECIP (Clustering using a hierarchical classification of the clusters - reciprocal neighbours method),
PARTI (Cut of the dendrogram produced by the previous step, and optimisation of the partition obtained),
MOTEX (crosstabulating the partition produced by step PARTI with words: the obtained contingency table is called a lexical table),
MOCAR (characteristic words, and characteristics responses for each class of the partition),
SELEC (selecting active and supplementary elements),
DECLA (systematic description of the classes of the partition produced by step PARTI using the other relevant categorical variables),
POSIT (illustrating the principal spaces of responses with supplementary categorical variables).

2.11) Click: "Execute a command file".

This step will run the basic computation steps present in the command file: archiving data and text, characteristic words and responses, correspondence analysis of the lexical table, thorough descriptions of clusters using both words and categorical variables.



Recap of the executed steps

3) Click the button: "Basic numerical results"

The button opens a created (and saved) html file named **"imp.html"** which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu. Note that this file is also saved under another name. The name "imp.html" is concatenated with the date and time of the analysis (continental notation). That file keeps as an archive the main numerical results whereas the file "imp.html" is replaced for each new analysis performed in the same directory.

This file in also saved under a simple text format, under the name **"imp.txt"** , and likewise with a name including the date and time of execution.

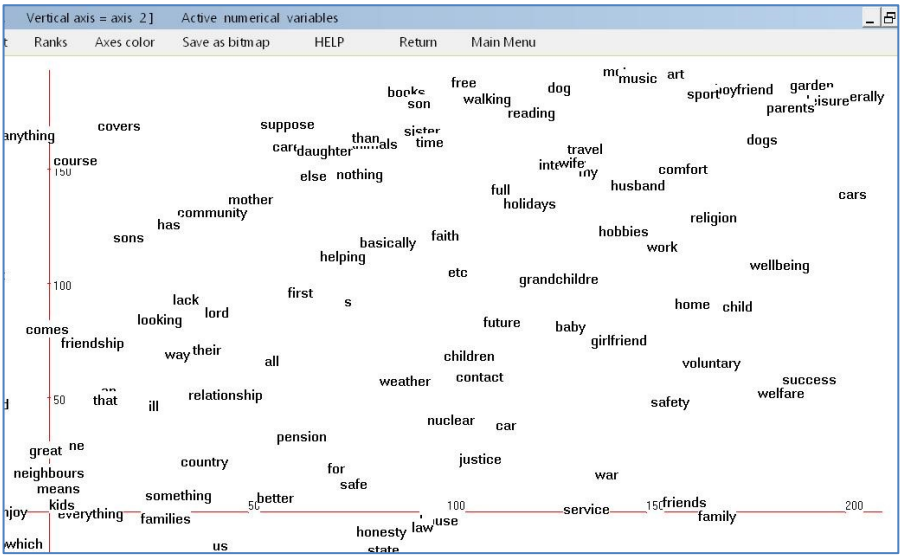
Perusing the complete list of words highlights some errors in the original text file (inevitable in real sized applications) : for instance, the symbol "]" was absent from the list of separators, and creates some new "words"…

4) At this stage, we click on one of the lower buttons of the basic steps panel (Steps: "VIC")

5) Click the button "ViewAxes"

... and follow the sub-menus. In fact, three tabs are relevant for this example: **"Active variables"** [= words in the case of the analysis: ""VISURECA"], **"Individuals (observations)** [= respondents]" and **"Supplementary Categories"**. After clicking on **"ViewAxes"** in each case, one obtains the set of principal coordinates along each axis. Clicking on a column header produce a ranking of all the rows according to the values of that column. In this particular example, this is somewhat redundant with the printed results of the step CLAIR ". **Return.**

6) Click the button: PlaneView Research , and follow the sub-menus...
In this example, six items of the menu are relevant **"Active columns (variables or categories)"** (principal coordinates of the active words), **"Supplementary categories"** (coordinates of the supplementary categories derived from the step " POSIT "), **"Active rows (individuals, observations)"**, (coordinates of the respondents), **"Active columns + Active rows"**, **"Active individuals (density)"** and **"Active columns + Supplementary categories"**. The graphical displays of chosen pairs of axes are then produced.
Return.



Upper right part of the plane (after clicking "Zoom") in which the coordinates have been converted into ranks (button "Rank") to obtain a more legible display (the display contains 398 words).

7) Click on the button: "BootstrapView"

This button opens the **"DtmVic: Bootstrap - Validation - Stability – Inference"** windows.

7.1 Click on: **"LoadData"** . In this case (specific partial bootstrap), the replicated coordinates file to be opened is named: **"ngus_dir_var_boot.txt"** . (The set of possible files is given by the panel).

7.2 Click on: **"Confidence Areas"** submenu, and choose the pair of axes to be displayed (select axes 1 and 2 to begin with).

7.3 We obtain the list of the identifiers of active columns (words). Tick some cases to select some words, and press the button: **"Select"**.

– Click on: **"Confidence Ellipses"** to obtain the graphical display of the chosen column points.

– Close the display window, and, again in the blue window, press: **"Convex hulls"**. The ellipses are now replaced with the convex hulls of the replicates for each point. The convex hulls take into account the peripheral points, whereas the ellipses are drawn using the density of the clouds of replicates. The two pieces of information are complementary. **Return.**

8) Click on "ClusterView "

8.1 Choose the axes (1 and 2 to begin with), and "Continue".

8.2 Click on **"View"**. The centroids of the 30 clusters (produced by the Step PARTI) appears on the first principal plane.

8.3 Activate the button: **"Words"**, and , pointing with the mouse on a specific cluster, press the right button of the mouse. A description of the cluster involving the most characteristic words of the cluster appears. This description is somewhat redundant with that of the Step MOCAR . But this display exhibits the pattern of clusters and their relative locations.

8.4 Activate the button **"Texts"**. Pointing with the mouse on a specific cluster, and pressing the right button of the mouse, we can read the most characteristic responses of the selected cluster.

8.5 Activate the button: **"Categorical"**. Pointing with the mouse on a specific category, and pressing the right button of the mouse, we can read the most characteristic categories of the selected category. This description is somewhat redundant with that provided in the results file (file "imp.txt)

by the step DECLA. But we do have simultaneously in front of us the pattern of categories and their relative locations.

9) Click on "Kohonen map"
Select the type of coordinate.

9.1 Select: **"Variables (columns)"**: these active variables are the words in this example.

9.2 Select a (5 x 5) map, and continue.

9.3 After clicking on two small check-boxes, press **"Draw"** on the menu of the large green windows entitled Kohonen map.

9.4 You can change the font size (**"Font"**) and dilate the obtained Kohonen map (**"Dilat."**) to make it more legible. The words appearing in the same cell are often associated in the same responses. This property holds, at a lesser degree, for contiguous cells.

9.5 Pressing **"AxeView"**, and selecting one axis allows one to enrich the display with pieces of information about a specified principal axis : large positive coordinates in red colour, large negative coordinates in green, with some transitional hues.

9.6 Go back to the main menu, click on **"Kohonen map"** and choose the item **"Observations"**.

9.7 Select a (10 x 10) map, and redo the operations 9.3 to 9.5 for the observations.

End of example A.6

Chapter IV

Three more examples to practise DtmVic with textual data

Unlike chapters II and III, chapter IV contains examples which use existing command files (or: parameter files). The examples correspond to the directory **"DtmVic_Examples_B_Texts"** *that has been downloaded with DtmVic.*

(Application examples B.1—B.3)

IV.1 Open questions in a survey: First exploration

Example B.1. EX_B01.Text-Responses_Corda

First processing of the responses to an open-ended question. Examples of modification of the frequency threshold for words. Example of concordances (syntactic context) for some words. Correspondence Analysis (CA) of the sparse lexical table words x respondents, clustering of the responses, and description of the obtained clusters through their characteristic words and responses. [Similar to the analysis VISURESP in the menu "Create a command file"]

IV.2 Open questions and MCA in a sample survey

Example B.2. EX_B02.Text-Responses_MCA

Multiple Correspondence Analysis and Clustering of respondents using closed questions. Processing of aggregated [and lemmatised] responses to open questions. Example B.3 illustrates another technique for grouping and processing responses to open question in a sample survey. In a first phase, a multiple correspondence analysis is performed on a set of selected categorical variables (i.e.: responses to closed-end questions). The MCA is complemented with a clustering, followed by an automatic description of the clusters. These clusters are then used to aggregate the responses to an open question.

IV.3 Analysis of the Semantic network of French verbs

Example B.3. EX_B03.Text-Semantic.

Visualization of the semantic links existing between 829 French verbs. Each verb is described by a list of "synonyms". The "respondents" are here the 829 verbs. The (fictitious) open-ended question is "Which are your synonyms ?", and the textual response is constituted by a list of synonyms.

IV.1 Open questions in a survey: First exploration

Example B.1: EX_B01.Text-Responses_Corda

Example B.1 aims at describing the responses to an open-ended question in a sample survey. The principal axes visualization is complemented by a clustering, with an automatic description of the clusters. Example of modification of the frequency threshold for words. Example of concordances (syntactic context) for some words. This is a typical first outlook on the set of responses: to detect and describe the main groupings of responses. Such outlook is by no means an achieved processing.
Example A.6, above, provided another point of view, making use of other pieces of information about the respondents.
To have a look at the data, search for the directory **DtmVic_Examples.**
In this directory, open the sub-directory **DtmVic_Examples_B_Texts**.
In that directory, open the directory of Example B.1, named **"EX_B01.Text-Responses_Corda"** .
It is recommended to use one directory for each application, since DtmVic produces a lot of intermediate txt-files related to the application. At the outset, such directory must contain only 2 files :

- a) the text file,
- b) the command file.

(in this particular context, there are neither data file nor dictionary file: the questionnaire comprises three open-ended questions, without considering the closed-end questions)

a) Text file: TDA_TEX.txt

This file has already served as an example for Examples A.5 and A.6 Chapter 3. It contains the free responses of 1043 individuals to three open-ended questions.
Firstly, the following open-ended question was asked: *"What is the single most important thing in life for you?"*
It was followed by the probe: *"What other things are very important to you?".*
A third question has also been asked: *"What means to you the culture of your own country"* We analyse here the responses to this third question.
See examples A.5 and A.6 for a description of both data and corresponding files.

b) Command file: "EX_B01_Param.txt"

As shown in Chapter 3, another "command file" similar to the "command file " **"EX_B01_Param.txt"** can be also generated by clicking on the button: **"Create a command file"** of the main menu (Basic Steps). A window **"Choosing among some basic analysis"** appears. Click in this case on the button: **"VISURESP– Visualization of Responses"** – located in the paragraph **"Textual data"**, and follow the instructions as indicated in Chapter 3.
The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: **"Help about command parameters"**) and, with more details, below (Appendix B.1).

Running the example B.1 and reading the results

1) Click on the button: "Open an existing command file" (main menu)

2) Then, search for the sub- directory:

DtmVic_Examples_B_Texts in: **DtmVic_Examples.**

3) In that directory, open the directory of Example B.01: " EX_B01.Text-Responses_Corda" .

4) Open the command file:EX_B01_par.txt

After identifying the textual data file, 11 "steps" are performed:

ARTEX (Archiving texts),
SELOX (selecting the open question),
NUMER (numerical coding of the text: now, all the words are kept),
CORDA (concordance for some selected words),
SETEX (introducing a new threshold for the frequencies of words),
ASPAR (correspondence analysis of the [sparse] contingency table "respondents words"),
CLAIR (Brief description of factorial axes),
RECIP (Clustering using a hierarchical classification of the clusters - reciprocal neighbours method),
PARTI (Cut of the dendrogram produced by the previous step, and optimisation of the partition obtained),
MOTEX (crosstabulating the partition produced by step PARTI with words: the obtained contingency table is called a lexical table),
MOCAR (characteristic words, and characteristic responses for each class of the partition).

We will comment later on this command file (Appendix of the section) which commands the basic computation steps. Instead of editing this file, we directly go back to the main menu and execute the basic computation steps.

5) Return to the main menu (“Return to execute”)

6) Click on the button:“Execute a command file”

This step will run the basic computation steps present in the command file: archiving text, correspondence analysis of the lexical table, brief description of the axes, clustering procedure, thorough descriptions of clusters using characteristic words and responses.

7) Click the button: “Basic numerical results”

The button opens a created (and saved) html file named **“imp.html”** which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu.

Note that this file is also saved under another name. The name “imp.html” is concatenated with the date and time of the analysis (continental notation): “imp_06.07.12_14.45.html” means June 8th, 2012, at 2:45 p.m. That file keeps as an archive the main numerical results whereas the file: “imp.html” is replaced for each new analysis performed in the same directory. This file in also saved under a simple text format , under the name **“imp.txt”** , and likewise with a name including the date and time of execution.

From the step NUMER, we learn for instance that we have 1043 responses, with a total number of words (occurrences or token) of 9148, involving 1629 distinct words (or: types) . Using a frequency threshold of 8 (see STEP SETEX in the command file below) the total number of kept words reduces to 11559, whereas the number of distinct kept word reduces (drastically) to 170.
From the step CORDA, we can observe the contexts of the selected words (see the command file in appendix B.1) *life, money, love, museum, fish*. Note that from the button "**Create a command file**" of the main menu, we can build the command file leading to the step **"CORDA"** (button: **"Other analyses"** and button **"CORDA"**).

8) At this stage, we click on one of the lower buttons of the basic steps panel (Steps: “VIC”)

9) Click the button: “ViewAxes”

and ... follow the sub-menus. In fact, only two tabs are relevant for this example: **“Active variables”** [= words in the case of step ASPAR] , **“Individuals**

(observations) [= respondents]" . After clicking on **"View"** in both cases, one obtains the set of principal coordinates along each axis.
Clicking on a column header produce a ranking of all the rows according to the values of that column. In this particular example, this is somewhat redundant with the printed results of the step "CLAIR". Evidently, the use of the ViewAxes menu is justified when the data set is large, which is the case here. **Return.**

10) Click the button: "PlaneView Research"

and follow the sub-menus...
In this example, four items of the menu are relevant: **"Active columns (variables or categories)", "Active rows (individuals, observations)", "Active columns + Active rows", " Active individuals (density)".** The graphical displays of chosen pairs of axes are then produced.
The roles of the different buttons are straightforward, except perhaps the button: **"Rank"**, which is useful only in the case of very intricate displays, (which is the case here). Since the set "individual" has 1043 elements, it is possible to test, with this example, partial printings of the individuals in two subsets of 50% or four subsets of 25%…(subsets randomly drawn without replacement). **Return.**

11) Click on the button: "BootstrapView"

This button opens the **"DtmVic: Bootstrap - Validation - Stability – Inference"** windows.

11.1 Click on: **"LoadData"** . In this case (partial bootstrap), the replicated coordinates file to be opened is named: **"ngus_dir_var_boot.txt"**. (The set of possible files is given by the panel).

11.2 Click on: **"Confidence Areas"** submenu, and choose the pair of axes to be displayed (select axes 1 and 2 to begin with).

11.3 We obtain the list of the identifiers of active columns (words). Tick some cases to select some words, and press the button: **"Select"**.

– Click on: **"Confidence Ellipses"** to obtain the graphical display of the chosen column points.

– Close the display window, and, again in the blue window, press: **"Convex hulls"**. As in the previous examples, the ellipses are now replaced with the convex hulls of the replicates for each point. **Return.**

12) Click on: "ClusterView"

12.1 Choose the axes (1 and 2 to begin with), and **"Continue"**.

12.2 Click on: **"View"**. The centroids of the 12 clusters (Step PARTI) appears on the first principal plane.

12.3 Activate the button: **"Words"**, and, pointing with the mouse on a specific cluster, press the right button of the mouse. A description of the cluster involving the most characteristic words of the cluster appears. This description is somewhat redundant with that of the Step MOCAR. But we do have in front of us the pattern of clusters and their relative locations.

12.4 Activate the button: **"Texts"**. Pointing with the mouse on a specific cluster, and pressing the right button of the mouse, we can read the most characteristic responses of the selected cluster.

13) Click on: "Kohonen map".

Select the type of coordinate.

13.1 Select: **"Active variables (columns)"**: these active variables are the words in this example.

13.2 Select a (5 x 5) map, and continue.

13.3 After clicking on two small check-boxes, press: **"Draw"** on the menu of the large green windows entitled Kohonen map.

13.4 You can change the font size (**"Font"**) and dilate the obtained Kohonen map: (**"Dilat."**) to make it more legible. The words appearing in the same cell are often associated in the same responses. This property holds, at a lesser degree, for contiguous cells.

13.5 Pressing **"Axes View"**, and selecting one axis allows one to enrich the display with pieces of information about a specified principal axis: large positive coordinates in red colour, large negative coordinates in green, with some transitional hues.

13.6 Go back to the main menu, click on **"Kohonen map"** and choose the item: **"Active observations"**.

13.7 Select a (10 x 10) map, and redo the operations 13.3 to 13.5 for the observations.
In the context of this example, the other items of the menu are not relevant.

Appendix B1: *(for advanced users)*

The command file can be generated using the button: "Create_a command file" (The involved analyses are "VISUESP" and "CORDA"**) Therefore, freshman practitioners could skip this appendix.**

The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: **"Help about parameters"**).

