

# An Introduction to DtmVic

## *Five elementary examples to discover DtmVic*

*The following five 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.*

## Application examples A.1—A.5

[To select the examples, press the right button of the mouse](#)

- Example A.1.**      **[EX\\_A01.PrinCompAnalysis.](#)**  
*(Principal Components Analysis)*      Page A.2  
Active and supplementary variables. Supplementary categories. Bootstrap validation.  
PCA is followed by a clustering of observations, and a description of the obtained clusters.
- Example A.2.**      **[EX\\_A02.SimpleCorAnalysis.](#)**  
*(Correspondence Analysis)*      Page A.7  
Correspondence Analysis of a small contingency table. Bootstrap validation.
- Example A.3.**      **[EX\\_A03.MultCorAnalysis.](#)**  
*(Multiple Correspondence Analysis)*      Page A.11  
Active and supplementary categories. Bootstrap validation.  
MCA is followed by a clustering of observations, and a description of the obtained clusters.
- Example A.4.**      **[EX\\_A04.Text-Poems.](#)**  
*(Correspondence Analysis of a lexical table)*      Page A.16  
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.
- Example A.5.**      **[EX\\_A05.Text-Responses](#)**  
*(Open questions in a sample survey)*      Page A.21  
Using both numerical and textual data. Processing of the responses to an open-ended question using a specific categorical variable. 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. Simultaneous Kohonen map for words and categories.

## Example A.1: **EX\_A01.PrinCompAnalysis** (*Principal Components Analysis*)

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](http://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.

### 1) Looking at the two files: 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”** (click directly on the file in text format or use your text editor (notepad, notepad ++, ultraedit, TotalEdit, etc.). 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 Tutorial D: “Importation”). 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 ( $\leq 6$ ) describe the cross-tabulation “gender – age”, the second digit ( $\leq 2$ ) the activity, the third digit ( $\leq 3$ ) 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”)

Open DtmVic.

2.1) Click the button: **“Create”** of the main menu: Basic Steps, line **“Command File”**.

A window **“Choosing among some basic analyses”** appears.

2.2) Click then the button: **“PCA– Principal Components Analysis”** – 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 is 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 variables)”**.

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 to select 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”** (line: **“Command file”**) to open directly **“Param\_PCA.txt”**, and, in so doing, reach this point of the process, using 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<sup>th</sup>, 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 is also saved under a simple text format, under the name **“imp.txt”**, and likewise with a name including the date and time of execution.

## 4) Steps VIC (Visualization, Inference, Classification)

### 4.1) Click **“Axesview”** 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 Axesview 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.

#### 4.2) Click “**PlaneView**” 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.

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.

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).

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).

*[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, 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.

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.
- 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” 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” 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 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 A1**

## Example A.2: **EX\_A02.SimpleCorAnalysis** (*two-way Correspondence Analysis*)

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

DtmVic generates several intermediate text-files related to the 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](http://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,

#### 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 - also known as teletype font - such as “courier” 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**” of the main menu “Basic Steps”, 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 is 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 variables)”**

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: **“Select the observations from a list”**. Select then the first eight rows (occupations) as “active observations” 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 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”** (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”**.

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\_08.07.09\_14.45.html”** means July 8<sup>th</sup>, 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 is also saved under a simple text format, under the name **“imp.txt”**, and likewise with a name including the date and time of execution.

## 4) Steps VIC (Visualization, Inference, Classification)

### 4.1) Click the **“Axesview”** 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 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”** printed in the log-file **“imp.txt”**.

Evidently, the use of the Axesview menu makes sense when the data set is very large.

### 4.2) Click the **“PlaneView”** 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 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 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).

### 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.*

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" 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" 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 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 A2**

## Example A.3: **EX\_A03.MultCorAnalysis** (*Multiple Correspondence Analysis*)

Example A.3 aims at describing a set of categorical variables through MCA.

The corresponding data are located in the subdirectory: **EX\_A03.MultCorAnalysis** of 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](http://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. In this version of DtmVic, the identifiers of categories (in the internal DtmVic format) 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]. 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 labels corresponding to each response items. An optional “short identifier” occupies columns 1 to 5. A numerical variable (such as “age”) has 0 category.

#### 2.1) 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”** of the main menu: Basic Steps, line **“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”**.

#### 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 is 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 variables)”.**

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.

**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”.

**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 the 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”** (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”**.

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\_08.07.09\_14.45.html”** means July 8<sup>th</sup>, 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 is also saved under a simple text format, under the name **“imp.txt”**, and likewise with a name including the date and time of execution.

### 4) Steps VIC (Visualization, Inference, Classification)

**4.1) Click “Axesview” 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 **“Main basic results”**). The use of the **Axview** menu is profitable when the data set is very large.

**4.2) Click PlaneView 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.

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. 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). 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).