Now, we exhibit the command file that contains comments (preceded by #). As seen previously, comments are also allowed in the (mandatory) line that immediately follows a statement "STEP xxxxx"

Command file : "EX_B01_Param.txt"

```
# The Program DtmVic needs 2 files in this "open survey case"
# ----------------------------------------
# 1) The present file of commands, whatever its name.
# 2) The text file (NTEXZ).
#   Syntax:  ">"= continuation, "#"= comments
#-----------------------------------------------------------

LISTP = yes, LISTF = no  #  Global parameters(leave as it is)
#
NTEXZ = 'TDA_tex.txt'  # name of text file (free name)
#
STEP ARTEX
==== Archive - Texts or responses to open ended questions
ITYP=2  NBQT=3

#---------------------- Comments about step ARTEX
# - ITYP:  type of textual data file NTEXZ
#  ITYP = 2 ==> type of file = responses to open questions
# - NBQT: number of questions per respondent
#    NBQT = 3 ==> there are 3 open questions
#-----------------------------------------------------------
#
STEP SELOX
==== Selection of open questions (and of individuals)
NUMQ = LIST
3
```

```
#---------------------- Comments about step SELOX
# - NUMQ: index of the selected question
#     if NUMQ = -1 or NUMQ = LIST : several questions
#    will be merged (the list of question numbers
#    must follow immediately next line)
#    here: question 3 is selected
#-------------------------------------------------------------

STEP NUMER
==== Numerical coding of words
NSEU = 0 LEDIT=TOT
weak  -
strong . ; : ( ) ! ? ,
end
#---------------------- Comments about step NUMER
# - NSEU:  frequency threshold of the kept words
#    (here, only the frequencies > 8 will be kept)
# - LEDIT: printing the words (0=no; 1=alphabetical order;
#    2=frequency order; 3= both 1 and 2).
# --- key-words:
# - weak   (weak separators) followed by those separators
#    [separators of words]
# - strong (strong separators) followed by those separators
#    [separators of segments, for step SEGME]
# - end ... indicates the end of key-words statements.
#-------------------------------------------------------------

STEP CORDA  # concordances
==========
LEDIT = 1
FORME  life  money love museum fish
END
#---------------------- Comments about step CORDA
#LEDIT: printing identifiers of individuals
#     (0 = no printing, 1 = identifiers of
#      respondents are printed, default = 0)
# --- key-word of headings :
#  FORME  must be followed by the selected words
#  END  end of the selection.
#-------------------------------------------------------------

#---- selecting a new threshold for words (SETEX) -

NSPB ='NSPB'

# the file NSPB created by SETEX is given the name: 'NSPB'

STEP SETEX
===========  Change of threshold for the frequency of words
NSEU =8 NMOMI=0 NREMI=2 LEDIT =NEW
#---------------------- Comments about step SETEX
# NSEU:  threshold of frequency for selecting words.
# NMOMI: minimum number of letters of a kept word.
# NREMI: minimum number of words of a kept response.
# LEDIT: printing the dictionaries (0=no, 1=new, 2=tot).
#-------------------------------------------------------------

NSPA = 'NSPB'
```

```
#----- the file 'NSPB' created by SETEX  is substituted to
# the file NSPA that was created by NUMER.


STEP ASPAR
==== Correspondence analysis of the table: Words X Responses
NAXE=8 LEDIT=0 NGRAF=5 NROWS=60 NPAGE=1 NBASE=12 NITER=20
#---------------------- Comments about step ASPAR
# - NAXE:  number of requested principal coordinates
# - LEDIT: printing the responses
#      (0 = no; 1 = coordinates of variables;
#       2 = 1 + coordinates of respondents)
# - NGRAF: number of requested printer graphics
#    in the results file "imp.txt"
#   NGRAF = 5 means that we will get the printouts of
#    the planes spanned by the following pairs of axes:
#    (1, 2), (2, 3), (3, 4), (4, 5), (4, 6).
# - NPAGE: number of pages of these graphics
# - NROWS: number of lines of these graphics
#  The two following parameters concern an option
#  for diagonalizing very large matrices: (if NBASE > 0)
# - NBASE: dimension of the approximation space
#      (NBASE = 0: main core diagonalization)
# - NITER: number of iterations (if NBASE > 0)
#-----------------------------------------------------

STEP CLAIR
==== Brief description of NAXE principal axes
NAXE=6 LIGN=no NMAX=40
#---------------------- Comments about step CLAIR
# - NAXE = ... number of axes to be described
# - LIGN = no means that lines (or rows, or individuals
#    or respondents are excluded)
# - NMAX = ... Maximum number of elements that will
#    be sorted to describe each axis
#-----------------------------------------------------

STEP RECIP
==== Clustering of respondents using reciprocal neighbours
NAXU=7 LDEND=DENSE NTERM=20 LDESC=no
#---------------------- Comments about step RECIP
# This step carries out a hierarchical clustering
# using the reciprocal neighbours technique (recommended
# when dealing with less than 1000 individuals.
# - naxu...  number of axes kept from the
#  previous MCA .
#  - LDEND... printing dendrogram (0=no, 1=dense,
#  2=large).
# - nterm... number of kept terminal elements
#  NTERM = TOT means that all the elements are kept.
# - LDESC... describing nodes of the tree (0=no, 1=yes).
#-----------------------------------------------------

STEP PARTI
==== Cut of the dendrogram to obtain 9 clusters
NITER=7 LEDIN=3
12  # number of classes of the partition
#---------------------- Comments about step PARTI
# - NITER... number of "consolidation" iterations (0=no).
```

```
#  - LEDIN... printing the correspondences classes-
# individuals (3 = printing of the correspondence
# classes->individuals and the correspondence
# individuals-->classes).
# The line immediately following the command must
# contain the sizes of the desired final partition
# (here: 9).
#-----------------------------------------------

STEP MOTEX
==== cross-tabulating words and clusters
NVSEL=-1 LEDIT = 0
#---------------------- Comments about step MOTEX
#  NVSEL: index of the categorical variable defining
#    the groupings of texts
#    the conventional value NVSEL = -1 means that
#    the categorical variable coincides with the
#    previously computed partition.
#  LEDIT: parameter for printing the table words*texts
#    (0=no, 1=yes).
#-----------------------------------------------

STEP MOCAR
==== Characteristics words for each cluster
NOMOT=10   NOREP=6
#---------------------- Comments about step MOCAR
#  NOMOT: number of requested characteristic words  for
#     each text (i.e: for each cluster)
#  NOREP: number of characteristic responses for each text.
# MOCAR considers as a characteristic response for a category
# a response containing as many characteristic words as possible.
# (with penalties for anti-characteristic words).
#-----------------------------------------------
STOP
#------------------------------------------------------------
```

End of example B.1

IV.2 Open questions and MCA in a survey

Example B.2. EX_B02.Text-Responses_MCA

Example B.2 illustrates another technique for grouping and processing responses to open question in a sample survey. In a first phase, a multiple correspondence analysis is performed on a set of selected categorical variables (i.e: responses to closed-end questions). The principal axes visualisation is complemented with a clustering, followed by an automatic description of the clusters. These clusters are then used to aggregate the responses to an open question. The survey, the closed-end questions and the textual responses are the same as those of previous examples[4].

The sequence of steps is enriched by the following computations:
As in Example B.1, the numerical coding (step NUMER) is performed with a frequency threshold of 0 : all the words (types) are kept. We can then carry out the new step CORTE , allowing us to perform a "primary lemmatization" of the text. (see also the procedure **CORTEX** from the menu invoked by the button "Create" of the main menu, and the complementary command files whose names ends by "TEX").
We can now take advantage of the presence of both open-ended and closed-end questions to describe the clusters, not only with characteristic words and responses (as done previously in Example B.1), but also with categories. Another new step: POLEX , describes the location of the words in the plane spanned by the first principal axes.

To have a look at the data, search for the directory: **DtmVic_Examples.**
In this directory, open the sub-directory **DtmVic_Examples_B_Texts** .
In that sub-directory, open the directory of Example B.2, named **"EX_B02.Text-Responses_MCA"** .

[4] More explanation about this type of example and the corresponding methodology can be found in the book: "Exploring Textual data" (L. Lebart, A. Salem, L. Berry; Kluwer Academic Publisher, 1998).

It is recommended to use one directory for each application, since DtmVic produces a lot of intermediate txt-files related to the application.

At the outset, such directory must contain 4 files :

- a) the data file,
- b) the dictionary file,
- c) the text file,
- d) the command file.

---

a) Data file: TDA_dat.txt (same as that of Example B.2)

This file contains responses to questions which were included in the multinational survey [see also Examples A.5 and A.6] conducted in seven countries (Japan, France, Germany, United Kingdom, USA, Netherlands, Italy) in the late nineteen eighties (Hayashi *et al*., 1992)..

The data file **" TDA_dat.txt"** comprises 1043 rows and 15 columns (identifier of rows [between quotes] + 14 values [corresponding either to numerical variables or to item numbers of categorical variables] separated by at least one blank space).

b) Dictionary file: TDA_dic.txt (same as that of Examples A.5 and A.6)
The dictionary file **"TDA_dic.txt"** contains the identifiers of these 14 variables. In this version of DtmVic, the identifiers of categories must begin at: "column 6" [using a fixed interval font such as "courier"].

c) Text file: TDA_TEX.txt (same as that of examples A.5, A.6, and B.1)
We refer to previous example for comments about the questionnaire and the data format.

d) Command file: EX_B02_Param.txt
The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: **"Help about command parameters"**) and, with more details, below.

Note that another "command file", similar to the provided "command file **"EX_B02_Param.txt,** can be also generated by clicking on the button: **"Create a command file"** of the main menu (Basic Steps). A window **"Choosing among some basic analysis"** appears. Click then on the button: **"MCA_Texts – Visualization of Responses"** located in the paragraph **"textual and numerical data"**, and follow the instructions.

Running the example B.2 and reading the results

1) Click on the button: "Open an existing command file" (Main menu)

2) Then, search for the sub-directory: "DtmVic_Examples_B_Texts" in **"DtmVic_Examples".**

3) In that sub-directory, open the directory of Example B.2 named "EX_B02.Text-Responses_MCA"

4) Open the existing command file: EX_B02_Param.txt.
After identifying the textual data file, 16 "steps" are performed:
ARDAT (archiving data),
ARTEX (Archiving texts),
SELOX (selecting the open question),
NUMER (numerical coding of the text: now, all the words are kept),
CORTE (deleting some function words [or empty words], declaring as equivalent flections of a same lemma),
SETEX (introducing a new threshold for the frequencies of words),
SELEC (selecting active and supplementary elements),
MULTM (Multiple correspondence analysis),
DEFAC (Brief description of factorial axes),
POLEX (projecting the words of the responses as supplementary elements in the principal planes),
RECIP (Clustering using a hierarchical classification of the clusters - reciprocal neighbours method),
PARTI (Cut of the dendrogram produced by the previous step, and optimisation of the partition obtained),
DECLA (systematic description of the classes of the partition produced by step PARTI using the other relevant categorical variables),
MOTEX (crosstabulating the partition produced by step PARTI with words: the obtained contingency table is a "lexical table"),
MOCAR (characteristic words, and characteristic responses for each class of the partition),
RECAR (characteristic responses for each class of the partition using a different criterion of selection, allowing for lengthy responses).

We will comment later on this command file (Appendix B.2 of the section) which commands the basic computation steps. Instead of editing this file, we will directly go back to the main menu and execute the basic computation steps.

5) Return to the main menu ("Return to execute")

6) Click on the button: "Execute a command file"
This step will run the basic computation steps present in the command file.

7) Click the button: "Basic numerical results"
The button opens a created (and saved) html file named **"imp.html"** which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu.

Note that this file is also saved under another name. The name "imp.html" is concatenated with the date and time of the analysis (continental notation). That file keeps as an archive the main numerical results whereas the file "imp.html" is replaced for each new analysis performed in the same directory. This file in also saved under a simple text format , under the name **"imp.txt"** , and likewise with a name including the date and time of execution.

From the step NUMER , with the new threshold of "0", we check for instance that we still have 1043 responses, with a total number of words (occurrences or token) of 13 918, involving 1 368 distinct words (or: types). In this version of DtmVic, the results of the new step CORTE are confined to this "result file". **Return.**

8) At this stage, we click on one of the lower buttons of the basic steps panel (Steps: "VIC")

9) Click the button "ViewAxes"
and ... follow the sub-menus. Here, four tabs are relevant for this example: **"Active variables"** [= categories in this MCA case], **"Supplementary categories", "Individuals (observations) [= respondents]"** , and **"supplementary lexical units"** (provided by step POLEX = projections of words onto the axes of the MCA). After clicking on " **View** " in both cases, one obtains the set of principal coordinates along each axis. Clicking on a column header produce a ranking of all the rows according to the values of that column. **Return.**

10) Click the button: "PlaneView Research"
and follow the sub-menus...

In this example, seven items of the menu are relevant **"Active columns (variables or categories)"** ,(Active categories of the Multiple Correspondence Analysis), **"Supplementary categories"** , (Supplementary categories of the same MCA), **"Active rows (individuals, observations)"**, **"Active columns + Active rows", "Supplementary lexical units"** (projection of the words used by the respondent in their responses to the open question), provided by step POLEX), **"Active individuals (density)"** and **"Active columns + Supplementary categories"** . The graphical displays of the chosen pairs of axes are then produced.
Return.

11) Click the button: "BootstrapView" [The Bootstrap concerns here the Phase of Multiple Correspondence Analysis. See example A3]

This button opens the DtmVic-Bootstrap-Stability windows.

11.1 Click: "**LoadData**" . In this case (partial bootstrap), the two replicated coordinates file to be opened are named **"ngus_var_boot.txt"** and **"ngus_sup_cat_boot.txt"** (see the panel reminding the names of the relevant files below the menu bar).
In fact, **ngus_var_boot.txt** contains both active and supplementary categories. The file **ngus_sup_cat_boot.txt** contains only supplementary categories, for which the bootstrap procedure is all the more meaningful.

11.2 Click on **"Confidence areas"**, submenu, and choose the pair of axes to be displayed (choose axes 1 and 2 to beginwith).

11.3 Tick the chosen white cases to select the elements the location of which should be assessed, and press the button **"Select"**. Select, for instance, the supplementary elements "male, female, less than 30 years old with high level of education, over 55 with high, and also with low level of education.

11 .4 Click on: **"Confidence Ellipses"** to obtain the graphical display of the active category points (in blue colour), and of the supplementary category points (in red).
In this display, we learn for example that in this principal space (built as a "space of opinions", due to the selection of active questions), male and female do not occupy statistically distinct locations (ellipses overlapping). As shown by the locations of other categories, age and education lead to distinct patterns of opinions.

11.5 Close the display window, and press **"Convex hulls"**. The ellipses are now replaced by the convex hulls of the replicates for each point. The convex hulls take into account the peripheral points, whereas the ellipses are drawn using the density of the clouds of replicates. The two pieces of information are complementary. Go back to the main menu.

12) Click on: "ClusterView "

12.1 Choose the axes (1 and 2 to begin with), and: **"Continue"**.

12.2 Click on: **"View"**. The centroids of the 7 clusters (produced by Step PARTI) appear on the first principal plane.

12.3 Activate the button: **"Categorical"**. Pointing with the mouse on a specific category, and pressing the right button of the mouse, we can read the most characteristic categories of the selected category. This description is somewhat redundant with that provided in the results file (file "imp.txt) by the step DECLA. But we do have simultaneously in front of us the pattern of categories and their relative locations.

12.4 Activate the button **"Words"**, and , pointing with the mouse on a specific cluster, press the right button of the mouse. A description of the cluster involving the most characteristic words of the cluster appears. This description is somewhat redundant with that of the Step MOCAR. But, again, we do have in front of us the pattern of clusters and their relative locations.

12.5 Activate the button: **"Texts"**. Pointing with the mouse on a specific cluster, and pressing the right button of the mouse, we can read the most characteristic responses of the selected cluster.
Return.

13) Click on "Kohonen map"
Select the type of coordinate.

13.1 Select: **"Columns (variables)"**: these active variables are the categories in this example.

13.2 Select a (4 x 4) map, and continue.

13.3 After clicking on some check-boxes, press: **"Draw"** on the menu of the large green windows entitled Kohonen map.

13.4 You can change the font size: (**"Font"**) and dilate the obtained Kohonen map: (**"Dilat."**) to make it more legible. The categories appearing in the same cell are often associated in the same responses. This property holds, at a lesser degree, for contiguous cells.

13.5 Pressing **"AxeView"**, and selecting one axis allows one to enrich the display with pieces of information about a specified principal axis: large positive coordinates in red colour, large negative coordinates in green, with some transitional hues.

13.6 Go back to the main menu, click on: **"Kohonen map"** and choose the item **"Active observations".**

13.7 Select a (12 x 12) map, and redo the previous operations for the observations (the button **"Dilat."** is now indispensable).

Appendix B.2 *(for advanced users)*

A similar (but not identical) command file can be generated using the menu "Create a command file". Therefore, beginners could skip this appendix

The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: "Help about parameters").

Command file: EX_B02_Param.txt
Now, we will exhibit the command file that contains comments (preceded by #).

```
# ---------------- EX_B02_Param.txt : Textual Data Analysis  ----
# The Program DtmVic needs 4 files in this "open survey case"
# ------------------------------------------------------------
# 1) The present file of commands, whatever its name.
# 2) The text file (NTEXZ).
# 3) The dictionary file (NDICZ).
# 4) The data file (NDONZ).
#   Syntax:  ">"= continuation, "#"= comments
#-------------------------------------------------------------

LISTP = yes, LISTF = no  # leave as it is...
```

```
NTEXZ = 'TDA_tex.txt'    # text file (same as in example TDA1)
NDICZ = 'TDA_dic.txt'    # dictionary file
NDONZ = 'TDA_dat.txt'    # data file

STEP ARDAT # Archiving data and dictionary
==========
 NQEXA =14 , NIDI = 1,  NIEXA =1043
#-------------------------------------------------------------
#  NQEXA:  number of variables in the  dictionary.
#  NIDI:  number of groups of 4 characters
#             (identifier of individuals) (0=no).
#  NIEXA:  number of individuals in file ndonz.
#-------------------------------------------------

STEP ARTEX  # Archiving responses to 3 open questions
===========
ityp = 2 nbqt = 3 nlig=5

# See Appendix B1 above for the comments about this step
# or the "Help about Command Parameters" (Main menu and Editor "Open an
existing command file").
#-------------------------------------------------


STEP SELOX  # Selecting responses  to questions 1 and 2
===========
NUMQ=LIST     LDONA=1
1,2

# See Appendix B1 for the comments about this step
#-------------------------------------------------


STEP NUMER # extracting words : threshold= 0
===========
NSEU = 0, LEDIT = TOT NXMAX = 20000 coef = 10
weak  -
strong . ? ; ( ) : , '
end

# See Appendix B1 for the comments about this step
#-------------------------------------------------

#---------------- example of pre-processing texts --

NSPC = 'NSPC'

# the file NSPC created by CORTE is given the name: 'NSPC'

step CORTE
========== deletion and equivalence between words
LEDIT = 2
delet  a an and at but by etc for from if in into of on or >
   out over pp than the to up
equiv  two 2
equiv  be am m are re is been being was
equiv  child children
equiv  content contented
equiv  can could
```

```
equiv  would d
equiv  do doing don
equiv  enjoy enjoying
equiv  family families
equiv  get got getting
equiv  go going
equiv  have having ve
equiv  help helping
equiv  holiday holidays
equiv  job jobs
equiv  keep keeping
equiv  live living
equiv  look looking
equiv  see seeing
equiv  son sons
equiv  sport sports
equiv  thing things
equiv  work working
equiv  worry worries
end

#---------------------- Comments about step CORTE
# step CORTE (correction of texts) helps us to perform
# what we may term a manual lemmatisation.
# In fact, the frequency threshold NSEU should be "0"
# in the preceding step NUMER…
# The deletions concerns mainly function words (or tool
# words, or auxiliary words, or grammatical words…).
# Many equivalences are found simply by looking at the
# alphabetical list of words provided by step NUMER.
#  ledit: printing of words  (0=no, 1=nspc, 2=tot).
#  lclas: printing sorted words (0=no, 1=yes).
#
# CORTE uses 3 key-words whose meanings are straightforward:
# delet,  equiv,  end
#-------------------------------------------------------
# IMPORTANT NOTE
# The previous series of deletions and equivalences can be
# generated via the step CORTEX:
# Click on the button "Create a command file" of the main
# menu (Basic Steps) and follow the proposed instructions
# (button: CORTEX, in the paragraph "Textual data").
#-------------------------------------------------------

NSPA = 'NSPC'

#----- the file 'NSPC' created by CORTE is substituted to
# the file NSPA that was created by NUMER.

#---- selecting a new threshold for words -

NSPB ='NSPB'

# the file NSPB created by SETEX is given the name: 'NSPB'

STEP SETEX
============================
NSEU =15 NMOMI=0 NREMI=2 LEDIT =NEW
```

```
#---------------------- Comments about step SETEX
# NSEU:  threshold of frequency for selecting words.
# nmomi: minimum number of letters of a kept word.
# nremi: minimum number of words of a kept response.
# ledit: printing the dictionaries (0=no, 1=new, 2=tot).
#-----------------------------------------------------------

NSPA = 'NSPB'

#----- the file 'NSPB' created by SETEX  is substituted to
# the file NSPA that was created by NUMER and modified by CORTE.

STEP SELEC
========== Selects active, supplementary variables and observations
LSELI = TOT,  IMASS = UNIF, LZERO = REC, LEDIT = short
NOMI ILL 1 2 11 14
NOMI ACT 4--10
END

#---------------------------------------------------
# LSELI:    mode of selection of individuals
#         (0=all, 1=list, 2=filt).
# IMASS:    index of variable 'weight of individuals'
#           (0=uniform weights).
# LZERO:    coding missing responses (0=norec, 1=rec).
# LEDIT:    printing dictionary of selected variables
#           (0=no, 1=short,2=long).
#   The selection commands use following key words :
# ACT (active)  ILL (illustrative, or supplementary)
# NOMI (nominal, or categorical), END (end of selection)
#---------------------------------------------------

STEP MULTM
========== Multiple correspondence analysis
NAXE  = 7, PCMIN = 2. , LBURT = TOT, LEDCO = yes NSIMU=10

#---------------------- Comments about step MULTM
# - NAXE = ... number of computed principal axes
# - PCMIN ... threshold for "cleaning" the active
# categories (in percent). This means that the low-
# frequency active categories (less than 2% in this
# case) are eliminated, and the corresponding
# individuals are dispatched at random among the
# other categories of the same variable (to remedy
# a well known weakness of the chi-square distance).
#
# - LBURT... printing the Burt contingency table
#     (0=NO, 1=MASS, 2=TOT, 3=PROF).
# - LEDCO... printing the correlations variable-
#    axes (0=no, 1=yes).
# - NSIMU...number of bootstrap replication (less than 30)
#    (0 = no bootstrap)
#---------------------------------------------------

STEP DEFAC   # Description of factorial axes
========== Multiple correspondence analysis
SEUIL = 40., LCRIM = VTEST, VTMIN = 2.0
VEC = 1--2 / MOD
end
```

```
#---------------------- Comments about step DEFAC
# SEUIL = ... Maximum number of elements that will
#  be sorted to describe each axis
# LCRIM = ... Criterion for sorting the elements
#  (here VTEST means "test-values" (signed number
#  of standard deviations)
# VEC = ... list of axes to be described
# CONT = continuous variables , MOD = categories
# The key-word END indicates the end of the list.
#-------------------------------------------------

STEP POLEX
==== projecting supplementary words
ngraf = 2

#---------------------- Comments about step POLEX
# POLEX aims at positioning words on principal space
# (here: principal space provided by MCA of closed questions)
#  ngraf = number of requested graphics (on file imp.txt)
#-------------------------------------------------

STEP RECIP
==== Clustering of respondents using reciprocal neighbours
NAXU=7 LDEND=DENSE NTERM=20 LDESC=no

# See Appendix B1 for the comments about this step
#-------------------------------------------------


STEP PARTI
==== Cut of the dendrogram to obtain 7 clusters
NITER=10 LEDIN=3
7  # number of classes of the partition

# See Appendix B1 for the comments about this step
#-------------------------------------------------

STEP DECLA
========== Systematic description of clusters
CMODA = 5.0, PCMIN = 2.0, LSUPR = no, CCONT = 5.0 >
LPNOM = no, EDNOM = no, EDCON = no
7 # list of numbers of classes of requested partitions
#-------------------------------------------------
# EDNOM:  printing the tables (classes * questions)
#         (0=no).
# LPNOM:  describing partition with questions
#         (0=no, 1=yes).
# CMODA:  describing classes with categories (0=no).
# PCMIN:  minimum relative ( % ) weight for a category.
# LSUPR:  characteristic category if
#         %(cat./class) > %(cat./total)(0=no,1=yes).
# EDCON:  describing partition with numerical variables
#        (0=no, 1=yes).
# CCONT:  describing classes with numerical variables
#         (0=no).
#-------------------------------------------------

STEP MOTEX
```

```
===========  Cross-tabulating words and partition
NVSEL = -1, LEDIT = 1

#---------------------- Comments about step MOTEX
# See Appendix B1 for the comments about this step
#-------------------------------------------------

STEP MOCAR
==== Characteristic words for each cluster (criterion 1)
NOMOT=10   NOREP=6

# See Appendix B1 for the comments about this step
#-------------------------------------------------

STEP RECAR
===========  characteristic responses (criterion 2)
NOREP = 4

#---------------------- Comments about step RECAR
#  NOREP: number of characteristic responses for each text.
# RECAR, for each cluster or category, computes the Chi-square
# distances between the responses and the mean-point of the category.
# Responses having the shortest distances are considered as
# characteristic of the category. This criterion is favourable
# to lengthy responses.
#-------------------------------------------------
STOP
#-------------------------------------------------
```

End of example B.2

IV.3 Analysis of a Semantic network (French verbs)

Example B.3. EX_B03.Text-Semantic.

Example B.3 provides a visualisation of the semantic links existing between 829 French verbs. Each verb is described by a list of synonyms. This example is in fact very similar to Example B.1 (Responses to an open question). The "respondents" are here the 829 verbs. The fictitious open-ended question is "Which are your synonyms?", and the textual "response" is constituted by a list of synonyms. The example is also similar to the "Japan Map" example, pertaining to Example C.3 (Descriptions of graphs) from chapter V.
The principal axes visualization is complemented by a clustering, with an automatic description of the clusters. This is a typical first outlook on the set of responses: to detect and describe the main groupings of responses. Such outlook is by no means an achieved processing.

For a similar application, please refer to the book: "The Semiometric Challenge" (2014) by L. Lebart, J.F. Steiner; M. Piron, J. Wisdom. Publisher: L2C, (The book can be freely downloaded from www.dtmvic.com).

To have a look at the data, search for the directory **DtmVic_Examples.**
In this directory, open the sub-directory **DtmVic_Examples_B_Texts** .
In that directory, open the directory of Example B.03, named " **EX_B03.Text-Semantic".**

It is recommended to use one directory for each application, since DtmVic produces a lot of intermediate txt-files related to the application. At the outset, such directory must contain 2 files:

- a) the text file, **synotex.txt**

- b) the command file: **"syno_par.txt"**

(in this particular context, there are neither data file nor dictionary file: the fictitious questionnaire comprises one open-ended question, without closed-end questions).

a) Text file: synotex.txt

The format is typical of responses to open questions (see examples A.5, B.1, B.2). Since the “responses” (here: lists of synonym verbs) may have different lengths, separators are used to distinguish between these lists. Lists (in fact : responses) are separated by the chain of characters “----“ (starting column 1) possibly followed by an identifier. Like all the data files involved in DtmVic as input files, that file is a raw text file (.txt). If the text file comes from a text processing phase, it must be saved beforehand as a “.txt file”.

b)Command file: EX_B03_Param.txt
The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: **"Help about command parameters"**) and, with more details, below.

Note that this “command file” **“EX_B03_Param.txt”** can be also generated by clicking on the button **“Create a command file”** of the main menu (DTM: Basic Steps). A window **“Choosing among some basic analysis”** appears. Click then on the button: **VISURESP** – Visualization of Responses – located in the paragraph **“Textual data”** , and follow the instructions.

Running the example B.3 and reading the results

1) Click on the button: “Open an existing command file” (Main menu)

2) Then, search for the sub-directory DtmVic_Examples_B_Texts in: **DtmVic_Examples.**

3) In that directory, open the directory of Example B.03, named “EX_B03.Text-Semantic” .

4) Open then the command file: EX_B03_Param.txt
After identifying the textual data file, seven "steps" are performed:

ARTEX (Archiving texts),
SELOX (selecting the open question),
NUMER (numerical coding of the text),
ASPAR (correspondence analysis of the [sparse] contingency table “respondents - words”),
CLAIR (Brief description of factorial axes),
RECIP (Clustering using a hierarchical classification of the clusters - reciprocal

neighbours method),
PARTI (Cut of the dendrogram produced by the previous step, and optimisation of the partition obtained),
MOTEX (cross-tabulating the partition produced by step PARTI with words: the obtained contingency table is called a lexical table),
MOCAR (characteristic words, and characteristics responses for each class of the partition).

Instead of editing this file, we go back directly to the main menu and execute the basic computation steps.
Return to the main menu **("Return to execute")**

5) Click on the button: "Execute a command file"
This step will run the basic computation steps present in the command file.

6) Click the button: "Basic numerical results"
The button opens a created (and saved) html file named **"imp.html"** which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu. (Note that this file is also saved under another name. See previous examples)
From the step NUMER, we learn for instance that we have 829 "responses", with a total number of words (occurrences or token) of 17 446, involving 3 839 distinct words (or: types). Using a frequency threshold of 12, the total number of kept words reduces to 5 013, whereas the number of distinct kept word reduces (more drastically) to 280. **Return.**

7) At this stage, we click on one of the lower buttons of the basic steps panel (Steps: "VIC")
8) Click the button: "ViewAxes"
and ... follow the sub-menus. In fact, only two tabs are relevant for this example: **"Active variables"** [= synonyms in the case of step ASPAR], **"Individuals (observations)"** [= 829 initial words]. After clicking on **"View"** in both cases, one obtains the set of principal coordinates along each axis.
Clicking on a column header produce a ranking of all the rows according to the values of that column. In this particular example, this is somewhat redundant with the printed results of the step "CLAIR" .

9) Click the button: "PlaneView Research" ... and follow the sub-menus.
In this example, four items of the menu are relevant **"Active columns (variables or categories)"** (= synonyms, here), **"Active Rows"** (individuals, observations)"

(= 829 original words), **"Active columns + Rows"** , **"Individuals (density)"**. The graphical display of chosen pairs of axes are then produced.
The roles of the different buttons are straightforward, except perhaps the button: **"Rank"**, which is useful only in the case of very intricate displays, (which is the case here). Since the set "individual" has 829 elements, it is possible to test, with this example, partial printings of the individuals in two subsets of 50% or four subsets of 25%…(subsets randomly drawn without replacement)

10) Click the button: "BootstrapView"
The use of the bootstrap is similar to Example B1 above..

11) Click on "ClusterView "

11.1 Choose the axes (1 and 2 to begin with), and **"Continue"**.

11.2 Click on **"View"**. The centroids of the 20 clusters (Step PARTI) appears on the first principal plane.

11.3 Activate the button **"Words"**, and, pointing with the mouse on a specific cluster, press the right button of the mouse. A description of the cluster involving the most characteristic words of the cluster appears. This description is somewhat redundant with that of the Step MOCAR . But we do have in front of us the pattern of clusters and their relative locations.

11.4 Activate the button **"Texts"**. Pointing with the mouse on a specific cluster, and pressing the right button of the mouse, we can read the most characteristic responses of the selected cluster.

12) Click on: "Kohonen map"
Select the type of coordinate.

12.1 Select: **"Active variables (columns)"** : these active variables are the words in this example.

12.2 Select a (8 x 8) map, and continue.

12.3 After clicking on two small check-boxes, press **"Draw"** on the menu of the large green windows entitled Kohonen map.

12.4 You can change the font size (**"Font"**) and dilate the obtained Kohonen map (**"Dilat."**) to make it more legible. The words appearing in the same cell are often associated in the same verbs. This property holds, at a lesser degree, for contiguous cells.

12.5 Pressing **"AxeView"**, and selecting one axis allows one to enrich the display with pieces of information about a specified principal axis : large positive coordinates in red colour, large negative coordinates in green, with some transitional hues.

12.6 Go back to the main menu, click on **"Kohonen map"** and choose the item **"Active observations".**

12.7 Select a (10 x 10) map, and redo the operations 12.3 to 12.5 for the observations.
In the context of this example, the other items of the main menu are not relevant.

13 Click on "Visualization"
A new window is displayed.

13.1 Click on: **"Load coordinate"**
In the corresponding sub-menu, choose the file: **"ngus_ind.txt"** . The principal coordinates of the individuals (rows) are selected.

13.2 Click then on **"Select or Create Partition"**
In the corresponding sub-menu, choose **"no partition"** .

13.3 Click on: **"MST"** (Minimum Spanning Tree). Choose then the number of axes that will serve to compute the Minimum Spanning Tree: full space (for example).

13.4 Click on: **" N.N."** (search for Nearest Neighbours – limited to 20 NN).
13.5 Click on: **"Graphics"** .
Choose the axes 1 and 2 (default) in the small window "Description of classes" and click on: **"Display"** .
In the new window entitled **"Visualisation-Graphics"** are displayed the individuals in the plane spanned by the selected axes. A random colour is attributed to each cluster (if any). The button **"Change colour"** allows you to try a new set of colour.

About the window "Visualisation - Graphics" (from the sub-menu Graphics)

On the vertical tool bar, you can press each button to activate it (red colour), and press it again to cancel the activation (original colour)

- The button **"Density"** , for sake of legibility, replaces the identifiers of individuals by a single character reminding the cluster (the identifier and the cluster number can be obtained by clicking on the left button of the mouse in the vicinity of each point).
- The button **"C.Hull"** (Convex hull) draws the convex hull of each cluster.
- The button **"MST"** (Minimum Spanning Tree) draws the minimum spanning tree.
- The button **"Ellipse"** perform a Principal Components Analysis of each cluster within the two-dimensional sub-space of visualisation and draws the corresponding ellipses (containing roughly 95% of the points).
- The button **"N.N."** (Nearest neighbours) joins each point to its nearest neighbours. Pressing afterwards the button **"N.N. up"** allows you to increment the number of neighbours up to 20 nearest neighbours.

Appendix B.3

The steps and the command file of example B.3 are included in those of Example B.1 (if we except the name of the data file containing the input text).
The reader should then refer to Appendix B.1 to obtain the corresponding comments. Remind that this command file can be also generated by clicking on the button **"Create a command file"** of the main menu (DTM: Basic Steps), and selecting the procedure: **VISURESP** – Visualization of Responses – in the paragraph **"Textual data".**

End of example B3

Chapter V

Numerical data: More examples

Each example corresponds to a directory included in
"DtmVic_Examples_NumDat"

Application examples C.1—C.4

V.1 Numerical data, PCA, Semiometry

Example C.1. EX_C01.PCA_Semio

Example C.1 aims at describing a set of numerical variables (an excerpt of semiometric data) through Principal Components Analysis complemented with a clustering. Bootstrap procedures, Kohonen maps are followed by the various tools of visualisation provided in the sub-menu "Visualization" of the phase "VIC": visualisation of clusters (or categories) using symbols or colours, convex hulls or density ellipses for clusters, Minimum spanning tree, drawing of various nearest neighbours graphs.

V.2 PCA, Contiguity, Discrimination (Fisher's Iris Data)

Example C.2. EX_C02.PCA_Contiguity

Example C.2 is devoted to a classical set of numerical variables (The Iris data set of Anderson and Fisher) through PCA, Clustering, Contiguity Analysis, Discriminant Analysis. The principal axes visualisation is complemented by a clustering. At the outset, example C.2 is very similar to example C.1: PCA and classification (clustering) of a set of numerical data, with various tools of visualisation, involving also a specific categorical data. It presents then the improvements provided by Contiguity Analysis and its particular case: Linear Discriminant Analysis.

V.3 Description of graphs through CA

Example C.3. EX_C03.Graphs

Example C.3 aims at describing some simple symmetrical planar graphs, mainly through correspondence analysis. The directory EX_03.Graphs contains several sub-directories and examples. The 3 graphs are planar graphs: a chessboard shaped graph, a cycle, and graphs supposed to represent maps of the regions of Japan and France. The examples provide a bridge between distinct facets of DtmVic: a same graph can lead to different input data.

V.4 Structural Compression of Images

Example C.4. EX_C04.Images

Example C.4 deals with the *Structural Compression of Images* through CA, Singular Values Decomposition, and Discrete Fourier Transforms. It could be viewed as a pedagogical appendix. It does not make use of data in DtmVic format, since it deals with digitalized images. A simple rectangular array of integers suffices: there is no need for identifiers of rows or column. A specialized interface is provided via the button "DtmVic Images" of the main menu.

V.1 Numerical data, PCA, Semiometry

Example C.1. EX_C01.PCA_Semio

Example C.1 aims at describing a set of numerical variables (an excerpt of "semiometric data") through Principal Components Analysis. The principal axes visualisation is complemented by a clustering, with an automatic description of the clusters. Bootstrap procedures, Kohonen maps are followed by the various tools of visualisation provided in the menu "Visualization" in the sub-window "Visualization, Inference, Classification": visualisation of clusters (or categories) using symbols or colours, convex hulls or density ellipses for clusters, Minimum spanning tree, drawing of various nearest neighbours graphs.

A new clustering of variables (or of observations/individuals) through a simple k-means method can be obtained and visualized from the sub-menu "Visualization".

About Semiometric data:

In most surveys in the field of marketing research, it is customary to include information about lifestyles and values. Such information is generally obtained through a set of questions describing the attitudes and opinions towards a list of sentences or statements. "Semiometry" is a technique introduced by Jean-François Steiner, a writer interested in marketing research, to tackle that problem in a more general way.

The basic idea is to insert in the questionnaire a series of questions consisting uniquely of words (a list of 210 words is currently used, but we will be dealing here with an abbreviated lists containing a subset of 70 words). The interviewees must rate these words according to a seven levels scale, the lowest level (mark = 1) relating to a "most disagreeable (or unpleasant) feeling about the word", the highest level (mark = 7) relating to a "most agreeable (or pleasant) feeling" about the word.

The processing of the filled questionnaires (mainly through Principal Component Analysis) produces a stable pattern (up to 8 stable principal axes). Very similar patterns are obtained in ten different countries, despite the problems posed by the translation of the list of words.

For more information, please refer to the book: "*The Semiometric Challenge*" (2014) by L. Lebart, J.F. Steiner; M. Piron, J. Wisdom. Publisher: L2C, (The book can be freely downloaded from www.dtmvic.com).

Semiometrics data files

To have a look at the data, search for the directory **DtmVic_Examples.**
In this directory, open the sub-directory **DtmVic_Examples_C_NumData** .
In that directory, open the directory of Example C.1, named **"EX_C01_PCA.Semio"** .
It is recommended to use one directory for each application, since DtmVic produces a lot of intermediate txt-files related to the application. At the outset, such directory must contain 3 files :
- a) the data file,
- b) the dictionary file,
- c) the command file.

a) Data file: "PCA_semio.dat.txt"

Our reduced-size example comprises 300 respondents (instead of 1000 or 2000 that are the usual sizes of semiometric survey samples) and 76 variables: 70 words (the marks given to the words are considered here as numerical variables) and 6 categorical variables describing the characteristics of the respondents. The data file "PCA.dat.txt" comprises 300 rows and 76 columns (identifier of rows [between quotes] + 75 values [corresponding either to numerical variables or to item numbers of categorical variables] separated by at least one blank space).

b) Dictionary file: "PCA_semio.dic.txt"

The dictionary file "PCA.dic.txt" contains the identifiers of these 75 variables. In this version of DtmVic, the identifiers of categories must begin at: "column 6" [a fixed interval font - also known as teletype font - such as "courier" can be used to facilitate this kind of format].

c) Command file: "EX_C01_Param.txt"

The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: **"Help about command parameters"**) or in the editor (button **"Open an existing command file"),** of the main menu.

Note that another "command file" similar (but not identical) to the "command file: **"EX_C01_Param.txt"** can be also generated by clicking on the button **"Create a command file"** of the main menu (DTM: Basic Steps). Proceed then as shown by the first example **"EX_A01.PrinCompAnalysis"** of chapter 2.

Running the example C.1 and reading the results

1) Click on the button : "Open an existing command file" (line: *Command file* of the main menu)
2) Search for the sub-directory " DtmVic_Examples_C_NumData" in **"DtmVic_Examples".**
3) In that directory, open the directory of Example C.1: " EX_C01.PCA_Semio"
4) Open the command file: "EX_C01_Param.txt"

After identifying the two data files, 10 "steps" are identified:
ARDAT (Archiving data),
SELEC (selecting active and supplementary elements),
STATS (some basic statistics),
PRICO (Principal components analysis),
DEFAC (Brief description of factorial axes),
RECIP (Clustering using a hierarchical classification of the clusters - reciprocal neighbours method),
PARTI (Cut of the dendrogram produced by the previous step, and optimisation of the obtained partition),
DECLA (Automatic description of the classes of the partition),
SELEC (selecting one categorical variable, in this case),
EXCAT (extracting one categorical variable - selected by step SELEC - to be used in some graphical displays).

In this example, as in most applications, the step SELEC plays a fundamental role, in deciding which set of variables will be active, and which set will be illustrative or supplementary.
In that command file, the step RECIP performs a hierarchical clustering of the elements using the "reciprocal neighbour algorithm" and the step PARTI that follows cuts the obtained tree according to the *a priori* fixed number of clusters. PARTI optimizes afterwards the corresponding partition through k-means iterations.

The methodology of this "hybrid algorithm" is presented in "Multivariate Descriptive Statistical Analysis" (L. Lebart, A. Morineau, K. Warwick; J. Wiley, New York, 1984).

We will comment later on this command file (Appendix of the section) which commands the basic computation steps. Instead of editing this file, we will go back to the main menu and execute the basic computation steps.

5) Return to the main menu (“Return to execute”)

6) Click on the button: “Execute a command file”
This step will run the basic computation steps present in the command file.

7) Click the button: “Basic numerical results”
The button opens a created (and saved) html file named **“imp.html”** which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu. Note that this file is also saved under another name.
[The name **“imp.html”** is concatenated with the date and time of the analysis (continental notation). That file keeps as an archive the main numerical results whereas the file **“imp.html”** is replaced for each new analysis performed in the same directory. This file in also saved under a simple text format , under the name **“imp.txt”** , and likewise with a name including the date and time of execution.]