### 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”.

In the same vein, we can tick the classes of age, 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 RECI and PARTI of the created command file “**Param\_MCA.txt**”, i.e.: Clustering using reciprocal neighbours (RECI), 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 DECI (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**” 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” 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**

## **Example A.4: EX\_A04.Text-poems**

*(Textual Data Analysis: simple series of texts)*

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/](http://www.shakespeare-online.com/sonnets/).

In this simple format, DtmVic can process up to 1000 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](http://www.dtmvic.com) (section "publication").

### **1) Looking at the text file**

Search for the examples directory: "**DtmVic\_Examples\_A\_Start**"

In that directory, open the directory 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:

Look at the text file: "**Sonnet\_LowerCase.txt**". (using a text editor such as Notepad, Notepad++, TotalEdit, Ultraedit, TextEdit, etc.)

The format is specific (DtmVic internal text format type 1. See the tutorial D: "Importation"). Since the texts may have very different lengths, separators \*\*\*\* (at the beginning of a line) are used to distinguish between texts. 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")**

**Open DtmVic.**

**2.1) Click the button: "Create" of the main menu: Basic Steps, line "Command File".**

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

**2.2) Click then the button: "VISUTEX – Visualization of Texts"** – located in the paragraph "Textual and numerical data".

**2.3) Press the button: "Open the 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**".

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).

#### 2.5) Click directly **“Vocabulary”**.

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 indeed a very small frequency, adapted to a very small corpus! It is just an opportunity of exploring the sequence of commands, without meaningful linguistic interpretation...).

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

#### 2.6) Then 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).

#### 2.7) Then click **“Create a first parameter file”**

A parameter file is displayed in the memo [A similar parameter file is commented in the Appendix A. 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 director.

*If you wish, you could now exit from DtmVic, and, later on, use the button of the main menu **“Open”** (line: **“Command file”**) to open directly **“Param\_VISUTEX.txt”**, and, in so doing, the user reaches this point of the process. You can then use afterwards the **“Execute”** command of the main menu.*

#### 2.8) Click on: **“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.

### 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 8<sup>th</sup>, 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 is 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).

## 4) Steps VIC (Visualization, Inference, Classification)

### 4.1) Click the button: “Axesview”

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 Axeview menu is justified when the data set is large, which is not the case here.

### 4.2) Click the button: “PlaneView” , 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).

### 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\_var\_boot.txt”. (The set of possible files is given by the 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 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 (or individuals or observations) in blue. We can see that many sonnets occupy significant locations (several confidence ellipses do not overlap) whereas the locations of the words is 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” 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 (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).

#### 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.

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 of the correspondence analysis of the lexical table.

#### 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.

#### *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” 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” 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**

## **Example A.5: EX\_A05.Text-Responses** *(Textual Data Analysis: Open and Closed Questions)*

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.

Those 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 tutorial, 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.

This example corresponds to the directory “**EX\_A05.Text-Responses**” included in: “**DtmVic\_Examples\_A\_Start**”.

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

**1.1) 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).

#### **1.2) Dictionary file: “TDA\_dic.txt”**

The dictionary file “TDA\_dic.txt” contains the identifiers of the 14 variables. In this internal 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” should be used to facilitate this kind of format]. 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.

#### **1.3) Text file: “TDA\_tex.txt”**

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”.

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).

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 and responses are computed for each category.

## 2) Generation of a command file (or: “parameter file”)

### 2.1) Click the button: “**Create**” of the main menu: Basic Steps, 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 the text file**”, then search for the directory:

“**DtmVic\_Examples\_A\_Start**”. In that directory, open the directory of example A.4, named “**EX\_A05.Text-Responses**”.

Open then the dictionary 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**”.

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 is 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.

We suggest to select the categorical variable number 14, (age - education). Only one active variable is selected in this 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.

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,

### 2.10) 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”** (line: **“Command file”**) to open directly **“Param\_ANALEX.txt”**, and, in so doing, reach directly this point of the process, using the **“Execute”** command of the main menu.*

### 2.11) 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 8<sup>th</sup>, 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 is 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 total number of kept words reduces to 10 738, whereas the number of distinct kept word 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: “AxesView”...** and follow the sub-menus. In fact, only one tab is relevant for this example: **“Active variables”** [= words and categories, in the case of step APLUM]. Rows and columns of the lexical table are merged. After clicking on **“View”**, 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. Evidently, the use of the Axesview menu is justified when the data set is large, which is the case here.

**4.2) Press the button: “PlaneView”...** and follow the sub-menus...

In this example, only one item of the menu is relevant **“Active columns (variables or categories)”**. In fact, this item concerns both rows and columns of the contingency table (lexical table). The graphical display 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\_var\_boot.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).

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) 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.

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 on **“ClusterView ”**

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).

#### 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 of the correspondence analysis of the lexical table.

#### 4.6) Click **“Seriation”**

The aim of seriation techniques has been briefly described in the section 5 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”** 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”** 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 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\_ANALEX.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 A5**

---

**End of tutorial A**

---