8) At this stage, we click on one of the lower buttons of the basic steps panel (Steps: “VIC”)

9) Click the button “ViewAxes” ... and follow the sub-menus. In fact, only two tabs are relevant for this example: **“Active variables”** [= the selected words] , **“Individuals (observations) [= respondents]”** and **"supplementary categories"** . After clicking on **“View”** in each case, the set of principal coordinates along each axis is displayed.
Clicking on a column header produces a ranking of all the rows according to the values of that column. In this particular example, this is redundant with the printed results of the step “DEFAC”.
In the case of this particular example, in which the first axes appear to be stable and to have an interpretation, the ViewAxes procedure is useful to observe at a glance the bundles of words occupying extreme locations along each axis. Example for active variables and axis 2: opposition between the words “sacred, God, perfection, soul” on the one hand, and “sensual, adventurer, nudity, island, desire” on the other. The supplementary categories characterize the respondents. The bootstrap, later on, will provide a validation of some aspects of the observed structure.

10) Click the button: PlaneView Research... and follow the sub-menus.
In this example, four items of the menu are relevant **"Active columns (variables or categories)"** and **"supplementary categories"**, **"Active rows (individuals, observations)"**, **"Active columns + Active rows"**, **"Active individuals (density)"** and **"Active columns + Supplementary categories"**. The graphical displays of chosen pairs of axes are then produced.
In the case of semiometric data, the so-called "first semiometric plane" is in fact the plane spanned by the axes 2 and 3. The first axis is referred to as a "purely methodological axis", linked to a "size effect" common in many PCA applications (a whole chapter of the [downloadable] book quoted previously: "The Semiometric Challenge" is devoted to this first axis).
In the case of PCA, the first menu item **"Active columns (variables or categories)"** may contain, in fact, both active numerical variables (in black) and supplementary numerical variables (in red). We have only active numerical variables in this particular example, but, later on, the reader can edit the command file (step SELEC) to withdraw some words from the active set and give them the status of "supplementary (or illustrative) elements". He or she can also use the **"Create a command file"** menu, exemplified in chapter 2, example A.1, to choose the procedure "PCA", allowing then for selecting more comfortably the active and supplementary elements.

Go back to the "VIC" set of buttons.

11) Click the button: "BootstrapView"
This button open the DtmVic-Bootstrap-Stability windows.

11.1 Click **"LoadData"** . In this case (partial bootstrap), the two replicated coordinates file to be opened are named **"ngus_var_boot.txt"** and **"ngus_sup_cat_boot.txt"** (see the panel reminding the names of the relevant files below the menu bar).
The file **"ngus_var_boot.txt"** contains only active variables. The file **"ngus_sup_cat_boot.txt"** contains only supplementary categories, for which the bootstrap procedure is also meaningful.

11.2 Click on **"Confidence areas"** submenu, and choose the pair of axes to be displayed (choose axes 2 and 3, to begin with).

11.3 Click on **"Loading"** in the blue window that appears then, to obtain the dictionaries of variables. Tick the chosen white boxes to select the elements the location of which should be assessed, and press the button **"Select"**. Select for instance among others, the categories Male and Female

11.4 Click on **"Confidence Ellipses"** to obtain the graphical display of the active variable points (if the file **ngus_var_boot.txt** has been loaded), or of the supplementary category points (if the file **ngus_sup_cat_boot.txt** has been loaded).

11.5 Close the display window, and press **"Convex hulls"**. The ellipses are now replaced by the convex hulls of the replicates for each point. The convex hulls take into account the peripheral points, whereas the ellipses are drawn using the density of the clouds of replicates. The two pieces of information are complementary.

Go back to the "VIC" set of buttons.

12. Click on "ClusterView "

12.1 Choose the axes (2 and 3 to begin with), and **"Continue"**.

12.2 Click on **"View"** . The centroids of the 7 clusters of individuals (Step PARTI) appear on the first principal plane.

12.3 Activate the button **"Categorical"** , and, pointing with the mouse on a specific cluster, press the right button of the mouse. A description of the cluster involving the most characteristic response items appears. This description is similar to that of the Step DECLA . But we can watch on this display the pattern of clusters and their relative locations. One can easily imagine the usefulness of the tool for a survey with thousands of individuals, hundreds of variables, and more clusters.

12.4 Activate the button **"Numerical"**. We will observe the link between the numerical variables (both active and supplementary variables) of the data file and the 5 clusters. Due to the small number of individuals, some clusters do not produce significant results.
Go back to the "VIC" menu.

13) Click on "Kohonen map"
Select the type of coordinate.

13.1 Select: **"Active variables (columns)"**: these active variables are the 70 words in this example.

13.2 Select a (4 x 4) map, and continue.

13.3 After clicking on two small check-boxes, press **"Draw"** on the menu of the large green windows entitled Kohonen map.

13.4 You can change the font size (**"Font"**) and dilate the obtained Kohonen map (**"Dilat."**) to make it more legible. The words appearing in the same cell are often associated in the same responses. This property holds, at a lesser degree, for contiguous cells.

13.5 Pressing **"AxeView"**, and selecting one axis allows one to enrich the display with pieces of information about a specified principal axis : large positive coordinates in red colour, large negative coordinates in green, with some transitional hues.

13.6 Go back to the main menu, click on **"Kohonen map"** and choose the item **"Observations"**

13.7 Select a (8 x 8) map, and redo the operations 13.3 to 13.5 for the observations.
Go back to the "VIC" set of buttons.

14. Click on "Visualization"
A new window entitled **"DTM-Visualization: Loading files, Selecting axes"** appears.

14.1 Click on **"Load coordinate"**
In the corresponding sub-menu, choose the file: **"ngus_ind.txt"**. The principal coordinates of the individuals (rows) are selected.

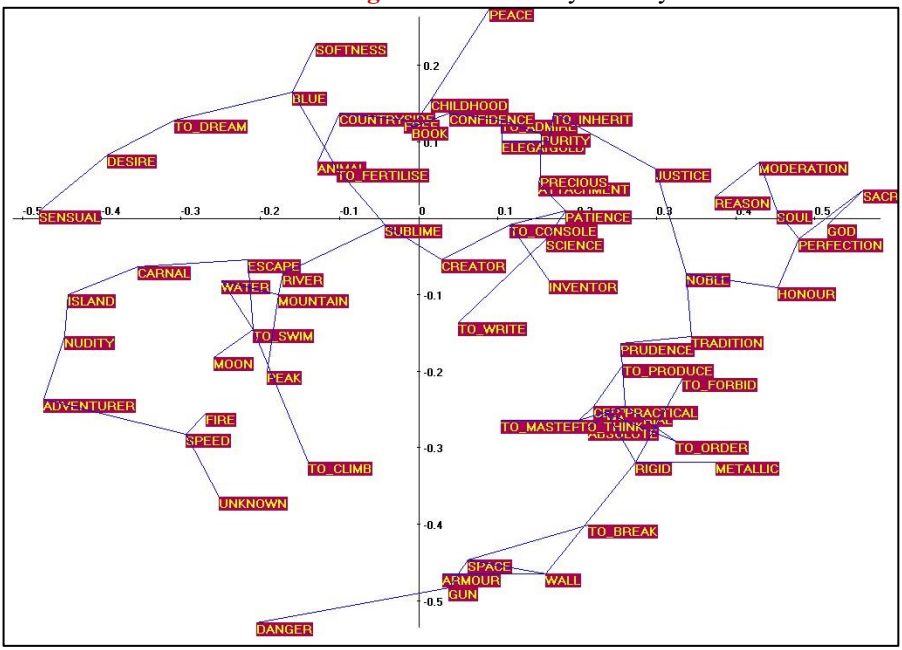
14.2 Click then on **"Load a partition file"**
In the corresponding sub-menu, choose **"Select a partition"** . The partition obtained previously from the computation step must then be loaded (its name: **"part_cla_ind.txt"**).

14.3 Click on **"MST"** (Minimum Spanning Tree). Choose then the number of axes that will serve to compute the Minimum Spanning Tree: 5 (for example).

14.4 Click on **"N.N."** (search for nearest neighbours – limited to 20 NN).

14.5 Click on **"Graphics"**.
Choose the axes 1 and 2 (default) in the small window **"Selection of Axes"** and click on **"Display"**.

In the new window entitled **"Visualisation, Graphics"** are displayed the individuals in the plane spanned by the selected axes. A random colour is attributed to each cluster. The button **"Change colour"** allows you to try a new set of colour.



Plane (2, 3) of the semiometric PCA, with the Minimum Spanning Tree

14.6 About the window **"Visualisation, Graphics"**

14.6.1 -- On the vertical tool bar, you can press each button to activate it (red colour), and press it again to cancel the activation (initial colour)

14.6.2 -- The button **"Density"**, for sake of clarity, replace the identifiers of individuals by a single character reminding the cluster (the identifier and the cluster number can be obtained by clicking on the left button of the mouse in the vicinity of each point).

14.6.3 -- The button **"C.Hull"** (Convex hull) draws the convex hull of each cluster.

14.6.4 -- The button **"MST"** (Minimum Spanning Tree) draws the minimum spanning tree.

14.6.5 -- The button **"Ellipse"** perform a Principal Components Analysis into each cluster within the two-dimensional sub-space of visualisation and draws the corresponding ellipses (containing roughly 90% of the points).

14.6.6 -- The button **"N.N."** (Nearest neighbours) joins each point to its nearest neighbours. Pressing afterwards the button **"N.N. up"** allows you to increment the number of neighbours up to 20 nearest neighbours.
14.6.7 -- We will see in section 16 below how to use the lower buttons of the left side vertical bar "IterKM", "Mean", "Clust" (useful to visualize the iteration of a k-means partition).

Go back to the "VIC" menu.

15. Click again on "Visualization"

15.1 We are going to redo the operation of paragraph 14, but instead of loading a partition provided by a clustering algorithm, we will load the partition induced by the categories of a specific categorical variable. Such partition correspond to the variable number 76 (gender), selected and extracted through the steps SELEC and EXCAT (at the end of the command file, see below).
15.2 In the window entitled **"DTM-Visualization: loading files, selecting axes "** , click on **"Load coordinate"**

15.3 In the corresponding sub-menu, choose again the file: **"ngus_ind.txt"** . The principal coordinates of the individuals (rows) are selected.

15.4 Click then on **"Load or create Partition"**

15.5 In the sub-menu **"Load or create Partition"** choose the file **"part_cat.txt"** . The partition induced by the categories of variable number 76 (gender) is loaded. After loading that partition, all the operations from 14.3 to 14.6 can be carried out again.

Comment: It is interesting to visualise the individuals in the plane spanned by the axes 2 and 3.

The two categories Male and Female are significantly linked to axis 3 (as it can be highlighted by looking at the bootstrap confidence areas). But this link is hardly visible when we look directly at the convex hulls of the two sub-clouds corresponding to these two categories of respondents. This (almost) paradoxical result exemplifies the difference between "statistically significant" (which is the case here) and "obviously different" (which is not the case).

16) Direct computation of a partition within the menu "Visualization"

DtmVic makes it possible now to build on line (i.e. outside the “command file”) a “k-means partition” of variables (or of individuals as well).
Click on “Visualization”

16.1 A new window entitled **“Visualization, Loading files, Selecting axes”** appears.

16.2 Click on **“Load coordinate”**

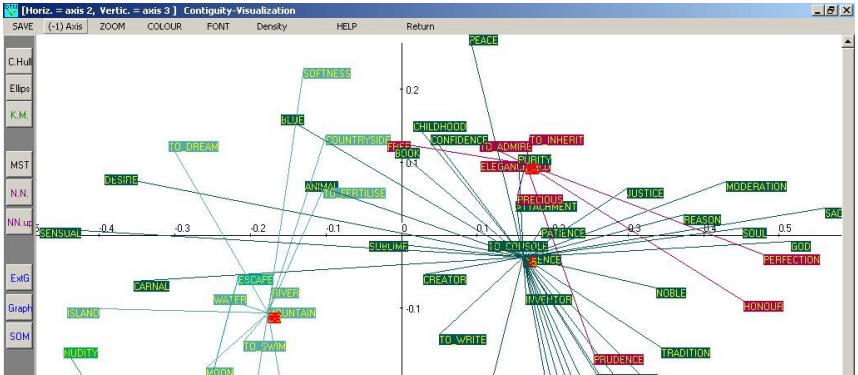
16.3 In the corresponding sub-menu, if you want a clustering of variables, choose the file: **“ngus_var_act.txt”** . The principal coordinates of the **active variables** are selected. If you want a clustering of individuals, select the file: **“ngus_ind.txt”** .

16.4 Click then on **“Load or create Partition”** .In the corresponding sub-menu, select the item **“Create a new k-means partition”** .

16.5 You have then to select the number of desired clusters , the number of principal coordinates to compute the distances, the maximum number of iteration (generally 12 iterations suffice).
You can also choose to visualize the iterations (answer: yes to the question “Do you want to look at the intermediate iterations?”).

16.6 The obtained partition will be automatically loaded and visualized. You can then redo the previous steps 14.3 to 14.6.

16.7 If you want to visualize the different steps, in the window **“Contiguity_Visualisation”** , click on “IterKM” , then click alternately on “Means” (computation of the centroids of the clusters) and on “Clust” (assignment of the elements to the new centroid) until the convergence is reached.



Visualization of the first iteration of the k-means method.

Note that the partition obtained through the classical k-means algorithm generally will not coincide with the partition induced by the parameters of the command file. In that command file, the step RECIP performs a hierarchical clustering of the elements using the "reciprocal neighbour algorithm" and the step PARTI that follow cuts the obtained tree according to the *a priori* fixed number of clusters. PARTI optimizes afterwards the corresponding partition through k-means iterations.

End of example C.1

V.2 PCA, Contiguity, Discrimination

(Fisher's Iris Data)

(Example C.2. EX_C02.PCA_Contiguity)

Example C2 aims at analysing a classical set of numerical variables (*The Iris data set of Anderson and Fisher*) through Principal Components Analysis, Classification, Contiguity Analysis, Discriminant Analysis. As in previous examples, the principal axes visualisation is complemented with a clustering, together with an automatic description of the clusters.
The first phases of Example C.2 are very similar to Example C.1: Principal components analysis and classification (clustering) of a set of numerical data, with various tools of visualisation, involving also a specific categorical data.
Subsection 12 and the following subsections present the improvements provided by Contiguity Analysis, with a comparison with Linear Discriminant Analysis.

Reminder about Contiguity Analysis

In Contiguity analysis, we consider the case of a set of multivariate observations, (*n* objects described by *p* variables, leading to a (*n, p*) matrix **X**), having an *a priori* graph structure. The *n* observations are the vertices of a symmetric graph G, whose associated (*n, n*) matrix is **M** (*mii' = 1* if vertices i and i' are joined by an edge, *mii' = 0* otherwise). Such situation occurs when vertices represent time-points, geographic areas. Contiguity Analysis, confronting local and global variances, provides a straightforward generalization of Linear Discriminant Analysis. It enables to point out the levels responsible of the observed patterns (*local* versus *global* level). In this example, we will deal with the situation in which **M** and the graph structure are not external, but derived from the data matrix **X** itself, **M** being for example the *k- nearest neighbours graph* derived from a distance between observations. Some interesting possibilities of exploration of data are sketched. Note that the idea of deriving a metric likely to highlight the existence of clusters dates back to the works of Art *et al.* (1982) and Gnanadesikan *et al.* (1982).

Some references

Art D., Gnanadesikan R., Kettenring J.R. (1982) Data Based Metrics for Cluster Analysis, *Utilitas Mathematica*, **21** A, 75-99.

Burtschy B., Lebart L. (1991) Contiguity analysis and projection pursuit. In : *Applied Stochastic Models and Data Analysis,* R. Gutierrez and M.J.M. Valderrama, Eds, World Scientific, Singapore, 117-128.
Gnanadesikan R., Kettenring J.R., Landwehr J.M. (1982) Projection Plots for Displaying Clusters, *in Statistics and Probability, Essays in Honor of C.R. Rao,* G. Kallianpur, P.R. Krishnaiah, J.K.Ghosh, eds, North-Holland.
Lebart L. (1969) Analyse statistique de la contiguité. *Publications de l'ISUP.* XVIII, 81-112.
Lebart , L. (2000): Contiguity Analysis and Classification, In: W. Gaul, O. Opitz and M. Schader (Eds): *Data Analysis*. Springer,Berlin, 233--244.
Lebart L. (2006): Assessing Self Organizing Maps via Contiguity Analysis. *Neural Networks* , 19, 847-854.

Looking at the data

To have a look at the data, search for the directory **DtmVic_Examples.**
In this directory, open the sub-directory **DtmVic_Examples_C_NumData**.
In that directory, open the directory of Example C.2, named **"EX_C02.PCA_Contiguity"**
It is recommended to use one directory for each application, since DtmVic produces a lot of intermediate txt-files related to the application. At the outset, such directory must contain 3 files:
- a) the data file,
- b) the dictionary file,
- c) the command file.

a) Data file: "iris_dat.txt"
Our example comprises 150 observations and 4 variables: 4 measurements (these numerical variables are the lengths of various constituents of the flowers: *Sepal Length, Sepal Width, Petal Length, Petal Width*) and one categorical variable describing the characteristics of the observations (three species of plants : setosa, versicolor, virginica).
The data file "iris_dat.txt" comprises 150 rows and 6 columns (the identifier of rows [between quotes] + 5 values [corresponding to 4 numerical variables and one categorical variable] separated by at least one blank space).
[Reference: Anderson, E. (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society* **59**, 2–5.]

b) Dictionary file: "iris_dic.txt"
The dictionary file "iris_dic.txt" contains the identifiers of these 5 variables. In this version of DtmVic dictionary, the identifiers of categories must begin at: "column 6" [a fixed interval font - also known as teletype font - such as "courier" can be used to facilitate this kind of format].

c) Command file: "EX_C02_Param.txt"
The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: **"Help about parameters"**) and below.

Note that another "command file" similar (but not identical) to the "command file" "iris_par.txt" can be also generated by clicking on the button "Create a command file" , line "Command file" of the main menu (DTM: Basic Steps). Proceed than as shown in the first example "EX_A01.PrinCompAnalysis" of chapter 2.

Running the example C.2 and reading the results

1) Click on the button: **"Open an existing command file"** (main menu)

2) Search for the sub-directory "DtmVic_Examples_C_NumData" in **"DtmVic_Examples".**

3) In that directory, open the directory of Example C.2 named "EX_C02.PCA_Contiguity" .

4) Open the command file: "EX_C02_Param.txt"

In that command file, we can read that after identifying the two files (data and dictionary) , 9 "steps" are performed:
ARDAT (Archiving data),
SELEC (selecting active and supplementary elements),
PRICO (Principal components analysis),
DEFAC (Brief description of factorial axes),
RECIP (Clustering using a hierarchical classification of the clusters - reciprocal neighbours method),
PARTI (Cut of the dendrogram produced by the previous step, and optimisation of the obtained partition),
DECLA (Automatic description of the classes of the partition),
SELEC (selecting one categorical variable, in this case),
EXCAT (extracting one categorical variable: the species of iris - selected by the previous step SELEC - to be used in some graphical displays).

We will comment later on this command file (Appendix of the section) which commands the basic computation steps. Instead of editing this file, we will go back to the main menu and execute the basic computation steps.

5) Return to the main menu (**"Return to execute "**)

6) Click on the button: "Execute a command file"

This step will run the basic computation steps present in the command file.

7) Click the button: "Basic numerical results"

The button opens a created (and saved) html file named **"imp.html"** which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu. Note that this file is also saved under another name. The name "imp.html" is concatenated with the date and time of the analysis (continental notation). That file keeps as an archive the main numerical results whereas the file "imp.html" is replaced for each new analysis performed in the same directory.

As usual, this file in also saved under a simple text format , under the name **"imp.txt"** , and likewise with a name including the date and time of execution. **Return.**

8) At this stage, we click on one of the buttons of the lower panel of the main menu (Steps: "VIC")

9) Click directly on the button: "BootstrapView"

This button open the DtmVic-Bootstrap-Stability windows.

9.1 Click **"LoadData"**. In this case (partial bootstrap), the two replicated coordinates file to be opened are named **"ngus_var_boot.txt"** and **"ngus_sup_cat_boot.txt"** (see the small panel reminding the names of the relevant files below the menu bar). In fact, **"ngus_var_boot.txt"** contains only active variables. The file **"ngus_sup_cat_boot.txt"** contains only supplementary categories, for which the bootstrap procedure is also meaningful.

9.2 Click on **"Confidence Areas"** submenu, and choose the pair of axes to be displayed (choose axes 1 and 2, to begin with).

9.3 Tick the chosen white boxes to select the elements the location of which should be assessed, and press the button **"Select"**. Select all the four variables when you open the file **"ngus_var_boot.txt"** , and, later on, the three species when you open the file **"ngus_sup_cat_boot.txt"** .

9.4 Click on **"Confidence Ellipses"** to obtain the graphical display of the active variable points (if the file **ngus_var_boot.txt** has been loaded), or of the supplementary category points (if the file **ngus_sup_cat_boot.txt** has been loaded). We can observe, for the variables, that for example, the "petal lengths" seem to be somewhat redundant with "petal widths", since their ellipses markedly overlap. We can see also that the three categories are significantly distinct (that does not mean that they can be linearly separated...).

9.5 Close the display window, and, again in the blue window, press **"Convex hulls"**. The ellipses are now replaced with the convex hulls of the replicates for each point. The convex hulls take into account the peripheral points, whereas the ellipses are drawn using the density of the clouds of replicates. The two pieces of information are complementary.

Go back to the "VIC" menu.

10. Click on "Visualization" [Visualization of the three species]

A new window entitled **"Visualization, Loading files, Selecting axes"** appears.

We are going to visualise the different species of flowers (categorical variable n° 5) in the plane spanned by the first principal components.

10.1 Click on **"Load coordinate"**

10.2 In the corresponding sub-menu, choose the file: **"ngus_ind.txt"** . The principal coordinates of the individuals (rows) are selected.

10.3 Click then on **"Load or create Partition"**

10.4 In the corresponding sub-menu, choose **"Load a partition file"**. The partition obtained previously from the computation step must then be loaded (its name: "**part_cat.txt** "). The partition induced by the 4 categories of variable number 5 (species of irises) is loaded. This partition has been selected and extracted through the steps SELEC and EXCAT (at the end of the command file; see the sequence of steps above).

10.5 Click on **"MST"** (Minimum Spanning Tree). Choose then the number of axes that will serve to compute the Minimum Spanning Tree: 5 (for example).

10.6 Choose then the number of axes that will serve to compute the Minimum Spanning Tree: 5 (for example).

10.7 Click on **"N.N."** (search for nearest neighbours – limited to 20 NN).

10.8 Click on **"Graphics"**.

10.9 Choose the axes 1 and 2 (default) in the small window **"Selection of axes"** and click on **"Display"**.

In the new window entitled **"Visualisation"** are displayed the individuals in the plane spanned by the selected axes. A random colour is attributed to each species. The button **"Change colour"** allows you to try a new set of colour.

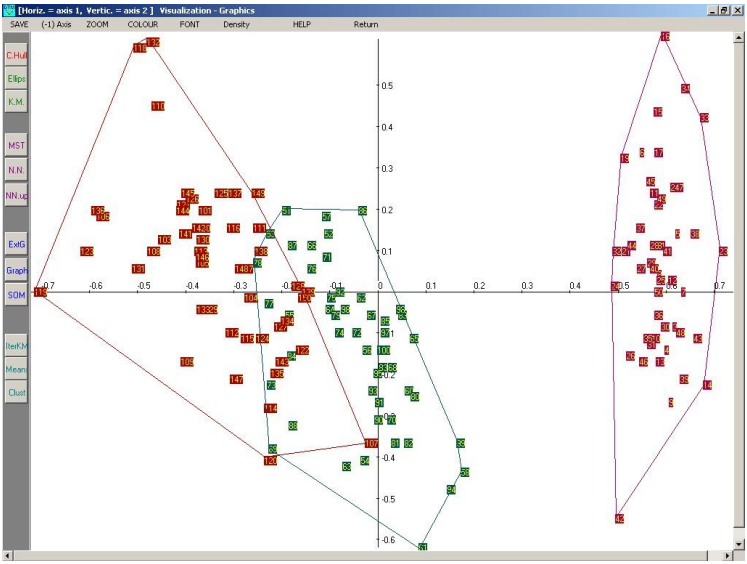
On the vertical tool bar, you can press each button to activate it (red colour), and press it again to cancel the activation (original colour)

-- The button **"Density"** , for sake of clarity, replaces the identifiers of individuals with a single character reminding the cluster (the identifier of individuals and the cluster number can be obtained by clicking on the left button of the mouse in the vicinity of each point).

-- The button **"C.Hull"** (Convex hull) draws the convex hull for each cluster.

-- The button **"MST"** (Minimum Spanning Tree) draws the minimum spanning tree.

-- The button **"Ellipse"** performs a Principal Components Analysis of each cluster within the two-dimensional sub-space of visualisation and draws the corresponding ellipses (containing roughly 95% of the points).



Principal plane from PCA (4 active variables) with convex Hulls of the species.

-- The button **"N.N."** (Nearest neighbours) joins each point to its nearest neighbours. Pressing afterwards the button **"N.N.up"** allows you to increment the number of neighbours up to the 20 nearest neighbours.

At this step, we have obtained a display of the 150 individuals, with either the convex hulls (or the ellipses) corresponding to the three species. This is a classical display of the Iris data in the principal plane of PCA, showing that the first species (number < 51) are well separated from the species 2 and 3.

Go back to the "VIC" menu.

11. Click again on "Visualization" [Visualization of the *clusters*]
We are going to redo the operation of paragraph 10, but instead of loading a partition induced by the 4 categories of variable number 5 (species of irises), we will load a partition produced by a clustering algorithm ignoring the species. Such partition correspond to the steps **RECIP** and **PARTI** (see the command file, below).

11.1 Click on **"Load coordinate"**

11.2 In the corresponding sub-menu, choose the file: **"ngus_ind.txt"** . The principal coordinates of the individuals (rows) are selected.

11.3 Click then on **"Load or create Partition"**

11.4 In the corresponding sub-menu, choose **"Load a partition file"**. The partition obtained previously from the computation step must then be loaded (its name: "**part_cla_ind.txt**"). This partition is derived from the steps RECIP and PARTI .
After loading that partition, all the operations from 10.5 to 10.9 can be carried out again.
It is interesting to visualise the individuals in the plane spanned by the axes 1 and 2.
As suspected, the partition obtained directly from the numerical measurements, ignoring the species, is unable to separate the three species. Only the species "setosa", well separated from the two other species, coincides with a cluster of the partition.
Go back to the "VIC" menu.

12. Click now on the button: "Contiguity" [Contiguity Analysis of Iris data]
We are now going to perform a "Contiguity analysis" using a "nearest neighbours graph" derived from the data. The partition into species is no more taken into account. The partition derived from the previous clustering is also ignored.

12.1 Click on **"Parameters/Edit"**
Choose the item **"Create"**
We are going to enter the parameters needed by a contiguity analysis:
- In the first block entitled **"ncoord = input coordinate file"** , tick "1" (File **ngus_ind.txt** : coordinates of individuals). The contiguity analysis will use the coordinates of individuals as input data.
- In the second block entitled **"npart = partition file"**, tick: "0" (no partition)
- In the third block entitled **"meth = method"**, tick "2" (Contiguity graph defined by nearest neighbours).
- Then we will have to enter the following numerical values :

- npas = 2 (increment for the number of nearest neighbours)
- Min = 4 (minimum number of nearest neighbours)
- Max = 8 (maximum number of nearest neighbours)

Three contiguity analysis will be performed three times for the three (symmetrised) graphs corresponding respectively to 4, 6, 8 neighbours (from Min = 4 up to Max = 8, with an increment of npas = 2).
Then: Click on: **"Validate"** . A summary of the parameters appears.

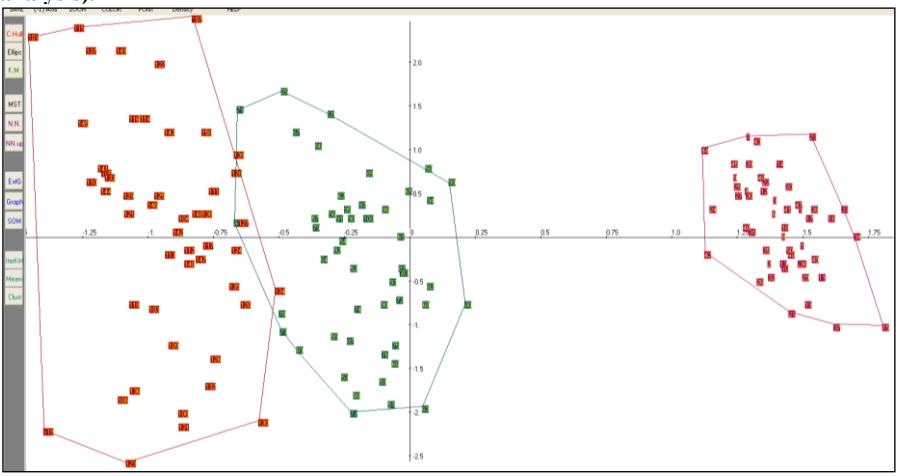
12.2 In the upper bar of the window, Click on **"Execute".** The computations are carried out.
The item **"Results"** of that bar contains technical details about the computations involved in Contiguity analysis.

12.3 Click on **"Contiguity View"**. We are led to the same window of visualisation than previously.
In the menu **"Load coordinates"**, of the new window, choose the file: **ngus_contig.txt**. Instead of using the principal coordinates of PCA (**ngus_ind.txt** as done previously), we use now the result of the Contiguity Analysis **ngus_contig.txt** .
From the menu **"Load or Create a Partition"**, choose the file: **part_cat.txt** . (this file identifies the species)
We cannot compute the Minimum Spanning Tree nor the Nearest Neighbours from the **"ngus_contig.txt"** coordinates.

12.4 Click on **"Graphics"**.
Then choose the axes 1 and 2 (default values)

Choose (tick) the contiguity level number 2, that correspond to 6 nearest neighbours. (level 1 corresponds to 4 nearest neighbours, and level 3 to 8 nearest neighbours).
Click on **"Display"**
Change the colours to obtain a good contrast between species.
Click on **"Convex Hull"** (vertical bar)
The three species are now better separated.

That means that the ("symmetrised") graph of 6 nearest neighbours allows for computing a "local covariance matrix" that can act, in this example, as a "within covariance matrix".
In this example, the principal plane of a contiguity analysis (unsupervised analysis) is similar to the principal plane of a Fisher Linear Discriminant Analysis (supervised analysis).



Contiguity Analysis provides a good separation between iris species.

We must keep in mind that the contiguity analysis did not use the *a priori* knowledge about the species. It is an *unsupervised* method.

Go back to the "VIC" menu.

13. Click again on "Contiguity"

We are now going to perform a "Contiguity analysis" that coincides exactly with a classical Linear Discriminant Analysis.

(Linear Discriminant Analysis is a particular case of Contiguity Analysis. In such a case, the graph involved in this Contiguity Analysis is made of k cliques (complete graphs) corresponding to the k classes of the Discriminant Analysis).

13.1 Click on **"Parameters/Edit"**
Choose the item **"Create"**
We are going to enter the parameters needed by a contiguity analysis:

- In the first block entitled **"ncoord = input coordinate file"** , tick "1" (File: **ngus_ind.txt:** coordinates of individuals). The contiguity analysis will use the coordinates of individuals as input data.
- In the second block entitled **"npart = partition file"** , tick "2" (**part_cat.txt** , categorical) (this partition will now be used to derive a graph).
- In the third block entitled **"meth = method"** , tick "3" (Classical Discriminant Analysis).
- In this case, the following parameters are meaningless. DtmVic asks you to skip them.
- The contiguity analysis will be performed using the graph associated with the partition into species. (all pairs of individual belonging to the same species are joined by an edge; no edge between individuals belonging to different species)

Then: Click on: **"Validate"**. A summary of the parameters appears.

13.2 In the upper bar of the window, Click on **"Execute"**. The computations are carried out.
The item **"Results"** of that upper bar contains technical details about the computations involved in Contiguity analysis. The matrix associated with the graph with its three diagonal blocks of "1" and with the value "0" elsewhere is visible in this listing of results.

13.3 Click on **"Contiguity View"**.
In the menu **"Load coordinates"**, of the new window, choose the file: **ngus_contig.txt** .
In the menu **"Load or Create a partition**, choose the file: **part_cart.txt** (we will identify the "species of iris")
We cannot compute the Minimum Spanning Tree nor the Nearest Neighbours from the **"ngus_contig.txt"** coordinates.

13.4 Click on **“Graphics”**.
Then choose the axes 1 and 2 (default values)
Click on **“Display”**.
Change the colours of the display (**“Colour”**) to obtain a good contrast between classes, then lock the colours.
Click on **“Convex Hull”** (vertical bar)

The three species of iris are well separated, too. But this is less of a surprise, since the Linear (Fisher) Discriminant Analysis aims precisely at separating the classes. We are here in a supervised case. The method uses the *a priori* knowledge of the species of iris to exhibit the coordinates (discriminant functions) that induce the best separations between the classes.

End of example C.2

V.3 Description of graphs through CA

Example C.3. EX_C03.Graphs

Example C.3 aims at describing four simple symmetrical planar graphs from their associated matrices, mainly through correspondence analysis. Unlike the previous example directories, the directory EX_C03.Graphs contains several sub-directories and examples.

Section 1 : Overview of the different directories and files

1.1 Search for the examples directory DtmVic_Examples

1.2 In that directory, open the directory of Example C.3, named EX_C03.Graphs .
This directory comprises three sub-directories.
The sub-directory named **"Chessboard"** relates to the description of a "chessboard shaped graph" (49 nodes corresponding to a square chessboard with 7 rows and 7 columns, the associated matrix being a 49 x 49 binary matrix).
The sub-directory named **"Cycle"** similarly relates to the description of a "cycle shaped graph" (49 nodes).
The sub-directory named **"Geography"** concerns the description of graphs associated with geographical maps (graphs of contiguous regions in Japan recorded under *textual form*, graphs of contiguous "departments" of France recorded under both *textual form* and "external form").

1.3 Open the sub-directory named "Chessboard" .
1.3.1 Open the sub-sub-directory **"Chessboard_numerical"** .
The file: **"Chessboard_7x7_dat.txt"** contains the data set representing the incidence matrix of the graph, with 49 rows and 49 columns. Like any classical data set of DtmVic, each row begins with its identifier. The entry cell *m(i, j)* of such a matrix **M** has the value 1 if the nodes i and j are joined by an edge, 0 otherwise. The identifiers of columns are to be found in the associated dictionary file: **"Chessboard_7x7_dic.txt"** .
That file will be analysed through Correspondence analysis (command file: **"Chessboard_CA.Param.txt"**) and also, through Principal Component Analysis

(command file: **“Chessboard_PCA.Param.txt”**) for the sake of a comparison. The comparison is not favourable to PCA in this particular case. [see, e.g.: *Exploring textual Data*(1998), by L. Lebart, A. Salem, L. Berry, Kluwer Academic Publisher].

Note that these command files can be generated from the button **"Create a command file"** of the main menu, as exemplified in chapters 2 and 3 relating to Principal Components Analysis and Correspondence Analysis.

1.3.2 In the sub-sub-directory **“Chessboard_textual”** .
The file: **“Chessboard_textual_7x7.txt”** contains the same basic information under a quite distinct form: the format relates to responses to open ended questions. Each node of the graph is considered as a respondent, answering to the fictitious open-ended question: “ Please, tell me which are your neighbours ?”. Instead of a binary matrix **M**, we are dealing here with a much smaller data matrix containing the address (column numbers) of the “1” in the matrix **M**. The command file **“Chessboard_Textual.Param.txt”** leads to the same results as those from the correspondence analysis of the previous paragraph, using however a quite distinct sequence of DtmVic steps. It is a “pedagogical example” of bridge between numerical and textual steps of DtmVic. In this type of data, the numbers are not considered as numbers in the mathematical meaning of the term, but as mere sequences of characters. [See below the example of the maps of Japan and France].

1.3.3 The file: **“Chessboard_Extern_7x7.txt”** is present in both preceding directories numerical and textual. It is another possible coding of the Chessboard graph, similar to the previous textual file. But in this case, the number are effectively read as integers, not as simple sequences of characters. The first line of the data set contains the number of nodes (49), then the length of the identifiers (4) and the maximum degree of the graph (upper bound of the numbers of edges adjacent to a single node) (10). Note that each row terminates with the dummy value 0.
Such specific format, the most compact one, can lead directly to a description of the graph in the sub-menu “Contiguity” of DtmVic, without command file.

1.4 Open the sub-directory named “Cycle”.
This sub-directory is the counterparts of the previous one relating to the chessboard graph. Only the shape of the graph is different. The textual coding and the PCA command file are omitted in this case.

1.5 Open the sub-directory named “Geography”.

The two sub-sub-directories files are the counterparts of those relating to the chessboard textual example. The directories **"Japan_text"** and **"France_text"** exemplify the "textual coding" in the case of a maps describing the different regions of Japan and the departments of France.
In the case of Japan, for example, the two first lines of the file Japan_map_text.txt set indicate that the provinces of *Akita* and *Iwate* are contiguous to the province of *Aomori*, etc. The file **"Japan_text_param.txt"** is the corresponding command file. It is identical to the file **"Chessboard_Textual.Param.txt"**, except for the name of the input data set.

Section 2: Running the example "Chessboard_numerical"

Click on the button **"Open an existing command file"** (main menu)

2.1 Reach again the "sub-sub-directory": "Chessboard_numerical".
We are in the framework of either a classical correspondence analysis or a Principal Components Analysis.
a) Data file: "Chessboard_7x7_dat.txt"
b) Dictionary file: "Chessboard_7x7_dic.txt".
c) Command file: "Chessboard_CA.Param.txt" [Correspondence Analysis] or, later on, "Chessboard_PCA.Param.txt" [Principal Components Analysis]

Note again that other "command files" similar to the previous ones, can be easily generated by clicking on the button "Create a command file" of the main menu (Basic Steps). A window "Choosing among some basic analysis " appears. Click then either on the button: SCA – Simple correspondence analysis– or on the button: PCA –Principal components analysis - both of them located in the paragraph " Numerical data", and follow the instructions as shown in Chapter 2.

We will start with correspondence analysis.

2.2 Open the command file: "Chessboard_CA.Param.txt"
After identifying the two data files, four "steps" are performed: ARDAT (Archiving data), SELEC (selecting active and supplementary elements), AFCOR (Correspondence analysis). (See, e.g., : Example A.2)

2.3 Return to the main menu ("Return to execute")

2.4 Click on the button: "Execute"

This step will run the basic computation steps present in the command file: archiving data and dictionary, selection of active elements, correspondence analysis of the selected table.

2.5 Click the button: "Basic numerical results"
The button opens a created (and saved) html file named **"imp.html"** which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu. Note that this file is also saved under another name (see previous examples).

2.6 At this stage, we click on one of the lower buttons of the basic steps panel (Steps: "VIC")

2.7 Click directly on the button: "Visualization" (we skip here the buttons AxeView, PlaneView, etc.)
We are going to visualise the graph.

2.7.1 A new window named **"DTM-Visualization: Loading files, Selecting axes"** appears.

2.7.2 Click on **"Load coordinate"**

2.7.3 In the corresponding sub-menu, choose the file: **"ngus_ind.txt" (individuals)"**. The principal coordinates of the individuals (rows) are selected. [Since the data matrix is symmetrical, it is equivalent to choose **"ngus_var_act.txt"**].

2.7.4 Click then on **"Load or create a Partition"**

2.7.5 In the corresponding sub-menu, select: **"No Partition".**

2.7.6 Click on **"MST"** (Minimum Spanning Tree). Choose then the number of axes that will serve to compute the Minimum Spanning Tree: 8 (for example).

2.7.7 Click on **"N.N."** (search for Nearest Neighbours – limited to 20 NN).

2.7.8 Click on **"Graphics"**.

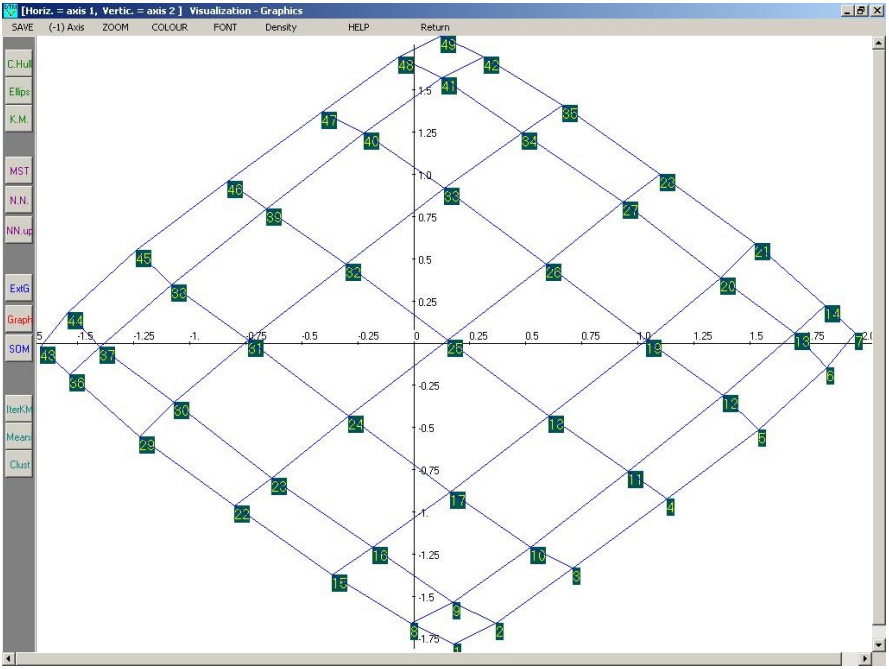
2.7.9 Choose the axes 1 and 2 (default) in the window **"Selection of axes"** and click on **"Display"**.

2.7.10 A new window entitled **"Visualisation, Graphics"** is displayed.

2.7.11 About the window **"Visualisation, Graphics"**

In the window entitled **"Visualisation, Graphics"** are displayed the nodes in the plane spanned by the selected axes. A random colour is attributed to the display. The button **"Change colour"** allows you to try a new set of colour.

On the vertical tool bar, you can press each button to activate it (red colour), and press it again to cancel the activation (initial colour)
– The button **"Density"** , for sake of clarity, replace the identifiers of nodes by a single character.
– The button **"C.Hull"** (Convex hull) is irrelevant here.
– The button **"MST"** (Minimum Spanning Tree) draws a possible minimum spanning tree.
– The button **"Ellipse"** is not relevant here.
– The button **"N.N."** (Nearest neighbours) joins each point to its nearest neighbours. Pressing afterwards the button **"N.N.up"** allows you to increment the number of neighbours up to the 20 nearest neighbours.
– The button **"ExtG"** allows you to load the graph in "External format".
– The button **"Graph"** (only when an extern graph has been loaded) allows you to draw the edges of the graph (Interesting to watch the distortions of the graph according to the selected pairs of axes.

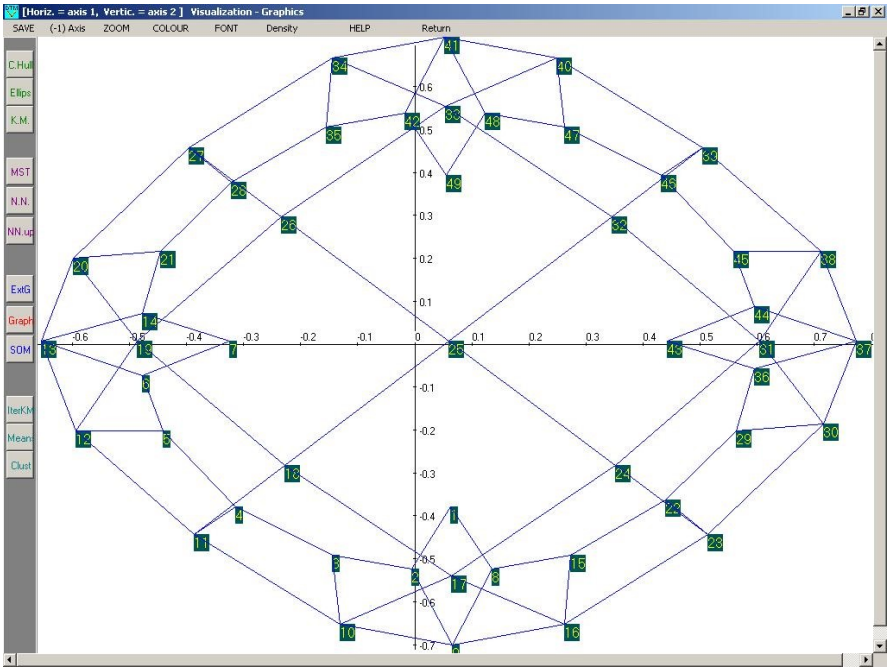


The 7 x 7 chessboard is well described by CA.

Important: in this particular application, the Minimum Spanning Tree and also the nearest neighbours are computed from the coordinates of the nodes in a space spanned by the first components.

2.8 Go back to the main menu.

2.9 Redo all the operations 2.2 to 2.7, opening now, during step 2.2, the command file: "Chessboard_PCA.Param.txt" (Principal Components Analysis). It will be seen through this example that PCA is less faithful than CA vis-à-vis the description of the graph structure.



The 7 x 7 chessboard is not so well described by PCA.

Section 3 : Running the example "Chessboard_textual"

Click on the button "Open an existing command file" (panel *Basic Steps* of the main menu)

3.1 Open the "sub-sub-directory": "Chessboard_textual":

We are in the framework of a textual analysis similar to the one of examples which aimed at describing the responses to an open ended question in a sample survey (examples A.5, A.6, B.1 to B.3).

We find in this directory the text file and the command file.
(in this particular context, there are neither data file nor dictionary file: the questionnaire comprises one pseudo open-ended question, put to each node: " Which are your neighbouring nodes ?")

3.1.1) Text file: **Chessboard_textual_7x7.txt**
The format is the same as in Example A.5. Since the responses may have very different lengths, separators are used to distinguish between individuals (or: respondents). Individuals (here: nodes) are separated by the chain of characters "----" (starting column 1) possibly followed by an identifier.

3.1.2) Command file: **Chessboard_Textual.Param.txt**
The computational phase of the analysis is decomposed into "steps". Each step requires some parameters briefly described in the main menu of DtmVic (button: **"Help about command parameters"**).

3.2) Open the command file: Chessboard_Textual.Param.txt
After identifying the input textual data file, four "steps" are performed:
ARTEX (Archiving texts),
SELOX (selecting the open question),
NUMER (numerical coding of the text),
ASPAR (correspondence analysis of the [sparse] contingency table "respondents - words").
We will not comment on this command file which commands the basic computation steps (see Example B.1). Instead of editing this file, we will content ourselves here in going back to the main menu and execute the basic computation steps.

Recall that such a command file can be generated by clicking on the button **"Create a command file"** *of the main menu (DTM: Basic Steps). A window* " Choosing among some basic analysis " *appears. Click then on the button:* **VISURESP** *, located in the paragraph "Textual data", and follow the instructions as shown in Chapter 3. Note also that in this simple data case (only one fictitious "open question"), it is possible to consider each response as a text. In such a case, the response separators "----" should be replaced with a text separator "****", as in example A.4. Instead he analysis* **"VISURESP"**, *it is then necessary to perform the analysis* **"VISUTEX"**.

3.3) Return to the main menu ("Return to execute")

3.4) Click "Execute a command file"
This step will run the basic computation steps present in the command file: archiving text, correspondence analysis of the lexical table.

3.5) Click the "Basic numerical results" button
The button opens a created (and saved) html file named **"imp.html"** which contains the main results of the previous basic computation steps. After perusing these numerical results, return to the main menu. Note that this file is also saved under another name. (see for example Chapter 2).

From the step NUMER, we learn for instance that we have 49 responses, with a total number of words (occurrences or token = here: edges of the graph) of 217, involving 49 distinct words (here: neighbours). Note that each node has been considered as its own neighbour.

3.6 Click the PlaneView button, and follow the sub-menus...
In this example, four items of the menu are relevant **"Active columns (variables or categories)", "Active rows (individuals, observations)", "Active columns + Active rows", "Active individuals (density)".** The graphical displays of chosen pairs of axes are then produced.

3.7 Click on "Visualization"
All the steps of the previous section 2.7 could be carried out likewise.

Section 4 : Running the example "Chessboard_Extern"

There are *neither command file, nor dictionary file* in this directory, since the specific type of coding of the graph ("external coding") provides a direct entry into the "Contiguity" menu.

In the menu "Visualization, Inference, Classification", click on the button: "Contiguity".

4.1 Click on "Parameters/Edit"
Choose the item **"Create"**
We are going to enter the parameters needed by a graph description:
- In the first block entitled **"ncoord = input coordinate file",** tick "0": "No coordinate file (simple description of an external graph)".
- In the second block entitled **"npart = partition file"**, tick "0" (No partition)
- In the third block entitled **"meth = method"**, tick "4" (External contiguity graph).

Then: Click on: **"Validate"** (as prompted by a message).

The parameter should be in the same directory as the external graph file (as suggested by a pop up message).

4.2 In the upper bar of the window, Click on "Execute".
A new window appears, and you are asked to choose the external graph file. It is in this example the file: **"Chessboard_Extern_7x7.txt".**
The computations are carried out.
The item **"Results"** of that bar contains some technical details about the computations involved in the correspondence analysis of the associated matrix **M** (These results are saved in the file **"imp_contig.txt"**).

4.3 Click on "Visualisation".
In the menu **"Load coordinates"**, of the new window, choose the file: **anagraf.txt** . (graph view through Correspondence Analysis)
In the menu **"Load or Create Partition"**, choose the item: **No partition**.
We can compute and load the Minimum Spanning Tree or the Nearest Neighbours from the "anagraf.txt" file coordinates, choosing for instance 12 axes (maximum number allowed in this version = 30).

4.4 Click on "Graphics".
Then choose the axes 1 and 2 (default values)
Click on **"Display"**. Change the colours if necessary.
Once again, all the steps of the previous section 2.7 could be carried out likewise.

4.5 About the window "Visualisation, Graphics"
To represent the edges of the original graph, click on the button **"ExtG"** (External Graph) of the vertical bar.
Open then again the file **"Chessboard_Extern_7x7.txt".** .
Click on the button **"Graph".** The button **"Graph"** produces the original graph as recorded in the file. This allows you to observe the distortion of the planar graph in the spaces spanned by the axes 3 to 12.
[It is the multidimensional Guttmann effect (See Benzécri, (1973),(in French) "L'analyse des données", Tome II B, Chapter 10, "Sur l'analyse de la correspondance définie par un graphe", pp 244-261)].

Section 5 : Running the example "Cycle_Numerical"

This section is identical to Section 2 (Running the example "Chessboard_Numerical"). The graph has now the shape of a cycle, with the same number of nodes.

The homologues of the files **“Chessboard_7x7_dat.txt”**, **“Chessboard_7x7_dic.txt”** and **“Chessboard_CA_Param.txt”** are now respectively **“Cycle_49_dat.txt”, “Cycle_49_dic.txt”** and **“Cycle_CA_Param.txt”** .
They can be found in the directory **“Cycle”**.

Section 6 : Running the example “Cycle_Extern”

This section is identical to Section 4 (Running the example “Chessboard_Extern”). The graph has now the shape of a cycle, with the same number of nodes.
The homologue of the file **“Chessboard_7x7_Extern.txt”** is the file: **“Cycle_Extern_49.txt”** .

Section 7 : Running the example “Japan_map”

This section is identical to Section 3 (Running the example “Chessboard_Textual”). The graph is now a sketch of a map of Japan, presented as a set of responses to the open question “Which are your neighbouring regions”, the “respondents” being the same regions of Japan…

The homologue of the directory **“Chessboard_Textual”** is : **“Japan_map”** whereas the homologue files of **“Chessboard_textual_7x7.txt”** and **“Chessboard_textual_Param.txt”** are respectively: **“Japan_map_Textual.tex.txt”** and **“Japan_map_Textual.Param.txt”** .

Section 8 : Running the examples “France_map”

This section is identical to Section 3 (Running the example “Chessboard_Textual”). The graph is now a sketch of a map of France, presented as responses to the open question “Which are your neighbouring departements (= counties)”, the “respondents” being also the departements of France…

The homologue of the directory **“Chessboard_Textual”** is : **“France_map”** whereas the homologue files of “ **Chessboard_textual_7x7.txt”** and **“Chessboard_textual_Param.txt”** are respectively : **“France.tex.txt”** and **“France.Param.txt”** .
The homologue of the file **“Chessboard_7x7_Extern.txt”** is the file: **“France_Extern.txt”** .

End of example C.3

V.4 Structural Compression of Images

Example C.4. EX_C04.Images

(Structural Compression of Images through SVD, CA and Discrete Fourier Transform)

Examples C.4 are mainly pedagogical examples which serve as an illustration for the compression effect of principal axes techniques (keeping a limited number of principal axes in Singular Value Decomposition and Correspondence Analysis) in the domain of image analysis (domain rather unexpected for most DtmVic users). Comparison is made with Discrete Fourier Transform (keeping a limited number of terms from the expansion) that takes into account the relative locations of the pixels.
It does not make use of data in internal DtmVic text format, since it deals with digitalized images. A simple rectangular array of integers suffices: there is no need for identifiers of rows or column.
In fact, three particular formats will be used: rectangular arrays of levels of grey (simple text format), plain “pgm” format (acronym derived from "Portable Gray Map") and, for colour images, plain “ppm” format (acronym derived from "Portable Pixel Map")
A specialized interface is provided via the button **“DtmVic Images”** of the main menu.

1. About the data (some image formats)

To have a look at the data,

1.1 Search for the directory “ DtmVic_Examples”.

1.2 Search for the sub-directory “DtmVic_Examples_C_NumData” in “DtmVic_Examples”.

1.3 In that directory, open the directory of Example C.4: “EX_C04.Images”.

1.4 Four sub-directories correspond to four examples:

“1_Cheetah_txt”,
“2_Baalbeck_pgm”,
“3_Cardinal_ppm_color”,
“4_Extra_pgm_ppm”

All these file can be examined via a text editor (such as “notepad” included in Windows, or a free software such that “notepad++”, or “TotalEdit”, etc.).

1.5 For greyscale *(US: grayscale)* images, two input format are available:

1.5.1 Simple text format :

The data table contains positive integer s <= 255 that are the values of the level of grey for each pixel (no identifiers). This is the case of the image **“cheetah.txt”** in the folder **“1_Cheetah_txt”** (adapted from "The Data Compression Book", Mark Nelson, M&T Publishing Inc., 1992). Such a format that does not contain explicitly the size of the image is the simplest one. Because of its rusticity, this format is neither used nor provided by the usual image processing software.

1.5.2 The "pgm" format : (Portable Grayscale Map) (look at the example: **"2_Baalbeck.pgm"**, using a text editor or a notepad)

The PGM format is a simple and transparent greyscale file format.
The differences with the plain format are:
There is one image in a file (general pgm format can cope with several images).
The first line contains the format identifier: P2.
The second and the third lines contain three integers: number of columns, number of rows, and the maximum value (255).
Then the table is displayed row-wise.
Each pixel in the table is represented as an ASCII decimal number (<255).
Each pixel in the table has at least one white space before and after it.
No line should exceed 72 characters.
For more information about such a format, please consult (e.g.): http://netpbm.sourceforge.net/doc/pgm.html

1.5.3 The "ppm" format for colour images:

For (small)colour images, the input format is the ppm text format (acronym for: portable pixel map).
Look at the example **"3_Cardinal.ppm"** , via a text editor or a notepad.
The three integers (levels of: Red, Green, Blue) describing each pixels are located consecutively in the same row.

Both pgm and ppm files can be obtained through exportation from the free software "Open office", using a jpeg file as an input.

2. Running a first example (simple greyscale format)

In the Main Menu, Click on the button **"SVD and CA of images"** (in the section:

"DtmVic-Images").
The first thing to do is to select an image. One of the three buttons, on the left hand side of the window, have to be selected to open the image, according to its format.

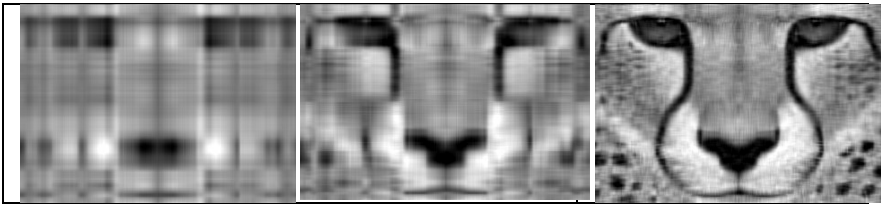
2.1 Click on the first button **"Read (formatted txt file)"** in the section **"Open greyscale image"**.

2.2 In the directory **"EX_CO4_Image"**, open the sub-directory **"1_Cheetah_txt"**. Within **"1_Cheetah_txt"**, open the file **"Cheetah.txt"**.
A message-box recalls the size of the image file. If you wish to visualize the original image, in the section **"Visualization"**, click on: **"Image (greyscale)"**.

2.3 Then, in the lower left part of the window, in the section **"Compression techniques"**, click the button: **"Correspondence Analysis"** (to begin with).

2.4 If you wish to obtain an overview of the data reconstitution, from 1 to 100 axes, Click directly on the button: **"Series from first term to total"**, in the right hand side panel. You can then observe the progressive reconstitution of the original data table (i.e.: the image).

2.5 If you are interested in focusing on a specific number of axes, then select the required number of axes in the vertical corresponding list, and visualize each image. Note that all the created images are saved in bitmap format (extension: ".bmp") in the directory of the analysed image file.



Description via CA (respectively 1, 4, and 16 axes)

2.6 Instead of Correspondence Analysis, you can choose **"Singular Value Decomposition"**, and redo all the operations 2.4 and 2.5.

2.7 If you select the lower button: **"Discrete Fourier Transform"**, a new window is displayed.

2.8 You have then to select the mode of computation of the Fourier series ("Row-wise" or "Column-wise"). Select **"Row-wise"**, for example.

2.9 Then, as previously, you can go directly to the right-hand side panel, and press the button: **"Series from the first term to total (greyscale)"** . The comparison of the obtained reconstitution (according to the number of kept terms in the Fourier decomposition) with the preceding reconstitution (using CA or SVD) is quite interesting.

2.10 If you are interested in focusing on a specific number of terms, then select the required number of terms in the vertical corresponding list, and visualize each image.

Note 1: Incidently, the graphical display of levels of grey for each row can be obtained from the button **"Curves of grey levels"** (press it several times to scan the whole image).

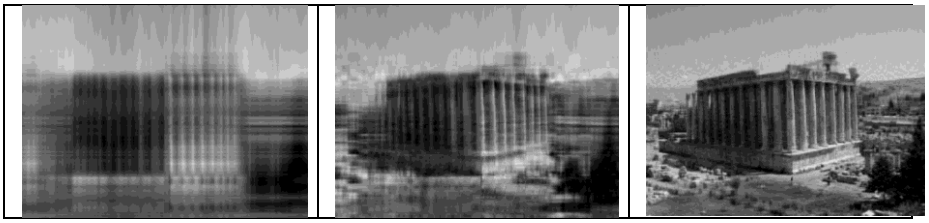
Note 2: All created images are saved in bitmap format (extension: ".bmp") in the directory of the analysed image file.

Note 3: The compression through SVD or CA does not depends on the order of rows and columns of the table (unlike the Fourier compression). Nevertheless, the "structural compression" (i.e. ignoring the relative locations of the pixels) gives worthwhile results.

3. Running other examples:

3.1 **Baalbeck Temple example.**
Click on the second button **"Read (pgm format)"**, always in the section **"Open greyscale image"**.
In the directory **"EX_CO4_Image"**, open the sub-directory **"2_Baalbeck_pgm"**. Within **"2_Baalbeck_pgm"**, open the file **"Baalbeck.pgm"**. A message-box recalls the size of the image file.
If you wish to visualize the original image, in the section **" Visualization"**, click on: **"Image (greyscale)"**. Then redo all the operations 2.3 to 2.10.
This example is interesting since it emphasize the fact a strong pattern (here: the columns of the temple) can contaminate the reconstitution through principal axes

methods. Such a contamination does not exist in the case of Fourier reconstitution row-wise, as expected...).



Baalbek temple: reconstitution with 2, 9, and 50 axes

3.2 **Cardinal (of Mauritius) example.**
Click on the third button **"Read (ppm format)"**, in the section **"Open colour image"**.
In the directory **"EX_CO4_Image"**, open the sub-directory **"3_cardinal_ppm_colour"**.
Within **"3_cardinal_ppm_colour"**, open the file **"cardinal.ppm"**.

A message-box recalls the size of the image file. If you wish to visualize the original image, in the section **"Visualization"**, click on: **"Image (colour)"**.
Then redo all the operations 2.3 to 2.10.

Cardinal of Mauritius: reconstitution with 2, 10, and 100 axes

3.3 **Extra_pgm_ppm example.**
The folder **"4_Extra_pgm_ppm"** contains two versions (colour and grey) of an image of a young boy using a broom.
Proceed as in section 3.1 for the pgm image, and as shown in section 3.2 for the ppm image.

Note: Remind that in the ppm format, the three basic colours (RGB = Red, Green, Blue) corresponding to each pixel have consecutive locations in the same row (the length of which is consequently three times the number of pixels).

The compression through SVD or CA does not depends on the order of the columns, that means that we don't use the fact that the three colours are relative to a same pixel! Nevertheless, the "structural compression" works.

In this case, the Fourier series row-wise is not adapted (unless we choose to juxtapose column-wise the three tables Red, Green, Blue, and, in so doing, abandon the ppm format). The reader can compare the results from the two Fourier compressions row-wise and column-wise.

End of Example C.4

Chapter VI

Importation procedures

This chapter contains a series of examples of data importation that aim at capturing or transforming data to comply with the DtmVic *format files. Each example corresponds to a directory included in the sub-directory* "DtmVic_Examples_D_Import" *included in the directory* "DtmVic_Examples" *that has been downloaded with* DtmVic.

(Four Importation examples D.1—D.4)

VI.0 Capture of dictionary and data *(Dictionary and Data from the keyboard)*
Preliminary Example D.0.

VI.1 *Importation from an Excel File*

Example D.1. EX_D01.Importation.XL.

(Importation of dictionary, numerical and textual data from an Excel ® file)

VI.2 *Importation of Textual Data from a free format file*

Example D2. EX_D02.Importation.Text.Free

(Importation of Textual Data from a specific free format file)

VI.3 *Importation of both numerical and Textual Data from a XML format file.*

Example D3. EX_D03.Importation.Text.Num.XML.

VI.0 Capture of numerical data and dictionary

Preliminary Example D.0

Recording the dictionary presented above in the introduction and recording some data.

1. **Click on the button "Data Importation , Preprocessing, Data Capture, Exportation"**. (Basic Step from the main menu of DtmVic).
2. A new window appears.
3. Choose the item : **'Building the Dictionary (manually)"**.
4. In the new green windows, three yellow cases are ready to receive the information relating to the first variable.

   "**Variable number"** , : " 1 "(default value).

   "**Variable identifier"** , type: Gender ;

   "**Variable type"** : type " 2 " [the type of a variable is the number of its categories](special value: 0 for a numerical variable).
5. Since the number of categories is greater than 1, a second green window appears, inviting you to record the names of the categories: male, female .
6. A second variable is then proposed. It will be the age, the type of which is " 0 ", it is a numerical variable. No window appears since no categories are involved.
7. A third variable is proposed, you may record a categorical variable "age" in 4 categories… etc.
8. A report of the recorded data is printed is the lower window, while the right hand side window displays the dictionary in DtmVic internal format. It is that dictionary that will be recorded at the end of the capture process.

9. When all variables are recorded, one must click on the button **"Save dictionary"**. We suggest to build a new directory (or: folder) in a workspace that is convenient to you, to open that directory, and to save the dictionary as "dic.txt".

10. Then: **"Return"** .

We are back in the Data Capture window.

11. Choose now the item **"Creating the data file"**.

12. The window "Creating data source file" appears.

13. Click on **"Load Dictionary"**

    - Ignore, at this stage, the button **"Update an existing data file"** .

14. The previous chosen directory appears. Select the dictionary: "dic.txt".

15. That dictionary is displayed in the upper right window.

16. Simultaneously, the yellow upper left window is ready to receive the data relating to the first individual : its identifier, (type for example " Rita "),

17. Type then the value of the first variable for that individual: Gender. A click on the right border of the caption window displays the two possible values. Let us choose "female ", to be consistent with the identifier.

18. The second variable, Age, is then proposed to the user. A numerical value must be inserted in the window, etc.

19. At the end of the record, the second individual or observation is proposed…
    We suggest to record 3 or 4 individuals in this exercise (more if you wish...)

20. Then press the button **"Save Data"**.

    The same directory is again proposed. A name should be proposed for the data file. The extension ".txt" is recommended, to facilitate a quick access to the content of the file. Let us select for example the name "dat.txt".

    Press then the button: **"Create a first parameter file"**.

21. The window **"Creating a starting parameter file"** appears.

22. Click on **"Create a parameter file"**.

23. A DtmVic parameter file is displayed in the lower window.

24. That parameter file is automatically saved under the name: **"param_start.txt"**.

    The parameter file does not include any statistical analysis command, except basic counts of categories, together with a computation of extreme and average values for the purely numerical variables.

    It is only meant here as a check of the capture of the data.

    Comments about the "first parameter file"
    After an identification of the two input files, three "steps" of DtmVic are involved: The step "ARDAT" that archives data and dictionary. The step "SELEC" that selects the variables for the subsequent processing. In this case, all the available variables are selected. The step "STATS" that computes the basic statistics mentioned above.

Click on **"Execute"**. Read the results by clicking on the **"Basic numerical results"** item of the menu. These results are saved under the names: "imp.html" an "imp.txt" in the same directory.

End of example D.0

VI.1 Example D.1: EX_D01.Importation.XL

Importation of numerical and textual data in "Excel ® format".

Transforming a specific XL (csv) format file into DtmVic dictionary file, text file and data file.

This importation procedure can be applied to any text file (.txt) having the following features, for n individuals and p variables:
The first row (p + 1 elements) contains the generic name of the identifiers (for example: *ident*) and the p names of the variables (no blank space allowed within the name, less than 20 characters, preferably less than 10) separated with a semicolon (or a comma, or a tab). Blank spaces are allowed between names (free format).
The *n* remaining rows contain *p + 1* elements: the identifier of the individual (less than 20 characters) and the values of the *p* variables (for categorical variables, no blank space are allowed within the alphanumeric values; preferably less than 10 characters) separated with a semicolon (or a tab). Blank spaces are allowed between values (free format) and, evidently, within textual variables (responses to open-ended questions, for example).

Only one type of separator (semicolon or tab) can be used in a file. Such file can be obtained by saving an Excel file as either a "CSV file", or a "tab-separated text file".

1- Looking at the data, preliminary steps

The folder **"EX_D01.Importation.XL"** contains the file **"datbase_global.xls"**.

The file datbase_global.xls corresponds to a frequent situation: the first row of the table contains the variable identifiers, the first column comprises the observations identifiers.

To begin with, we will have a look (**outside DtmVic**) at the original file to be imported.
This file is under Microsoft Excel ® format. The reader who is not provided with that software should skip the next instructions… or use the free software "Open Office" instead.

1.1 Search for the examples directory DtmVic_Examples

1.2 In that directory, open the directory of example D.01, named EX_D01.Importation.XL

1.3 Click on the file: "datbase_classical.xls" (basic dictionary **and** data **and** texts) to obtain a view of the data through an Excel spreadsheet.
- The first row contains the names of the 17 variables (there are 18 columns, but the first one relates to the identifier of individuals).
Note again two important constraints:
a) the names of variables must have less than 20 characters,
b) these names should not contain blank spaces (replace them by underscores, if any).
Note that these names will be truncated down to 10 characters to build the identifiers of the categories. It is then important that these first 10 characters allow for identifying the variable.
The remaining rows consist of 1043 lines (it is the same sample of individuals from the socio-economic sample surveys serving as example in the applications A.5, B.2).
The sequence of characters in the first cell of each line is the identifier of individual, the following sequences being the values of the 17 variables. Blank cells means "no-answer" or "missing value".

1.4 We must save this file as a text file in ".csv" format. (command: File, then "Save as")We obtain a free format file with semicolons as separators. The file in "csv" format is provided in the example directory.

Important:
1.4.1 If there are some semicolons in the data file, they should be replaced by another symbol before saving the "Excel file" as a "csv file".

1.4.2 Note also that before saving the file, the format of the cells containing numerical values must be "standard", to avoid some additional small blank spaces in numbers of more than 3 digits that are misinterpreted by the csv file. In the French version and in some

European versions of Excel, the “decimal commas” should be replaced by the usual decimal dots.

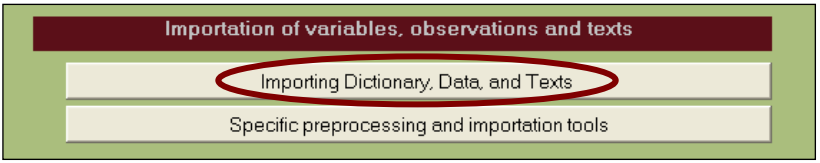
1.4.3 If your version of Excel does not allow for “saving as a csv file”, you can save the file using “tabs” as separators, and then, change the “tabs” into “semicolons”, alteration allowed by the button: **“Change tabs into semicolons”** (see below). This supposes that the initial data set does not already contain semicolons: if semicolons are present, you should replace them with another symbol before the importation process).

1.4.4 In many versions of Excel, the csv format uses commas as separators, instead of semicolons. You can then transform these commas into semicolons (provided that the initial data set does not already contain semicolons: you should replace then these semicolons with another symbol before the importation process).

2) Sequence of operations

2.1 Click on the button: “Data Importation, Preprocessing, Data Capture, Exportation”, (Basic Steps from the main menu of DtmVic). A new window appears.

2.2 Choose the item: “Importing Dictionary, Data and Texts”. The new window “Data Importation” is displayed.



2.3 Press the button entitled: “Excel® type files (saved as csv files)”.

A new window entitled “Data Importation from an Excel (r) file” appears.

If the Excel file has been saved using “tabs” or “commas” as separators, click on one of the optional buttons:

“Change tabs into semi-colons”.
“Change commas into semi-colons”.

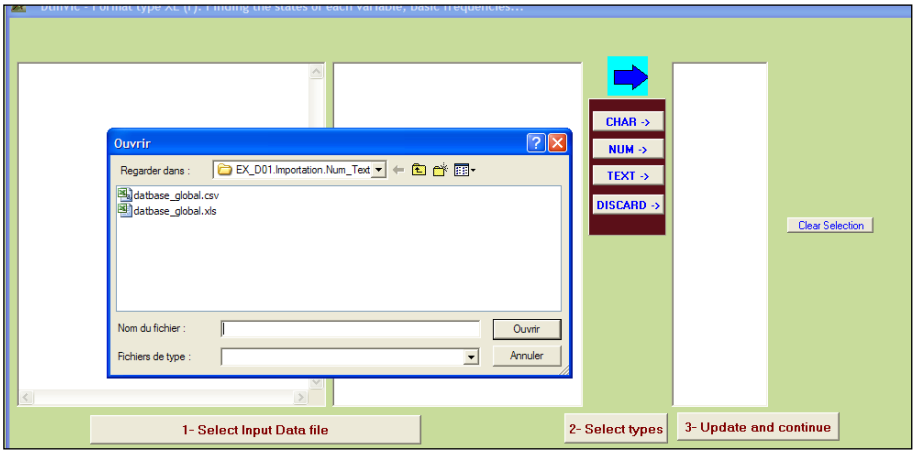
Select the file saved with tabs or commas, and convert it. Note that a new name is given to the created file. The importation process will continue using this new file.

2.4 Then, click on the button: "Start the Importation Process"

In the new window, click on: **"1.Select input data file"** (widen the window if necessary).

Select the previously saved file: **"datbase_global.csv"**(or the file produced by one of the previous buttons "0")
The left hand side memo contains, for each variable, all its observed values. In the case of continuous numerical variables, the number of values could be the same as the number of observations. In the case of textual data, the number of values is the number of "words" (separators : blank, periods, commas)
- The central memo is a summary of the previous one. For each variable, we can read within the brackets the number of distinct values observed in the file.
- The letter (A) in parenthesis means that some letters or non-numerical values have been observed.
- The letter (N) indicates that only numerical values have been obtained.
It is then easier to choose the types of the variables:

- categorical (**CHAR**),
- numerical (**NUM**),
- textual (**TEXT**),
- variables to be abandoned (**DISCARD**).



To choose these types, you have then to select one or several consecutive variables in the list, and choose, for each variable, one keyword among the four keywords {CHAR, NUM, TEXT, DISCARD}.

- " **CHAR** " means that we are dealing with a category of a nominal variable. Such variable could be coded with at most 6 characters. For instance, 'male' and 'female' for coding the gender (or "0" and "1", or "10" and "20" …). Conventionally, the first item (identifier) should be a "**CHAR**".

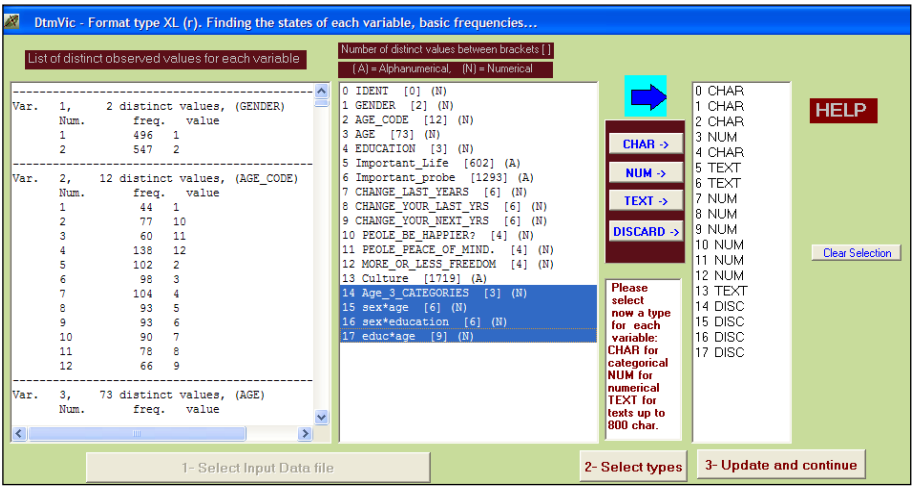
- " **NUM** " means that we are dealing with a purely numerical variable.

- " **TEXT** " means that the records (up to 8000 characters, another constraint) will feed the textual data file.

- " **DISCARD** " means that the records (whatever the prior status) will be suppressed in the imported file.

Clearly, a variable with a few distinct values containing letters (A) should be a categorical variable "CHAR".

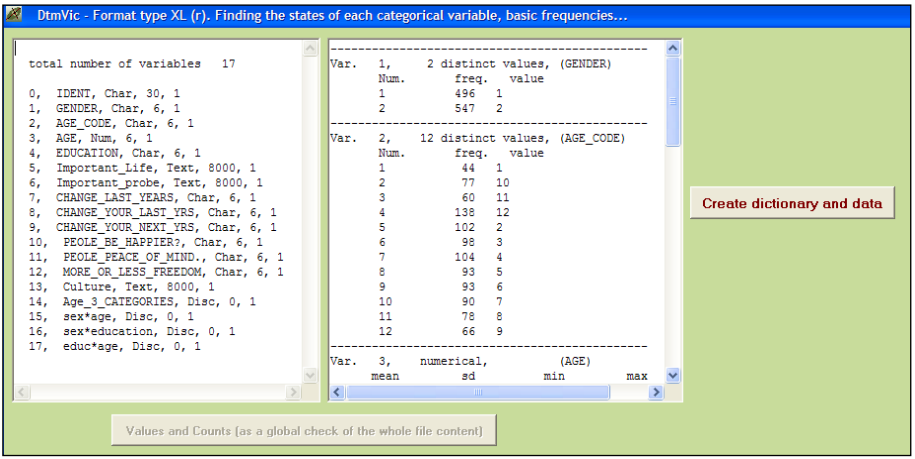
Similarly, a variable with hundreds of purely numerical values (N) will probably deserve the type: "NUM".



If expected numerical values contain letters (A), it could be than in the original Excel file, the missing values or ”Do not apply (DNA)” are represented by alphanumeric symbols. These symbols should be replaced with blank spaces in the original file, or directly in the “csv file” before the importation. If you give the status “NUM” to a variable whose values contain letters, the importation process will be stopped before being completed, entailing a waste of time.

2.5 Once the attribution of types is completed, click on the button “3.Updating and continue” .

2.6 In the new window, **Click on “Values and counts”**.
A further check of the consistency of the selected types or the variables.
A list of all the categories found in the data file, with the corresponding frequencies is displayed. Basic parameters are also provided for numerical variables. We will not dwell on this output serving mainly as a technical check.



2.7 Click then on "Create dictionary and data".

A new window entitled "Creating a dictionary and a data file" appears on the screen.

2.8 Click on "Name for the new dictionary".
You have to choose a name for the forthcoming DtmVic dictionary, always in the same directory (the extension ".txt" is recommended) select for example: "dtm_dic.txt".

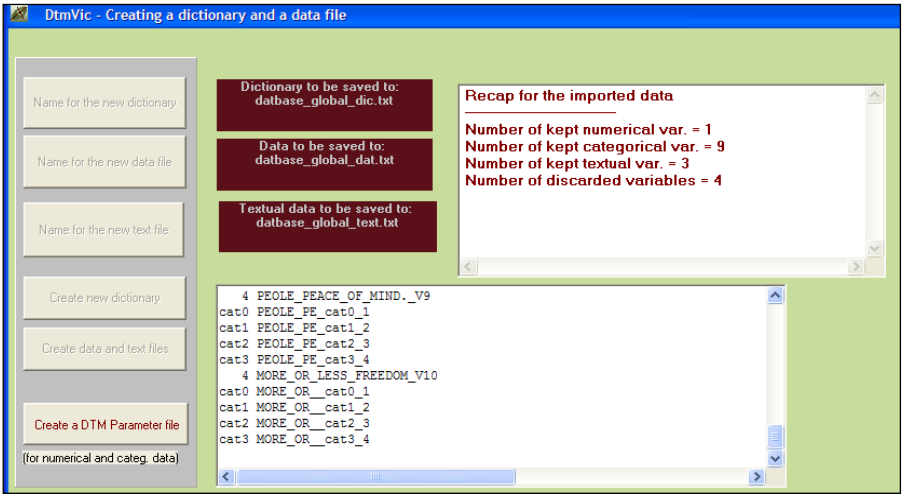
2.9 Click on "Name for the new data file"
You have to choose a name for the forthcoming DtmVic data file, always in the same directory (the extension ".txt " is recommended). Select for example: "dtm_dat.txt"

2.10 *[if textual data have been selected]* **Click on "Name for the new text file"**
You have to choose a name for the forthcoming DtmVic text file, always in the same directory (the extension ".txt" is recommended). Select for example: "dtm_text.txt"

2.11 Click on "Create new dictionary"
A DtmVic dictionary is created (number of lines = total number of variables + number of found categories). The DtmVic dictionary is displayed in the right hand side memo.

2.12 Click on "Create new data file".

2.13 *[if textual data have been selected]* **Click on "Create new text file".**



A message box producing the numbers of different types of variables is displayed.

2.14 Click on "Create a first parameter file" *(optional).*

The window "Creating a starting parameter file" appears.

[Reminder: In DtmVic, the phrases "Parameter file" and "Command file" are equivalent].

A DtmVic parameter file (or: command file) is displayed in the lower window.

The command file is automatically saved under the name: **"param_start.txt"**.
The command file does not include any statistical analysis command, except basic counts of categories, together with a computation of extreme and average values for the purely numerical variables.
It is only meant here as a check of the importation of the data.

Comments about the "first command file"

After an identification of the two input files, three "steps" of DtmVic are involved: The step "ARDAT" that archives data and dictionary. The step ">SELEC" that selects the variables for the subsequent processing. In this case, all the available variables are selected. The step "STATS" that computes the basic statistics mentioned above.

2.15 Click on "Execute". Back in the main menu window, the sequence of steps is displayed.

2.16 Click on the button: "Basic numerical results".
The button opens a created (and saved) html file named **"imp.html"** which contains the main results of the previous basic computation steps. After perusing these numerical results, **Return** to the main menu. Note that this file is also saved under another name: The name **"imp.html"** is concatenated with the date and time of the analysis (continental notation). That file keeps as an archive the main numerical results whereas the file **"imp.html"** is replaced for each new analysis performed in the same directory. Likewise, a simple text format file **"imp.txt"** is created and saved.

If the original Excel file contains textual variables (generally: responses to open-ended questions) a DtmVic textual file is created (the name of which has been given during the step 2.10). The step **"VISURESP"** (in the panel open by the button : **"Create a command file"** of the main menu) allows you to check the consistency of that textual file.

End of example D.1

VI.2 Example D.2:

EX_D02.Importation.Text.Free

Importation of textual data in "free format".

Transforming a specific free format text file into DtmVic text files (type2) .

The DtmVic format for textual data (type 2) is described in the first chapter.

It contains two types of separators: separators of individuals: "----" and separators of questions "++++", located in columns [1,2,3,4]. There is one constraint for the length of a line (200 characters) but, in principle, no constraint about the number of lines for one question or for one individual. However the number of open questions should not exceed 12, the number of closed questions should not exceed 1200, and the number of individuals is limited to 30,000 in the present version of DtmVic.

Remark about DtmVic text file type 1:
Another separator (separator of texts : "****") could be used in the case of DtmVic text file type 1, exemplified by Table 3 of the introduction.
This kind of internal format can be easily built directly from the original corpus of texts without using the importation procedure (see the example: EX_A04.Text-Poems of Chapter 3).
No importation procedure is needed in that case.

To begin with, we will have a look at the original textual data to be imported.

1- Looking at the data, preliminary steps

We will use the editor of the button **"Open an existing command file"** of the main menu of DtmVic as a simple text editor.

1. **Click on the button "Open an existing command file"**

2. **Search for the examples directory DtmVic_Examples .**
3. In that directory, **open the directory of example D.03** , named **EX_D03.Importation.Text.Free**
4. **Select the basic text file: "TDA1_text_free.txt"** .
   (the responses are those involved in application examples A.5, A.6, B.1, B.2).

The free format of that file is the following:
- Each line corresponds to an individual (a respondent) (up to 100,000 characters, no line-break or "end of line" allowed).
- The separators are the character #, which serves to separate the identifier of a respondent from its first response, and also to separate two consecutive responses.
We deal here with three open ended questions, since we have three # per line (a character # at the end of a line means an empty response to the last open question).

5. **Nature of the importation process**

The importation process consists in building a DtmVic text file from the original text file.
It consists in inserting the different separators.
The DtmVic format is closer to the usual format of texts in everyday life, easier to consult and peruse. However, the matching of both textual and numerical files is more easily carried out with the basic textual format (one individual = one row).

2 - Sequence of operations:

2.1 Click on the button "Data Importation , Preprocessing, Data Capture, Exportation", (Basic Step from the main menu of DtmVic). A new window appears.

2.2 Choose the item : "Importing Dictionary, Data and Texts".

The window "Data Importation" is displayed.

2.3 Press the button: "Textual data (free format)" .
The window " Importation of a text file" is displayed.

2.4 Click on : "Open text file".
You have to select the file "TDA1_text_free.txt" in the directory: **EXD03.Importation.Text.Free.**

2.5 Click on "Convert into DtmVic file" .
The 100 first line of the new DtmVic text file are displayed in the right hand side memo.
A prudent message is given "Conversion apparently completed".
Two message boxes give successively the number of individuals (1043) and the maximum length of the identifiers (10).

- The DtmVic text file is automatically saved under the name: **"DtmTextFile.txt"**

2.6 Click then on the button : "Create a first parameter file".
The window "Creating a starting parameter file" appears.

[Reminder: In DtmVic, the phrases "Parameter file" and "Command file" are equivalent].

2.7 Click on "Create a parameter file".
A DtmVic parameter file (or: command file) is displayed in the lower window.
The command file is automatically saved under the name: **"param_tex_start.txt"**.
The command file does not include any statistical analysis command, except basic counts of words for the first open question (parameter NUMQ = 1 in the step SELOX).
It is only meant here as a check of the conversion of the data.

Optional comments about the "first parameter file"
After an identification of the two input files, three "steps" of DtmVic are involved:
The step "ARTEX" that archives the three sets of responses to the three open questions.
The step "SELOX" that selects the open questions for the subsequent processing. In this case, by default, the first question is selected.
The step "NUMER" that performs the numerical coding of the selected text.
The right hand side memo indicates how to run that parameter file.

2.8 Click on "Execute a command file".

2.9 Click on the button: "Basic numerical results" .
The button opens a created (and saved) html file named **"imp.html"** which contains the main results of the previous basic computation steps. After perusing these numerical results, **Return** to the main menu. Note that this file is also saved under another name: The name **"imp.html"** is concatenated with the date and time of the analysis (continental notation). That file keeps as an archive the main numerical results whereas the file **"imp.html"** is replaced for each new analysis performed in the same directory. Likewise, a simple text format file **"imp.txt"** is created and saved.

End of Example D.2

VI.3 Example D.3:

EX_D03.Importation.Text.num.XML

Importation of numerical and Textual data in "XML format".

A specific XML format allows for dealing with both numerical data and textual data in a unique file. Such format could be generated from some online questionnaires in the framework of mySQL databases.

The DtmVic format for textual data is described in chapter 1. It contains two types of separators: separators of individuals: "----" and separators of questions "++++", located in columns [1,2,3,4]. There is one constraint for the length of a line (200 characters) but, in principle, no constraint about the number of lines for one question or for one individual. However the number of open questions should not exceed 12, the number of closed questions should not exceed 1000, and the number of individuals is limited to 30000 in this version of DtmVic.

To begin with, we will have a look at the original XML data file to be imported.

1- Looking at the data, preliminary steps

We will use the editor of the button **"Open an existing command file"** from the main menu of DtmVic as a simple text editor.

1. Click on the button **"Open an existing command file"**
2. Search for the examples directory **DtmVic_Examples**
3. In it, open the directory of Example D.05, named **EX_D05. Importation.Text.num.XML**
4. Select the unique file: **"TDA2__dtm.xml"** .

(The data are the same as those of example A.5 of Chapter 3 : instead of three files – dictionary, numerical data, textual data - we have now only one (rather large) file).
The structure is schematised below: </div

```
<FileName.xml>

<individual>
<id > identifier1 </id >
< question1 > response1 < /question1 >
< question2 > response2 </question2 >
.....
< open >
<Open_quest_1> free response 1 </Open_quest_1 >
<Open_quest_2> free response 2 </Open_quest_2>
.....
</ open >
</ individual >

< individual >
<id> identifier2 </id>
< question1 > response2.1 </question1 >
< question2 > response2.2 </question2 >
.....
< open >
< Open_quest_1 > free response 2.1 </Open_quest_1>
< Open_quest_2 > free response 2.2 < /Open_quest_2 >
.....
</ open >
</ individual >
.....
.....
< individual >
.....etc.
</ individual >

</FileName.xml>
```

All the tags can be chosen by the user, except the tag < **individual** > that indicates an end of record. However; that keyword " **individual** " is only a default value. It can be changed during the importation process.
For an individual, a missing tag means "no-response". It is advisable, but not necessary, to put the tags in the same order.

The first tag after < **individual** > must be the identifier of the individual (tag <id> in the example, but any other name is acceptable).
Comments complying with XML syntax are possible anywhere in the file.
Note that this simple format is directly provided by saving the mySQL data bases derived from "on line surveys" as a simple XML file (without attributes, the tags being nested as shown before).

The drawbacks are the following:
- An obvious drawback of the XML structure is the size of the file, owing to the presence of opening and closing tags for each variable and individual.
- Another problem is the presence of XML-dedicated symbols such as: &, <, >, ', " The dictionary must not contain such characters.

The advantages are the following:
- A unique file replaces three files (dictionary, data and text).
- The order of variables can change from an individual to another.
- The order of individuals is no more important, since it is not necessary to match the data file and the text file.
- The length of individual records can vary, since the absence of a tag means "no response" to the corresponding variable.

The importation process consists in building the three internal DtmVic files (dictionary file, data file and, possibly, text file) from the original XML file.

2 - Sequence of operations:

2.1 Choose the item : " Data Importation, Preprocessing, Data Capture, Exportation".
(from the main menu of DtmVic). A new window appears.

2.2 Select the item: "Importing Dictionary, Data and Texts".

The window "Data Importation" is displayed.

2.3 Press the button: "XML specific file".
The window "Find and select the tags, Import XML data file" is displayed.

2.4 If the tag separating the individuals in your XML file is not the keyword " individual ", type your own tag in the first small white window. Press **"enter"** to register the new tag.

2.5 As explained in the pop-up purple window, two thresholds are necessary.

Threshold1 is the minimal number of respondent to an open question (default value: 40)
(that default value means that if the sample size is 1000, we tolerate 960 non-responses for some open questions). The question is discarded if the number of response is less than *Threshold1*.

Threshold2 is the minimal length of the lengthiest response to an open question (default value : 60) (that default value means that if all the responses to a question have less than 60 **characters**, this question will not be selected as an open question, and discarded)

Remind that in this version of DtmVic, the number of open-questions should be less than 12, whereas the number of closed question should be less than 1200.

If you wish to change the previous default values *Threshold1* and *Threshold2*, enter the new values in the two windows below (don't forget to press the **"enter"** button afterwards).

2.6 Click on: "List of tags and content"
You have to select the file **"TDA2__dtm.xml"** in the directory: **EXD04.Importation.Text.num.XML.**
Some messages are produced, describing the different steps of the process.

2.7 If the XML file contains responses to open-ended questions:
Click on **"Create the textual data file to be imported"** .

2.8 Always in the case in which the XML file contains responses to open-ended questions:
Click on: **"import as an internal DTM file"**
The 100 first lines of the new DtmVic text file are displayed in the right hand side memo.
A message box gives the number of individuals.
- The DtmVic text file is automatically saved under the name : **"Dtm_final_text_TDA2__dtm.xml.txt"**

- The DtmVic data file and the DtmVic dictionary file **remain to be imported** from the created csv file:
“ Dtm_import_num_TDA2__dtm.xml.txt” (importation as an Excel file).
- An intermediate file, **“Dtm_import_text_TDA2__dtm.xml.txt”** is also created, as a mere check. It is the text file in importation format. In fact, the importation of text has been completed and the final text has already been provided (**”Dtm_final_text_TDA2__dtm.xml.txt”**)

2.9 About the control files

Five control files are created to check the different steps of the process.

- The (huge) file: **“Check1_data_TDA2__dtm.xml.txt”** contains a list of all the tags encountered for all the individuals.
- The file: **“Check2_Tags_TDA2__dtm.xml.txt”** contains a list of all the encountered tags, with the parameters characterizing these tags (frequency, mean rank, average length of the content, minimum length maximum length). In this case, all the tags are present and have the same position.
- The file **“Check3_Dict_TDA2__dtm.xml.txt”** contains all the tags sorted according to their average rank (in this case the same rank for each individual), the tags corresponding to textual responses, the tags corresponding to numerical responses.
- The file **“Check4_Textual_TDA2__dtm.xml.txt”** contains all the encountered responses to open questions.
- The file **“Check5_final_text_TDA2__dtm.xml.txt”** complements the previous one.
- The file **“Check6_import_text_TDA2__dtm.xml.txt”** contains the open questions in the free text format.

These five files could be suppressed after checking the whole process.

As a conclusion:

The DtmVic file is created: (**”Dtm_final_text_TDA2_dtm.xml.txt”**).

The dictionary file and the data file have been converted from the XML file into a unique csv file.
They remain to be imported through a standard Excel importation process, using the created file: **“Dtm_import_num_TDA2__dtm.xml.txt”** as an input (see:

example D.1) (you may replace the extension ".txt" with the extension ".csv" to obtain an Excel file for the numerical data).

End of Example D.3

Some references

Alvarez R., Bécue M., Lanero J. J., Valencia O. (2002). Results stability in Textual Analysis: its Application to the Study of the Spanish Investiture Speeches (1979-2000). In*: JADT-2002, 6-th Intern. Conf. on Textual Data Analysis*, Morin A., Sébillot P., (eds), INRIA-IRISA, Rennes , 1-12.

Alvarez, R., Bécue et M., Valencia O. (2004). Etude de la stabilité des valeurs propres de l'AFC d'un tableau lexical au moyen de procédures de rééchantillonnage. In*: « Le poids des mots »,* Purnelle, G., Fairon, C., Dister, A., editors, PUL, Louvain, 42-51.

Balbi S. (1995). Confidence regions in factorial representations for textual data with non symmetrical correspondence analysis, in S. Bolasco, L. Lebart, A. Salem (eds.), III Giornate Internazionali di Analisi dei dati testuali, Roma, tome 2, 5-12.

Becue M. (1991) *Analisis de Datos Textuales*. CISIA, Saint-Mandé.

Benzécri J-P. (1973) *L'Analyse des Données*, Tome 1: *La Taxinomie*, Tome 2: *L'Analyse des Correspondances*, Dunod, Paris (2de. éd. 1976).

Benzécri J-P. (1992) *Correspondence Analysis Handbook*. Marcel Dekker New York.

Bird S., Klein E., Loper E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Sebastopol (Ca, USA).

Bolasco S. (2013). L'analisi automatica dei testi. Fare ricerca con il text mining. Carocci, Roma.

Bouroche J.-M., Saporta G. (1980) *L'analyse des Données.* Coll. Que Sais-je ?, PUF, Paris.

Bry X. (1995) *Analyses Factorielles Simples*. Economica, Paris.

Burt C. (1950). The factorial analysis of qualitative data. *British J. of Statist. psychol*. 3, 3, p 166-185.

Burt C. (1953). Scale Analysis and factor analysis. Comments on Dr Guttman paper. *British J. of Statist. psychol*. 6, p 5-20.

Chateau, F. and Lebart, L. (1996). Assessing sample variability in visualization techniques related to principal component analysis: *bootstrap* and alternative simulation methods. In : *COMPSTAT96*, A., Prats, editor, Physica Verlag, Heidelberg, 205-210.

Cox D. R. (1977). The role of significance tests. *Scandinavian Journal of Statist*., 4, p 49-70.

Efron B. (1979) Bootstraps methods : another look at the Jackknife, *Ann. Statist*., 7, p 1-26.

Efron, B. and Tibshirani, R, J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Escofier B., Pagès J. (1988) *Analyses factorielle simple et multiple*. Dunod, Paris.

Geary R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5, 3, 115-145.

Gifi A. (1990) *Nonlinear Multivariate Analysis*. Wiley, Chichester.

Gower J.C., Hand D.J. (1996) *Biplots.* Chapman and Hall, London.

Gower J.C., Ross G. (1969) Minimum spanning trees and single linkage cluster analysis. *Appl. Statistics,* 18, 54-64.

Gower, J., C. and Dijksterhuis, G. B. (2004). *Procrustes Problems*, Oxford Univ. Press, Oxford.

Greenacre M., Blasius J. (editors) (2006*) Multiple Correspondence Analysis and Related Methods*. Chapman and Hall/CRC, London.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Guttman L. (1941). The quantification of a class of attributes: a theory and method of a scale construction. In : *The prediction of personal adjustment* (Horst P., ed.) p 321 -348, SSCR New York.

Habert B., Nazarenko A., Salem A. (1997) *Les linguistiques de Corpus*. Armand Colin, Paris.

Hastie, T., Tibshirani R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer, New York.

Hayashi C., Suzuki T., Sasaki M. (1992) *Data Analysis for Social Comparative research: International Perspective*, North-Holland, Amsterdam

Hirschfeld H.O. (1935) - A Connection between correlation and contingency. *Proc. Camb. Phil. Soc.* 31, p 520-524.

Jambu M. , Lebeaux M-O. (1978) *Classification Automatique pour l'Analyse des Données*. Tome 1*: Méthodes et Algorithmes*, Tome 2*: Logiciels*. Dunod, Paris.

Kohonen T. (1989) *Self-Organization and Associative Memory*, Springer-Verlag, Berlin.

Lambert T. (1986) *Réalisation d'un Logiciel d'Analyse de Données*. (Thèse) Université de Paris-Sud, Dép. Statistique, Orsay.

LeRoux B., Rouanet M. (2009) *Multiple Correspondence Analysis*. Vol. 163, Sage Publication Inc.

Lebart L., Morineau A. (1982) *SPAD Système Portable pour l'Analyse des Données*. CESIA, Paris.

Lebart L., Morineau A. Bécue M. (1989) *SPAD.T Système Portable pour l'Analyse des Données Textuelles*, Manuel de Référence. CISIA, Paris.

Lebart L., Morineau A. Pleuvret P., Brian E., Aluja T. (1983) *SPAD Système Portable pour l'Analyse des Données*, Tome II. CESIA

Lebart L., Morineau A., Tabard N. (1977) *Techniques de la Description Statistique, Méthodes et Logiciels pour l'Analyse des Grands Tableaux*. Dunod, Paris.

Lebart L., Piron M., Morineau A., (2006) *Statistique Exploratoire Multidi-mensionnelle, Visualisation et Inférence en Fouille de Données.* Dunod, Paris. (4ème édition, refondue). [à consulter pour une bibliographie plus complète]

Lebart L., Piron M., Steiner J.-F. (2003) *La Sémiométrie*, Dunod, Paris.

Lebart L., Salem A. (1994*) Statistique Textuelle*. Dunod, Paris.

Lebart, L. (2003). Validation Techniques in Text Mining, in: *Text Mining and its Applications*, Spiros Sirmakessis, editor, Springer. 169-178.

Lebart, L. (2007). Which *bootstrap* for principal axes methods? In: *Selected Contributions in Data Analysis and Classification*, P., Brito *et al*., editors, Springer, 581 – 588.

Lebart, L., Morineau, A. and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis,* Wiley, New York.

Lebart, L., Salem, A. and Berry, E. (1998). *Exploring Textual Data*, Kluwer Academic Publishers, Dordrecht.

Lerman I. C. (1981). *Classification et Analyse Ordinale des Données*. Dunod. Paris.

Marano P. (1972) Applications de l'analyse factorielle des correspondances à la compression de signaux d'images. *Annals of Telecommunications*, vol. 27, n° 5-6, 163-172.

Marchand P. (1998) *L'Analyse de Discours Assisté par Ordinateur*. Armand Colin, Paris.

Milan, L. and Whittaker, J. (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Applied Statistics*, 44, 1, 31-49.

Murtagh F. (2005) . *Correspondence Analysis and Data Coding with R*. Chapman and Hall, Boca Raton, USA.

Ratinaud P. (2011). [http://www.iramuteq.org/Members/pierre.ratinaud].

Reinert, M. (1983). "Une méthode de classification descendante hiérarchique: Application à l'analyse lexicale par contexte". *Cahiers de l'Analyse des Données*, 3, 187-198.

Reinert, M. (1986). Un logiciel d'analyse lexicale: [ALCESTE]. *Cahiers de l'Analyse des Données*, 4, 471–484.

Roux M. (1985) *Algorithmes de Classification*. Masson, Paris.

Salem A. (1987) *Pratique des segments répétés, Essai de satistique textuelle*, Klincksieck, Paris

Saporta G. (1990) *Probabilités, Analyse des Données et Statistique*. Technip, Paris.

Schmid H. (1994). Probabilistic part of speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Spearman C. (1904). General intelligence, objectively determined and measured. *Amer. Journal of Psychology,* 15, p 201-293.

Tenenhaus M. (2007) *Statistique*. Dunod, Paris.

Tuffery S. (2006) *Data Mining et Statistique Décisionnelle*. Technip, Paris

Tuzzi, A. and Tweedie, F., J. (2000). The best of both worlds: Comparing Mocar and Mcdisp. In: *JADT2000* (*Cinquièmes Journées Internationales sur l'Analyse des Données Textuelles*), Rajman, M., Chappelier, J-C., editors, EPFL, Lausanne, 271-276.

Vapnik W. (1998). *Statistical Learning Theory*. Wiley, New York.

Volle M. (1980) *Analyse des Données*, Economica, Paris.